Towards Autonomous PErceptual Systems

Stéphane Herbin

SORBONNE UNIVERSITÉ

ONERA
THE FRENCH AEROSPACE LAB

Towards Autonomous PErceptual Systems

Mémoire en vue de l'obtention de l'Habilitation à Diriger des Recherches de Sorbonne Université
Spécialité: Sciences pour l'Ingénieur

HDR présentée et soutenue le 6 juillet 2020 par

Stéphane Herbin

devant le jury composé de:

François Brémond *Directeur de Recherche, INRIA*	Rapporteur
Raja Chatila *Professeur, Sorbonne Université*	Président
Michel Crucianu *Professeur, CNAM*	Rapporteur
David Filliat *Professeur, ENSTA-ParisTech*	Rapporteur
Jean Ponce *Directeur de Recherche, INRIA*	Examinateur

Abstract

The objective of this document is to introduce the principle of Autonomous PErceptual System (APES) as an object of study.

The functionalities of artificial perception, in particular vision, have become both easier to design and more efficient through the use of a set of techniques and development environments grouped under the term “Deep Learning”. They have reached a certain level of maturity making it possible to envisage their use for real or even critical applications.

The research direction proposed here is to provide perception with a certain degree of autonomy envisaged as a means of guaranteeing its reliability.

The introduction of such a property implies to reconsider the status of perception no longer as passive functionality but as an activity involving as explicit stakeholders the environment to be perceived but also the recipient of the perceptual products with which the system maintains a contractual relationship determining the nature of the expected service and the means to guarantee it.

The study of autonomous perceptual systems thus leads to a research program organized along three axes: the design of a perceptual activity articulating functional dynamics and learning processes, the development of an inherent intelligibility of the mechanisms of perception for monitoring, specifying or justifying their behavior, and the implementation of a general approach to guarantee their safe and controlled use.

Résumé (Français)

L’objectif de ce mémoire est d’introduire le principe de système perceptif autonome comme objet d’étude.

Les fonctionnalités de perception artificielle, en particulier de vision, sont devenues à la fois plus faciles à concevoir et plus performantes par l’utilisation d’un ensemble

de techniques et d'environnements de développement regroupés sous l'expression apprentissage profond (Deep Learning). Elles ont atteint un certain niveau de maturité permettant d'envisager leur utilisation pour des application réelles voire critiques.

La direction de recherche proposée ici est de munir la perception d'un certain degré d'autonomie considéré comme moyen de garantir sa fiabilité.

L'introduction d'une telle propriété implique de reconsidérer le statut de la perception non plus comme fonctionnalité passive mais comme une activité impliquant comme parties prenantes explicites l'environnement à percevoir mais également le destinataire des produits perceptifs avec lequel le système entretient une relation contractuelle déterminant la nature du service attendu et les moyens de le garantir.

L'étude des systèmes perceptifs autonomes conduit ainsi à un programme de recherche organisé selon trois axes: la conception d'une activité perceptive articulant dynamique fonctionnelle et processus d'apprentissage, le développement d'une intelligibilité propre des mécanismes de perception pour surveiller, spécifier ou justifier leur comportement, et la mise en oeuvre d'une démarche générale permettant de garantir leur utilisation sûre et maîtrisée.

Remerciements

Tout d'abord, je remercie les membres du jury d'avoir accepté de participer à ma soutenance dans cette période particulière: François Brémond, Michel Crucianu et David Filliat, qui en outre ont pris de leur temps pour rapporter un manuscrit sans doute un peu long, Jean Ponce depuis New York et Raja Chatila qui a tenu le double rôle de président et de garant du bon fonctionnement des moyens modernes de communication, souvent capricieux. Je suis très honoré d'avoir pu compter sur leur présence – distancielle – et ressorti encouragé par les échanges bienveillants qui ont suivi mon exposé.

Ma motivation pour présenter cette HDR est de continuer à initier les futurs docteurs au domaine passionnant de la recherche et à ses codes. Je suis très heureux d'avoir pu accompagner dans leur thèse Benjamin Francesconi, Jonathan Guinet, Anne-Marie Tousch, Christophe Guilmart, Joseph Defretin, Isabelle Leang, Cédric Le Barz, Maxime Bucher et Alexis Lechat en collaboration avec leurs co-encadrants Bernard Chalmond, Sylvie Philipp-Foliguet, Jean-Yves Audibert, Patrick Pérez, Nicolas Vayatis, Benoît Girard, Jacques Droulez, Matthieu Cord et Frédéric Jurie. Je les remercie tous de m'avoir fait confiance.

Cette Habilitation à Diriger des Recherches vient après plus de vingt années passées à l'ONERA. Cette période a bien sûr été l'occasion d'interagir fructueusement avec de nombreux collègues: Adrien Chan Hon Tong, Guy Le Besnerais, Frédéric Champagnat, Philippe Cornic, Fabrice Janez, Elise Colin-Koeniguer, Anne Beaupère, Pauline Trouvé-Peloux, Martial Sanfourche, Valérie Leung, Alexandre Boulch, Bertrand Le Saux, Claire Pagetti, Frédéric Boniol, Benjamin Pannetier, Patrick Secchi, Fabrice Savignol, Gilles Foulon, Alain Michel, Jérôme Besombes, Olivier Poirel, Valentina Dragos, Jean-Christophe Sarrazin, Bruno Berbérian, Jean-Loup Farges. Ils ont influé, d'une manière ou d'une autre, les travaux et idées présentés dans ce mémoire. Mes remerciements vont également à Philippe Bidaud, Directeur Scientifique du domaine TIS, et à Virginie Wiels, Directrice du DTIS, pour leur soutien constant au développement des activités de recherche à l'ONERA auxquelles, j'espère, contribue cette HDR.

Contents

Introduction 1

Context

The “Deep Learning” era

This memoir is about artificial perception in a historical context where generic problems such as object detection or image classification, that had been previously estimated unsolved, are now considered for practical and even critical applications. This evolution is mostly caused by the surge of deep learning techniques that have been able to federate (or dominate) various subdomains of Artificial Intelligence (AI) under a common formalism – data-driven modeling and optimization or Machine Learning (ML).

Computer Vision (CV), and to a lesser extent Natural Language Processing (NLP), have been the core domains where this transformation was the most dramatic, with the breakthrough of the deep convolutional network now known as *AlexNet* at the ImageNet Large Scale Visual Recognition Challenge (Krizhevsky et al., 2012; Alom et al., 2018). Since then, ML has become the leading conceptual paradigm for almost all problems of CV, alternatives being forced to be justified and compared to it. As an indirect consequence, it also gave a new visibility to Artificial Intelligence, popularizing the expression as one of the keywords of many research fields.

The success of deep learning approaches is not only due to its empirical efficiency: their flexible modularity makes easier the integration and combination of multiple constraints, objectives and functional structures in a common framework, helped by the availability of programming environments and a large community.

A consequence of this ease of development is an increased variety of functional structures, adding closed-loop, sequential or interactive designs to feed-forward processing pipe-lines. Merging various sources of information in complex static or dynamic architectures, and optimizing globally their parameters from a data-driven “end-to-end” criterion, is now the standard routine.

Of course, Deep Learning is not the solution to every problem: good performance rely on the availability of large quantities of data, condition that may not be met in many practical situations, motivating the development of other settings able to address data scarcity.

The current hegemony of ML has also one side effect: that of overvaluing benchmarking on research practice. Most studies in current literature contain a large experimental section comparing proposed solutions on shared databases and eval-

uation metrics, but often with objectives that do not differentiate clearly scientific investigation from goal oriented engineering.

Maturing perception

In this context of unbounded creativity, the design of AI processes has evolved towards more functional complexity and possibilities, but has also given rise to new issues: formal reliability assessment, justification, explanation, verification, trustworthiness, certification, etc. are now currently used keywords revealing the level of maturity pursued for AI functions and also the difficulty to achieve such objectives when dealing with Machine Learning enabled approaches.

Perceptual *functions* follow the same trend but with some specificity: the *very* high dimension of input and inner spaces involved makes full knowledge of the environment impossible at design time, requiring working hypotheses, simplifying priors or online data gathering strategies to fill potential ignorance or uncertainty gaps.

However, *Perception* as a generic capacity is more than a collection of functions. It is indeed *the* interface with the external environment, but active and dynamic in the sense that it brings the world into existence with a selective and purposeful point of view. Perception is not mere passive signal processing.

The fundamental option taken here is that the right level of study for a mature perception is to address both agency and cognitive dimensions. Agency implies that the substratum of perception itself is a structure of dynamic components in interaction with an environment and other agents. By cognitive we refer to capacities such as memory management, reasoning, planning, learning, knowledge, decision, multi-modal integration, language, etc. which are usually seen as high-level but that are needed to implement perception.

Once the status of a cognitive agent is granted to perception, the question arises of its *autonomy*: how to exploit efficiently its own capacities and resources, but also what are its responsibilities about made commitments, for example in terms of accuracy, relevance, response time, energy expense, etc.

Besides the underlying scientific issues, considering perception as a cognitive agent is also an engineering ambition. The underlying assumption of the working program described in this document is that the road to a mature and reliable artificial perception is to address it as an Autonomous PErceptual System (APES).

This document

The general objective of this document is to introduce the idea of Autonomous PErceptual System and its implications in terms of research directions.

Its history started as a way to present in an organized and meaningful way my previous works, most of it accomplished at ONERA, and ended in a report between an essay, a research program or a roadmap, with (too) numerous references. I apologize for this evolution that has considerably increased the length of the original writing project.

The core of the document is divided in 4 chapters and a conclusion. Several contents are highlighted with two types of boxes: squared for previous work, light blue filled for research directions.

The goal of chapter 2 is to define more precisely the idea of APES as a means of building reliable artificial perceptual systems, but also to question current approaches in light of natural and cognitive science findings. In particular, it examines what it implies for perception to be *autonomous* per se as opposed to the idea of endowing an autonomous system with perceptual capacities.

Chapter 3 discusses the state of the art of formalism and techniques developed to implement APES's, and emphasizes two fundamental conceptual ingredients: active perception and machine learning. This chapter also contains short descriptions of most of my previous research activity and their underlying contribution to the development of APES's – a kind of fictive post-hoc story-telling.

One essential feature to ensure autonomy in perception is a capacity to express representations or signs of their inner states or processes that are *intelligible* to their recipient. Chapter 4 discusses the issues of interpretability or explainability – which are becoming major concerns of the AI community – and their instantiation for perceptual systems.

The boost in performance offered by deep learning approaches has paved the way for the actual integration of perceptual capacities in real and even critical systems. Chapter 5 reviews how to assess their safety and reliability, what are the current limitations and proposes several research actions.

2 Autonomous PErceptual Systems

This chapter introduces and justifies the concept of Autonomous PErceptual System (APES) as a research topic with the objective of identifying its main issues. The underlying idea of introducing such object is to examine the possibility of equipping perception with a certain level of autonomy and to study what is impacted by this property.

The chapter is organized in three sections: the first one examines what it means for perception to be considered as an autonomous system; a second section discusses several findings from cognitive science and philosophy that could inspire engineering; the last section describes general issues that should be addressed to specifically study APES.

2.1 What are APES?

A first idea to start investigations is to find out examples that could fit under APES category before precisely defining it:

- softwares able to describe in a textual or visual form the content of an image or a video (object localization and names, actions, events, etc.) as an answer to a request for information,
- smart sensors, for instance Pan-Tilt-Zoom (PTZ) cameras or Unmanned Aerial Vehicle (UAV) with sensors, combining controllable measurement device, viewpoint and processing for feature extraction, image quality enhancement, moving object detection,
- multi-sensor networks or robots that allocate and combine resources to complete multiple interpretation tasks such as people tracking, abnormal behavior detection, action recognition,
- multimedia database management systems, able to navigate in large data corpora, retrieve, organize and present to a user a set of data that corresponds to her/his needs,
- data mining software suites that visualize, combine, summarize, reveal regularities or accidents in heterogeneous datasets,
- search engine able to retrieve, suggest, notify digital content according to user requests or habits.

Although those examples target different application contexts and communities, they all share the same characteristics of *purposively* dealing with a *contingent*

environment, world and/or user with which they possibly *interact* by emitting actions or outputs and receiving inputs or requests.

The rest of this section will develop in details APES specificities: perceptual ability, systemic structure and autonomy.

APES are Perceptual Systems

The foremost feature of APES is their perceptual ability. We need to define with care what is implied by that faculty shared by most of living beings. A starting point is therefore to study how natural perception is addressed and defined before determining in what sense it can be understood for artificial entities.

Defining perception

Accurately defining *perception* is not a straightforward task as it potentially involves considerations from various knowledge areas. A starting exercise to figure out what is implicated when speaking of Perception is to compare the various definitions of the term that have been proposed in dictionaries and analyze their differences.

A first corpus of definitions — multilingual — can be found in Wikipedia, which itself refers to other sources:

English "Perception (from the Latin *perceptio*) is the organization, identification, and interpretation of sensory information in order to represent and understand the presented information, or the environment." [1]

French "La perception est l'activité par laquelle un sujet fait l'expérience d'objets ou de propriétés présents dans son environnement. Cette activité repose habituellement sur des informations délivrées par ses sens." [2]

Spanish "La percepción es la forma en la que el cerebro detecta las sensaciones que recibe a través de los sentidos para formar una impresión consciente de la realidad física de su entorno (interpretación)." [3]

Italian "La percezione è il processo psichico che opera la sintesi dei dati sensoriali in forme dotate di significato." [4]

[1] https://en.wikipedia.org/wiki/Perception.
[2] https://fr.wikipedia.org/wiki/Perception. Perception is the activity by which a subject experiences objects or properties present in his environment. This activity is usually based on information delivered by his senses.
[3] https://es.wikipedia.org/wiki/Percepcion. Perception is the way in which the brain detects the sensations it receives through the senses to form a conscious impression of the physical reality of its environment (interpretation).
[4] https://it.wikipedia.org/wiki/Percezione. Perception is the psychic process that operates the synthesis of sensory data in meaningful form.

German "Wahrnehmung (auch Perzeption genannt) ist der Prozess und das Ergebnis der Informationsgewinnung und verarbeitung von Reizen aus der Umwelt und dem Körperinnern eines Lebewesens." [5]

Other definitions can be found in several on-line dictionaries or encyclopedias:

Trésor de la Langue Française "Opération psychologique complexe par laquelle l'esprit, en organisant les données sensorielles, se forme une représentation des objets extérieurs et prend connaissance du réel." [6]

Académie Française "Acte par lequel le sujet se forme la représentation d'un objet appréhendé par les sens." [7]

Larousse "Événement cognitif dans lequel un stimulus ou un objet, présent dans l'environnement immédiat d'un individu, lui est représenté dans son activité psychologique interne, en principe de façon consciente ; fonction psychologique qui assure ces perceptions." [8]

Petit Robert "Fonction par laquelle l'esprit se représente les objets ; acte par lequel s'exerce cette fonction ; son résultat." [9]

Encyclopedia Britannica "Perception, in humans, [is] the process whereby sensory stimulation is translated into organized experience."

Several definitions are minimal: the Italian introduces the idea of meaningfulness, and the Encyclopedia Britannica relates senses to experience. Several mention the idea of interpretation. Other definitions speak of a subject or an individual hosting perception, and introduce concepts of mind and consciousness. Several definitions assume that senses hold information. Many introduce objects as the reference of representations. Most refer to an outer environment or a reality. One definition (Spanish) involves the brain as the entity hosting perception. The German definition mentions inner body perception.

None of the previous definitions are conceptually neutral: they rely on a certain interpretation of the nature of perception, and generate sometimes more issues than validated facts. For instance, what is exactly the status of the objects represented in perception, if any? Are they real (in what sense?) or mental constructions? Is perception a reliable account of reality, or a subjective construction, an unconscious

[5] https://de.wikipedia.org/wiki/Wahrnehmung. Perception is the process and result of information gathering and processing of stimuli from the environment and the internal body of a living being.
[6] http://www.cnrtl.fr/definition/perception. Complex psychological operation by which the mind, by organizing sensory data, forms a representation of external objects and becomes aware of reality.
[7] https://academie.atilf.fr/9/consulter/perception?page=1. Act by which the subject forms the representation of an object apprehended by the senses.
[8] https://www.larousse.fr/dictionnaires/francais/perception/59399?q=perception. A cognitive event in which a stimulus or object, present in the immediate environment of an individual, is represented to him in his internal psychological activity, in principle in a conscious manner; psychological function that ensures these perceptions.
[9] Function by which the mind represents the objects; act by which this function is exercised; its result.

inference (Helmholtz and Southall, 1924)? Those questions have been addressed and debated extensively in philosophy and psychology: the goal of this section, however, is not to present the various theories of perception, but only to point out that perception resists unquestionable definition in spite of centuries of studies. "In spite of the interest constantly aroused by the study of perception throughout the history of Western philosophy and despite the enormous contribution made on this subject by psychology since the moment when it tried to define itself as a science, this subject is typically a domain that resists both concrete observation and abstract analysis" [10] (Thinès, 2016).

For the following of this document, I will propose a minimal — i.e. reduced to essential ingredients — definition to avoid underlying and undeclared conceptual hypotheses, and also as an attempt to express a statement that can be shared between natural and artificial perception:

Minimal definition Perception is an activity informed of the external world through sensory input.

This definition only states that perception is the encounter of contingent data or stimuli originating in senses and of an active process that may find or extract information about an external world from it.

The important point that the proposed definition emphasizes is that perception is an *activity*, that may potentially involve complex mechanisms and may take various forms (construction process, synthesis, act, etc.).

The other specificities of this definition are the omissions: nothing is said about the purpose or result of perception, for instance a representation of the environment or world, nor that it has to be hosted by a mind or a consciousness which may or may not have access to its results.

The next section examines how this essential definition can be declined and updated for artificial perceptual systems.

The roles of artificial perception

A current difficulty when trying to describe and model perception when hosted by an artificial entity is the comparison with our own senses: seeing, hearing, smelling, tasting, touching are natural faculties that we experience "without knowing how they work". Humans have no difficulty in describing what they see in an image or in a video and in reasoning about the cause and consequences of the observed

[10] Original text in French: "*Malgré l'intérêt incessant qu'a suscité l'étude de la perception tout au long de l'histoire de la philosophie occidentale et malgré l'énorme contribution, sur ce sujet, de la psychologie depuis l'époque où celle-ci a tenté de se définir comme science, ce topique constitue, par excellence, un domaine qui résiste à la fois à l'observation concrète et à l'analyse abstraite.*"

phenomena. The holistic experience of natural perception is also an obstacle to its unequivocal and informative definition as was shown previously.

It is a platitude to state that this easiness and efficiency is not shared by artificial devices. The expression "semantic gap" has been coined to refer to this problem and expresses the fact that the information encoded in computers does not isomorphically match the inner structure of sensory data.

Another specificity of artificial perceptual entities is that they are required to externalize some content: if the nature or even existence as such content is debatable for natural perception and may lead to conceptual aporia, an artificial perceptual system is used for some goal. Engineering is more comfortable when a criterion to optimize or satisfy is exhibited, and when quantitative evaluation means and validation protocols can be applied.

I propose two possible distinct roles for perception when hosted by an artificial system: as a measuring device or as a sign producer.

Perception as a measuring device

In this first role, the objective for perception is to provide representations (topological, geometric, radiometric, spectral etc.) of the environment. This objective is consistent with the idea that senses provide us a direct and faithful account of the outside world, as proposed in several definitions of page 6 — a approach sometimes called *naive realism* in philosophy.

One clear advantage of this role is that perception production can be compared to an ideal content, a *ground truth*, validated by other measuring means. Perception can be seen as an inversion process, estimating or calculating physical features from given sensory data. The representations can be used as a formal substitute of the environment — a "digital twin" to invoke a fashionable expression — and exploited later on for various inference tasks.

Another advantage of this objective of perception is the expected universal character of the constructed representation obtained firstly by eliminating subjective idiosyncrasies — the norm of perception is the physical world, the reality — and secondly by ignoring the potential usage of representations, making perception without any specific purpose, ateleological. This lack of objective, however, may lead to uselessly detailed representations or processes.

A typical role of perception as a measuring device is visual pose estimation of an object or a person, i.e. a function able to describe the geometry and configuration of an object relatively to known landmarks or axes.

Perception as sign production

The second proposed role of perception is to consider it as a generator of objects that stand for some feature or property of the world and is meaningful to a dedicated user to serve her/his needs. This approach relaxes the ambition of universal and true representation of the environment, and involves the recipient of perception outputs, be it formal, material or human, as a key element: no perception without declared user that assigns meaning of it.

The idea that perception content is between world and user, ontologically associated with them, can be related to the concept of a *sign*, defined by Peirce (Chandler, 2007, page 29) as a triadic (three-part) model[11]:

1. The *representamen*: the form that the sign takes – the 'sign vehicle'.
2. An *object*: something to which the sign refers (a referent), or which it represents.
3. An *interpretant*: the effect produced by the sign or the *sense* made of it.

Considering perception as sign production, in the framework proposed by Peirce, leads to two consequences:

- perception is only "real" when instantiated, i.e. when the triadic semiotic relation becomes actual: the interpretant hosts a dynamic process — the *semiosis*;
- perception is inherently *to* someone (the *interpretant* in Peirce vocabulary), implying that its value depends on a final usage.

Peirce is also well known for having introduced three different ways to relate the sign vehicle and what it refers to (Chandler, 2007, page 41):

a. *Symbolic*: based on a relationship which is fundamentally unmotivated, arbitrary, and purely *conventional* (rather than being based on resemblance or direct connection to physical reality) – so that it must be agreed upon and learned.

b. *Iconic*: based on perceived *resemblance* or imitation (involving some recognizably similar quality such as appear - ance, sound, feeling, taste, or smell).

c. *Indexical*: based on *direct connection* (physical or causal). This link can be observed or inferred.

Usual perceptual outputs can be analyzed using those categories: in computer vision, object detection described as bounding boxes in an image, or even better as a mask, can be interpreted as having an iconic link to the object since its shape resembles the object geometric extension, a label is clearly symbolic since conventional, and the optical flow is indexical, as it is caused by the object motion. Measure values should

[11] There is another tradition of semiotics that are more comfortable with a dyadic model — signified/signifier — as proposed by Saussure (Chandler, 2007, page 13).

be considered as indexical since they are causally generated by the environment or by object properties. A bounding box used to encode the object size, not its shape, has therefore to be understood as an index, showing that a given formal representation can have two different meanings from a semiotic point of view.

Interpreting perception as sign production has a direct implication when trying to design and evaluate artificial perception: the relation between the sign-vehicle and the object can be seen as the usual function associating input sensory data (an image, a signal, etc.) to an output (a bounding box, a label, a flow field, a pose etc.). However, the triadic form involves a third stakeholder, the interpretant, i.e. the entity for which the sign meaningfully refers to an object. This implies that perception objectives should also be characterized by the types of sign that are produced (the representamen/object/intepretant triplets). One should not speak of visual object recognition in a neutral way, i.e. by only describing the input/output form of a function that produces "universal" representations, but should explicitly declare *for* what recipient: for a robot that will grasp the object, for a tracker that will maintain a label, for a database manager that will organize a dataset, etc.

One reason to introduce the idea of perception as sign production is when it has to do with *semantics*: in perceptual tasks such as object categorization, captioning, image segmentation, sound identification etc., the relation between the output and the object it refers to must be agreed upon and shared by users, i.e. *interpreters*, to be trusted.

One simple way to reveal the impact of involving an interpreter in "semantic" perception is when the perceptual task is to name observed objects. What is the best word to issue to describe the seen object? What categorization level target (basic, superordinate, subordinate or fine grained)? (Rosch, 1999; Tousch et al., 2012) For instance, when observing a car, it is not obvious to prefer to describe it as SUV, a family car, a Peugeot 3008, or simply as a vehicle. Identifying the right level is usually resolved by restricting the potential set of labels among which to choose, often in a rather arbitrary way. However, when the output is a caption, i.e. a free form text, it appears much more difficult to constrain the length or expressiveness level, and even define an ideal caption without knowing how it will be used, for what and by whom. Unequivocally defining ideal outputs is not straightforward, and we will see that, even on non semantic problems such as contour identification (chapter 5, pg. 144) that trying to eliminate the role of the interpreter introduces more difficulties than simplifications.

Perception as a system

The concept of perception as a system is not new and has been approached either to model natural perception or to design artificial algorithms and devices. Basically, what is defended is the idea that the classic feed-forward functional structure linking

sensor, processing algorithm and decision is perhaps not the most fruitful on the conceptual plan or the most technologically efficient way to address perception.

As for perception — as we have seen previously — and with a still greater degree of variability, the term system may refer to different ideas, is used in many knowledge and technical domains (engineering, biology, politics, economics, philosophy etc.), and may characterize either a concrete or an abstract object. For our use case — perception — we can define it, minimally, as:

System A circumscribed set of interrelated and interdependent elements that collectively and dynamically contribute to the achievement of a goal.

Several features of a system can be deduced from or added to this definition:

- Elements: a system is compound, its components playing a determined role;
- Interdependence: the existence of the elements depends on the others, making the system a whole *that is more than the sum of its parts*, i.e. a structure;
- Interrelation: the elements interact with each others, i.e. receive or emit information or actions;
- Circumscription: a system has an inside and an outside;
- Goal achievement: the system has an underlying final objective, not necessarily shared with its components;
- Temporality: a system is dynamical, and is therefore dependent on, or simply indexed by, some intrinsic or extrinsic sequential order.

The idea of perceptual system has been proposed extensively in natural perception studies, especially in visual neuroscience. There is an important body of knowledge about the visual system of primates, its neural architecture from the retina to the cortex, and about internal structures of the brain responsible for sensorimotor control. Fig. 2.1 depicts partial functional organization of the visual system, at two different levels: cortical and sub-cortical. What they show is that vision is spontaneously compound and mixes different natures of elements. A notable feature is the loopy structure of dependence between components, very far from a feed-forward processing scheme progressively transforming input sensory data to produce the final output.

This tight and loopy dependency between processing components has inspired several models and implementations of sensory-motor loops, especially in robotics. Fig. 2.2 shows a functional architecture intertwining data processing, gaze and motor control and sensory data acquisition to be hosted by a humanoid robot. The model is designed to account for developmental ability through environment interaction (grasping), but is a good example showing how difficult it is to restrict perception

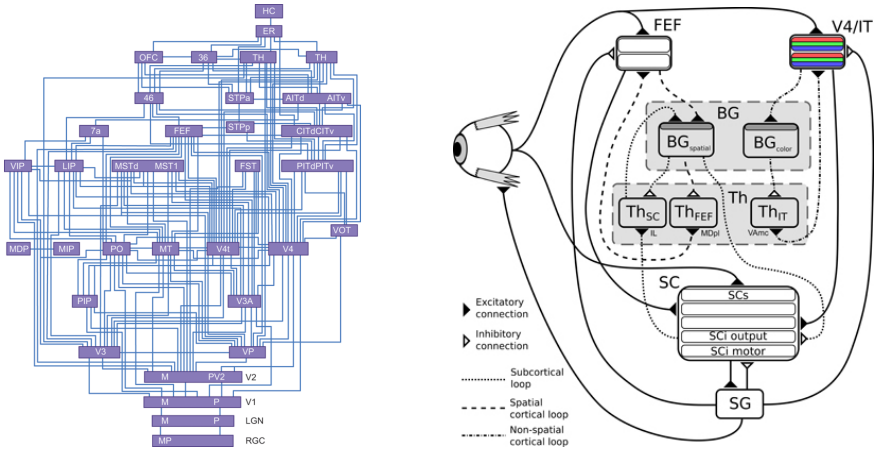


Fig. 2.1: (Left) Visual cortex areas of the macaque monkey. Each box is a separate cortical region that is known to play a role in vision, and the lines represent known pathways between these regions (Felleman and Van, 1991). (Right) Schematic representation of the relationships between cortical and subcortical loops for saccade generation of the primate visual system, involving brain inner structures such as basal ganglia (BG), thalamus (TH), superior colliculus (SC), and cortical areas such as frontal eye field (FEF) and V4/IT (N'Guyen et al., 2014).

to simple feed-forward processing when addressing complex multimodal tasks. This architecture also exemplifies the need of a systemic approach when addressing limited resource management for cognitive tasks and the role of attentional processes. This aspect will be developed in chapter 3.

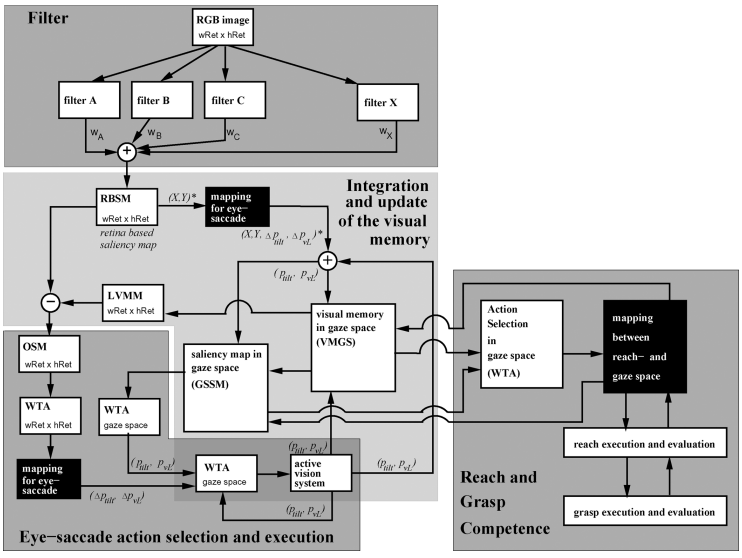


Fig. 2.2: Functional architecture of an object recognition system integrating vision and motor control (grasping, saccade) (Hülse et al., 2010).

The design and development of perceptual systems able to complete several tasks and combining various types of components with assigned roles (data processing, knowledge base, long-term memories, hypothesis management, reasoning, etc.) has a long tradition in artificial intelligence (Draper et al., 1989; Draper et al., 1999). The ambition of those systems is to be rather generic and are more oriented towards

exploiting knowledge representations, i.e. to be *cognitive*, rather than optimizing processing for a specific task. Fig. 2.3 shows a example of a rather recent proposition of this research trend.

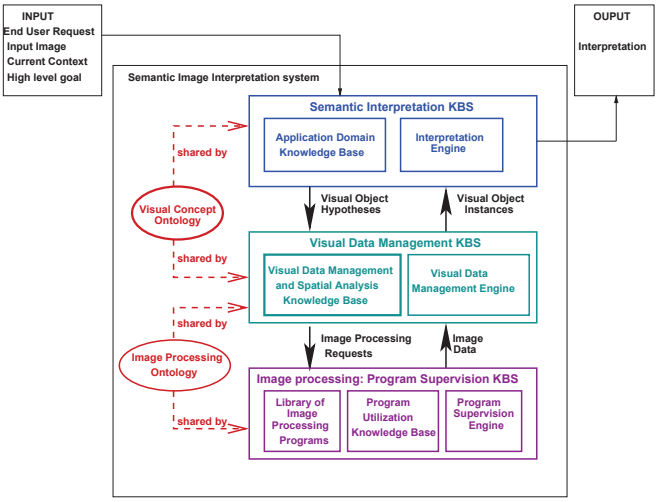


Fig. 2.3: Example of a vision architecture for image interpretation integrating several types of interacting "cognitive" components (Hudelot, 2005).

Finally, the deep network approach, and its capacity to accommodate several modules with specific roles under a unifying formalism, can be seen as the contemporary way to consider perception as a system. The availability of software development frameworks makes the design of interconnected components easy, with the possibility of combining classical feed-forward and recurrent (LSTM, GRU) local architectures in an "end-to-end" learnable framework. Fig. 2.4 shows an example of this type of approach, exploiting modules (e.g. "Visual Question Answering" and "captioning") playing the role of non declarative knowledge bases, similarly to cognitive vision architectures such as those of Fig. 2.3.

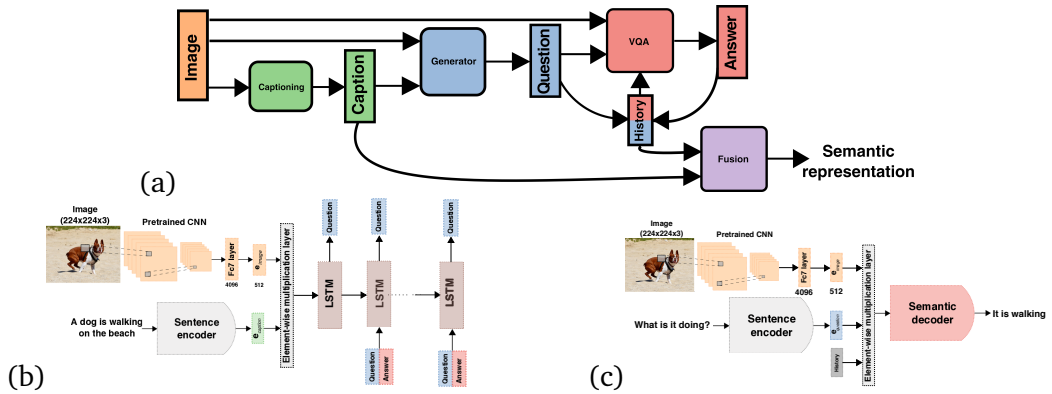


Fig. 2.4: (a) Global architecture of a semantic image encoder exploiting various sources of knowledge exploited as captioning and visual question answering modules. (b) Architecture of the question generator module (blue rounded rectangle). (c) Architecture of the history based VQA module (red rounded rectangle). (Bucher et al., 2018)

What could be the advantages of considering perception as a system, rather than as a feed-forward transducer transforming energy into representation? Three types of justifications can be given:

- it embeds the problem into a more general conceptual framework, allowing better expressiveness;
- it is more in line with models and findings from natural perception, where loopy functional architectures are commonly used to explain behaviors;
- it potentially brings better flexibility and adaptability through dynamical interactions between its components.

The main drawback of a systemic approach is of course that of complexity. We will see in a following section how this issue can be solved by transferring, partly, the complexity of design and operation to autonomy.

Boundaries of perceptual systems (Perception and cognition)

As has been proposed above, the role of perception is to provide either a measure or a sign referring to a property or a content of the environment for a user/recipient: one key feature of perception is that it *externalizes* something. Perception, as a system, has an inside and an outside divided in two categories of externalities: a contingent world and a user.

One consequence of a circumscribed perception is the need to define its functional boundaries. This can be addressed either as a scientific question — what is exactly the role and place of perception — or as an engineering problem — how dispatch tasks between components when designing complex artifacts such as robots or autonomous vehicles.

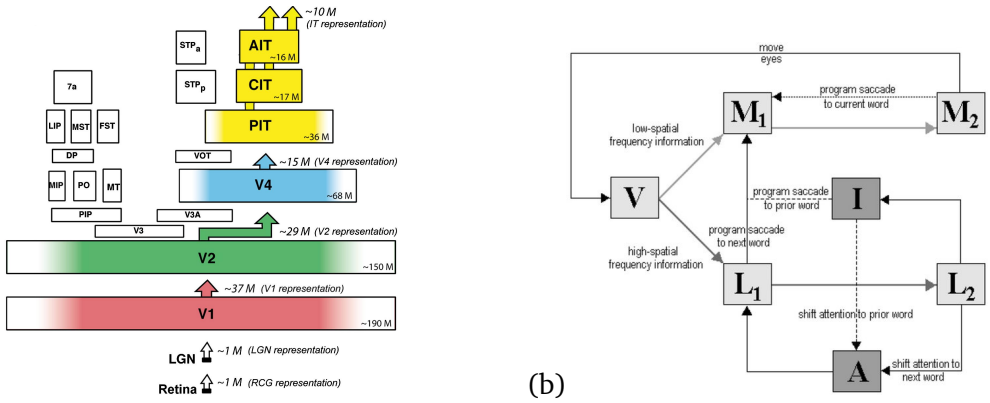


Fig. 2.5: (a) Feed-forward network for object recognition inspired by brain functional architecture (DiCarlo et al., 2012). (b) Diagram of functions involved in the EZ-reading model. V is preattentive visual processing; L1 is "familiarity check"; L2 is "lexical access"; A is "attention shift"; I is "postlexical integration"; M1 "labile saccadic programming"; and M2 "nonlabile saccadic programming". From (Reichle et al., 2012).

Fig. 2.5 shows two diagrams depicted natural vision models. The left diagram shows a feed-forward biologically inspired architecture for pre-attentive visual object recognition, i.e. a recognition not involving any saccade or eye movement (DiCarlo et al., 2012). The second diagram, purely functional, intends to describe the basic modules and actions involved in reading and contains dependent components, with multiple loops and interactions (Reichle et al., 2003).

Both diagrams describe cognitive *perceptual* tasks (object recognition and reading) but with very different options and objectives. The first one tries to isolate an elementary brick, the second one emphasizes the dynamic collaboration of elementary functions to complete a task. These diagrams reveal an underlying difference about what is thought to be the nature of perception.

The feed-forward architecture of Fig. 2.5(a) is in the quest of a model for "core visual recognition", i.e. a function providing through perception an all purpose representation for various tasks. Recent studies of this trend have proposed to interpret contemporary deep image feature networks — often used in the computer vision community as generic features — as plausible neural architectures (Kubilius et al., 2018; Rajalingham et al., 2018).

However, what the diagram of Fig. 2.5 does not show is the existence of "skip" efferent connections from the *early* visual cortex V1 directed towards other higher cortical and subcortical areas (Casagrande and Kaas, 1994) and afferent connections from the IT area (Fig. 2.6), implying that the visual brain is not processing inputs from the retina to interpretation in a strict feed-forward hierarchical way. The brain is the place of multiple bottom-up and top-down interactions, making the isolation in its global neural network of a core and universal perceptual function somehow conventional.

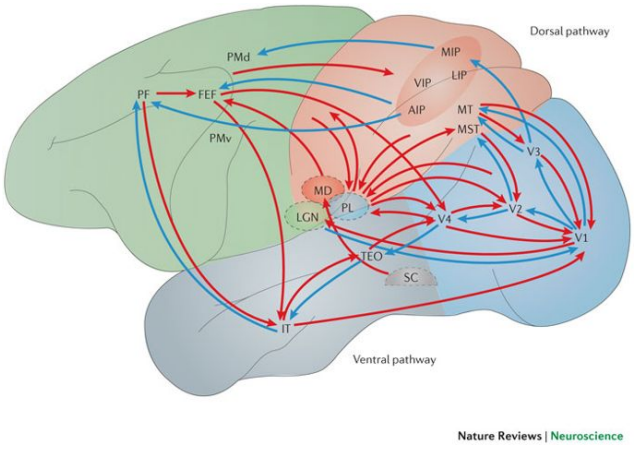


Fig. 2.6: Feedback pathways carrying top-down information to the visual areas. From (Gilbert and Li, 2013).

The contrasting roles of bottom-up and top-down information flows as explication of perceptual phenomena and experience has given rise to decades of discussions. Al-

though there have been numerous studies about the complex influence on perception of context, personal history, multiple sensory modalities, or of cognitive states such as emotion and motivation, the debate is still vivid. The target article published by Firestone and Scholl and its open peer commentaries (Firestone and Scholl, 2016), which disputes the fact that perception may be *penetrated* by cognition (Pylyshyn, 1999), is a recent instance of the current level of controversy.

The top-down vs. bottom-up opposition is also manifest in the usual distinction between low-level and high-level perception, mostly practiced in computer vision (Hildreth and Ullman, 1988), with the idea that the low-level part is automatic, unconditioned, signal processing based, whereas the high-level addresses more semantic tasks and potentially involves multiple contextual and global cognitive state conditioning. Here again, a clear segmentation between a low-level — more perceptual — and a high level — more cognitive — is difficult to keep, especially with the current deep learning approach where models, although mostly feed-forward, are shaped by "cognitive" tasks and objective functions (categorization, object detection, captioning etc.)

Another way to dispute the articulation between a core perception encapsulated as a component of a master cognition is to make perception fundamentally *active*, where attributes such as dynamical, selective, contextual, conditioned, indirect, loopy, top-down influenced, flexible, uncertain, adaptive become relevant. The idea of APES is clearly more comfortable with this way of considering perception.

APES are Autonomous Systems (Agents)

Perception, whether natural or artificial, is a capability with compound structure and which requires complexity management means to become truly usable, for scientific description or for engineering applications. This section examines how adding a degree of autonomy is a way to, partially, get around this question.

Autonomy is a property of a system having a capacity of action: an agent. We will therefore start by studying in what sense perception may fit with the idea of agency, and discuss later how autonomy is a complementary property that impacts the way perception should be addressed. We finally examine the usefulness, both conceptually and in practice, of considering perception as a service.

Perception and agents

Autonomous or intelligent systems or agents have been widely studied in artificial intelligence, and have given rise to specialized scientific communities. Research objectives in this field can be roughly divided in two groups: the first one puts the emphasis on the *agency* property, i.e. the fact that agents can act and modify their environment and inner states: (Poole and Mackworth, 2017) is an example of a

recent textbook of that category which "presents artificial intelligence as the study of the design of intelligent computational agents." The second group puts the accent on the study of interactions and organization between collection of agents either to model behaviors or to solve problems, and is usually referred to as "multi-agents" (Weiss, 2013).

What is the place or role of perceptual issues in those studies? To get an idea of this, Fig. 2.7 depicts schemata found in current textbooks where observation through sensors or stimuli are transformed into *percepts* to generate actions. The nature of this percept, however, is generally not developed in those books.

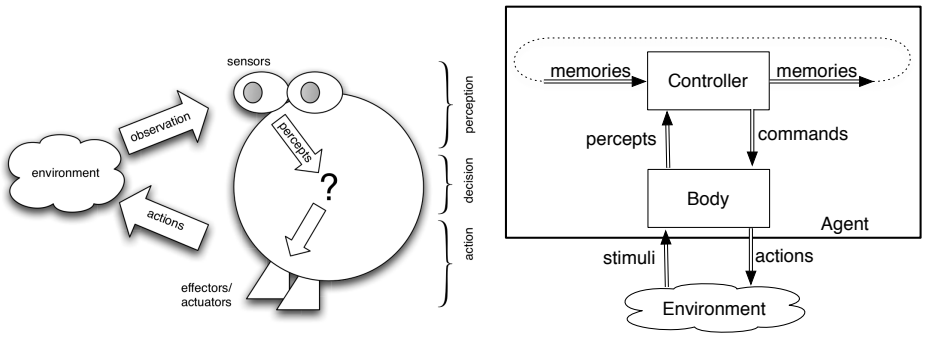


Fig. 2.7: (Left) "An agent in its environment. The agent takes sensory input in the form of percepts from the environment, and produces as output actions that affect it. The interaction is usually an ongoing, non-terminating one." From (Weiss, 2013, chapter 1). (Right) A similar diagram that introduces a memory based controller. From (Poole and Mackworth, 2017, chapter 2).

Another source of information are definitions found in the literature: (Franklin and Graesser, 1996) collect a series of them, and propose to essentially define autonomous agent as "a system situated within and a part of an environment that senses that environment and acts on it, over time, in pursuit of its own agenda and so as to effect what it senses in the future." There are several points that require comments in this definition: it introduces the verb *to sense* — not *to perceive* — conjugated at the active voice, and makes sensing a consequence of action, an *effect*. Senses appear to be at the same time a condition for action and a goal, since the agent is conditioned by what it should sense *next*, in the future.

What is common with the schemata and the definition discussed above is the fact that the agent is considered alone and enters a sensing/action loop with a unique source of contingency: an undifferentiated *environment*. Nothing specific is said about other agents that may directly interact with it and exploit or be influenced by its production. In particular, agents as depicted do not communicate, unless one considers communication as a sensing/action loop — which is not a meaningless option, but needs to be more clearly assessed.

This loneliness is of course not a property of the multi-agent approach where communication between entities — a user being a specific type of agent — is a

central concern (see (Weiss, 2013, part II)). However, the conceptual emphasis in this field is more about the management of the complexity induced by the "multiple" existence of agents than the interaction with a single user.

Perceptual agency

A central question that the idea of APES addresses is the possibility of considering perceptual agency, i.e. to study *agents whose purpose is to perceive* and that do not address perception solely as a resource for action.

Perceptual agency is the fact that perception at some point of its process has to select among several possibilities of action to construct its final state or outcome. It entails time inscription and sequentiality of operation, as for system, but adds *activity* as an essential feature.

The fact that perception for perceptual agents may be a final objective, and not a passive intermediate resource instigates two related questions: what is the nature of "perceptual" actions? what are the characteristics of the environment that receives them?

A first type of action one can think of is motor command. The idea of motion for perception is clearly inseparable from the sense of touch. When considering other senses such as vision of hearing, motor actions may also be involved to direct the head or control the visual gaze (see Fig. 2.5(b) and Fig. 2.2).

The fact that bodily movement are often entangled with perception has given rise to a global account of behavior or cognition as sensory-motor skills coupling in a loopy dependence sensor outputs and actuator commands. The extreme development of this trend is the idea that the world is mainly perceived as a source of action — affordances (Gibson, 2019)— or is constructed, hence perceived, through bodily interactions (O'Regan and Noë, 2001), i.e. the cognition is embodied (Varela et al., 1991) and generates its own world by acting (Gallagher, 2017).

Another remarkable active feature of perception, and probably also of cognition in general, is *attention*. It can be considered as the action of valuing, ordering, filtering or selecting available resources to complete a cognitive task. Attention has been kept for a long time in the neuroscience or psychology domains (Carrasco, 2011; Borji and Itti, 2013), but has been acclimated recently as a common tool in artificial intelligence for deep network design (Vaswani et al., 2017).

Visual attention is manifest through the "overt" phenomena of saccade or pursuit in primate vision. But attention studies also reveal the existence of "covert" phenomena, i.e. mental focus without eye movement (Richard D. Wright, 2008). Chun et al. (Chun et al., 2011) propose another categorization of attentional phenomena, and oppose *external* attention — that refers to the selection and modulation of sensory

information — to *internal* attention — that refers to the selection, modulation, and maintenance of internally generated information, sub-objectives, long-term memory, or working memory.

Attention phenomena show that the repertoire of perceptual actions target other objectives than motor control and may functionally influence "internal" processes. This means also that perception may involve operations and structures with various roles and types, and corresponding actions to control them, making APES a complex functional object by nature (see the diagrams of Fig. 2.2, 2.3 and 2.4 for instance).

Task oriented perception

So far, we have proposed that perception should be endowed with agency and that actions may be either external or internal. In the previous section, we have also questioned the relevance of drawing inalienable boundaries to perception in its relation to a hypothetical cognition that may functionally encapsulate it.

A simple idea to avoid endless discussions to decide who gets the primacy, a free mind or a contingent experience, instantiated by the top-down/bottom-up opposition, is to consider cognitive activity as a dual project of task *specification* — what is to be achieved — and *completion* — the operations actually involved.

The value of perception in this framework, which ontologically puts the notion task at the first place, depends on how and for what purpose its output is used. Perception "alone" is useless and only acquires value when contributing to an identified task.

Task specification is not the responsibility of perception, but perception has to "agree" to participate in the operations needed to complete it and to follow a set of rules and commitments. Conversely, cognition, considered as a task manager, enrolls available resources, including perception, based on a knowledge of what they can achieve.

In this task-oriented view of perception, no actor dictates what the other has to do without shared acknowledgement: cognition uses perception production to complete its objective but has to take into consideration available perceptual capacity; perception organizes its production so as to be useful to the task. We will come back later on this co-specification by introducing the idea of *service* in a further section.

All cognitive tasks do not necessarily involve perception. Remembering, solving a mathematical problem, judging, use skills that do not require the existence of perceptual inputs to operate. Tasks such as describing a scene, reading, navigating, grasping, singing etc., however, make use of sensory inputs somewhere in their process. Scene description produces from visual signal semantic representations, usually symbols, referencing its content. Reading requires visual sensing and gaze control (see Fig. 2.5). Navigating may exploit sensors to estimate a position, i.e. a measure, in order to build appropriate motor commands. Grasping uses visual

servoing to reach its target with good precision. Singing may use audio feedback to adjust pitch and level. In all those tasks, the role of perception is to provide outputs dedicated to their achievement.

Those examples suggest a rather general definition of a

Perceptual task : A goal-oriented activity that depends on measures or signs produced by perception from sensory inputs.

In this document, we will only be interested in describing how perceptual tasks are implemented as APES.

Binding perception to task implies multiplicity: there are, in principle, as many categories of perception as varieties of tasks: there is no such thing as a generic perception that provides versatile and universal representations of the world, but several instances of dedicated sign or measure producers that contribute *specifically* to the current task.

However the quest for common architecture or features in perception is an interesting scientific and technical question: it should result from discovering what can be shared between tasks, and not by positing a core architecture. A certain level of versatility can be a consequence of task multiplicity, as is proposed for instance in (Kokkinos, 2017; Doersch and Zisserman, 2017; Zamir et al., 2018), the problem being addressed is the existence of a structure organizing visual tasks and able to transfer latent features between them (Fig. 2.8). We will see in the following chapter what are the promises and the difficulties of conceiving such a project of joint control of multiple perceptual tasks.

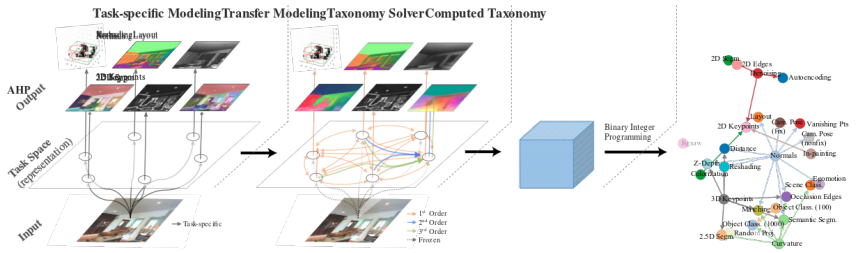


Fig. 2.8: Structure discovery between visual tasks (Zamir et al., 2018).

Adaptivity

As we have seen in the previous section, a desirable feature of APES's is their capacity of offering a repertoire of perceptual tasks for several recipients, i.e. their potential versatility. Addressing globally a task-user multiplicity is an over complex problem and motivates the development of *adaptation* capacities in APES's that will compensate for incomplete models. Roughly speaking, since it is impossible to describe exhaustively all the circumstances APES's will have to deal with, we are

forced to relax imperative modeling or programming towards more declarative way to express their expected behavior.

Considering adaptivity as "the capacity that certain systems possess to modify themselves in order to adjust to changes in the environment" (Barandiaran and Moreno, 2008), a first question is to identify what is the environment of an APES, or what is likely to change in it. Various aspects can be considered.

Adaptation to the situation, i.e. to the entities that signs or measures refer to. This is the usual way to consider run-time adaptivity where, for a given task, the perceptual system selects and combines its resources to adjust to the complexity of the outside world. For instance, image features or processing logic may differ in night or day vision; another type of adaptation is active vision that sequentially selects informative parts in the field of view (see Fig. 2.2 page 13).

Adaptation to the task(s), i.e. to the nature of expected output signs and measures. The structure of operations and resources depend on the perceptual tasks and on the system performance requirements. Several parts of the perceptual system can be generically shared in case of multiple tasks, e.g. deep image features as it is now customary in many computer vision chains.

Adaptation to the operational domain, i.e. to the *variety* of sensory inputs and tasks. An APES structure may globally adjust to the potential complexity or variability of the perceived world: for instance, image features and processing chains are different when detecting furniture in indoor scenes or when segmenting buildings in aerial images.

Adaptation to the user(s), i.e. to the nature of the potentially multiple agents that will exploit perceptual production. An APES may have to prioritize its resources according to the status of the recipient and to the nature of the perceptual output. For instance when the result of certain tasks are critical to the safety of a global entity hosting an APES as one of its subsystems, or when one of the tasks is to alert about some possible danger.

Those four different types of adaptivity do not have the same impact on APES design. The first one (adaption to the situation) is related to the systemic nature of perception as has been discussed previously (page 11) and conditions the dynamics of information flows. The other three are more global and have an influence on the functional structure of the system, i.e. the nature and relation between the various states, components, processes, etc.

Ideally, to be truly versatile, an APES should be able to adapt along those four directions: specifying and designing the rules or principles governing the internal

decisions and processes of this kind of system is a very challenging problem if addressed in a direct way. The next section discusses the role of autonomy as a necessary condition for efficient adaptivity.

Autonomy

One motivation for introducing the idea of APES is to endow perception with some degree of autonomy, and examine what this property implies. A first action is to clarify what is meant by autonomy.

As a preliminary precaution, we do not consider here the ontological problem of the self or identity that can result from the "operational closure" at the root of autopoiesis (Meincke, 2018; Barandiaran, 2017) or be a condition of autonomy. In other words, we do not question the existence of the entity — a perceptual agent — nor the boundaries with respect to its outside; we only try to address what it means to be autonomous.

The first property that comes to mind when speaking of autonomy is *independence*.

It arises rather naturally as a key feature of autonomy in moral or political philosophy. Autonomy is either considered as the right or condition of free self-government independently from external control or influence, or for instance in Kantian moral philosophy, as the "capacity to impose the (putatively objective) moral law on oneself" (Christman, 2018) and act in accordance with it. Autonomy is connected to free will, as a necessary condition, but adds to the capacity of making agent's choices of action a higher possibility of designing own objectives or motives. Another related concept is that of (moral) responsibility, as a dual or counter part of freedom of action: one way to answer to the problem is to ascribe responsibility to an agent only if the production of his/her actions can actually or potentially be demonstrated through explicit and verifiable statements identifying the underlying rules governing action.

All those questions of free-will, self-government, responsibility, statements, which are critical to define autonomy from a moral or legal point of view, are still debated and controversial. Whether they can be adapted and applied to artificial systems such as APES's requires detailed investigation and will not be addressed further in this document.

From a more focused artificial system perspective, thus, the issue of independence can be started by identifying its attributes: independence from what or who? For APES's, dependence can be reduced to a relation with two kinds of objects: the environment in which the perceptual agent lives, and other agents which may take the role of either users or prescribers of perceptual outputs.

The environment is the main source of contingency. A perceptual system cannot be independent from the environment since its principle is precisely to generate measures or signs that refer to it. However, freedom or self-government can be invoked in the way this reference is built and on what it refers to. An autonomous perceptual system has some flexibility on the form it generates a scene description — as a text, a mathematical object representing geometric configuration or radiative features, an image etc. It may also choose the elements to describe (object location or attributes, actions, events etc.) and will organize its resources accordingly. Having the freedom of choosing the nature of a description can however also be understood more negatively as a consequence of the limited capacity of APES's: choosing is selecting what can be achieved.

The relation to other agents however is potentially more constrained, in the sense that what makes the value of APES's outputs, is what a user or a sign recipient — the *interpretant* (cf. page 10) — will make of or expect from it. Somehow, an Autonomous PErceptual System must be aware of the recipient-user's objectives, at least partially, to guide its activity, making its behavior partly conditioned by these external constraints.

The global complex gathering APES and recipient-user involves two objective functions: the first one is hosted by the APES and aims at a building measures or signs that are faithful to the environment and valuable to the user, the second one represents the recipient-user own goals that makes use of APES production. The two ideally collaborate but may also compete: conflicts may be resolved by imposing a hierarchical precedence between the user and the perceptual system, restricting APES full independence.

However, an exclusive hierarchical relation with a single user which prescribes the current perceptual task may be inefficient and even dangerous for a global system, perception being the unique source of environmental information.

A first obstacle to exclusivity is the expected versatility of APES's that aims at providing outputs to several users that may have contradictory goals. A second difficulty that could make a hierarchical relation inefficient and even harmful is *inattentional blindness* (Mack and Rock, 1998) that makes a perceptual system univocally engaged in a given task for a given user, becoming blind to unexpected events. This kind of phenomena leads to problem of *attentional tunneling* (Régis et al., 2014), i.e. "the allocation of attention to a particular channel of information, diagnostic hypothesis or task goal, for a duration that is longer than optimal, given the expected cost of neglecting events on other channels, failing to consider other hypotheses, or failing to perform other tasks" (Wickens and Alexander, 2009). This behavior prevents a perceptual system from generating alerts that may inform the system of dangerous events, for instance, and from asking for revised objectives.

SAE level	Name	Narrative Definition	Execution of Steering and Acceleration/ Deceleration	*Monitoring* of Driving Environment	Fallback Performance of *Dynamic Driving Task*	System Capability *(Driving Modes)*
***Human driver* monitors the driving environment**						
0	No Automation	the full-time performance by the *human driver* of all aspects of the *dynamic driving task*, even when enhanced by warning or intervention systems	Human driver	Human driver	Human driver	n/a
1	Driver Assistance	the *driving mode*-specific execution by a driver assistance system of either steering or acceleration/deceleration using information about the driving environment and with the expectation that the *human driver* perform all remaining aspects of the *dynamic driving task*	Human driver and system	Human driver	Human driver	Some driving modes
2	Partial Automation	the *driving mode*-specific execution by one or more driver assistance systems of both steering and acceleration/ deceleration using information about the driving environment and with the expectation that the *human driver* perform all remaining aspects of the *dynamic driving task*	**System**	Human driver	Human driver	Some driving modes
***Automated driving system* ("system") monitors the driving environment**						
3	Conditional Automation	the *driving mode*-specific performance by an *automated driving system* of all aspects of the dynamic driving task with the expectation that the *human driver* will respond appropriately to a *request to intervene*	System	**System**	Human driver	Some driving modes
4	High Automation	the *driving mode*-specific performance by an automated driving system of all aspects of the *dynamic driving task*, even if a *human driver* does not respond appropriately to a *request to intervene*	System	System	**System**	Some driving modes
5	Full Automation	the full-time performance by an *automated driving system* of all aspects of the *dynamic driving task* under all roadway and environmental conditions that can be managed by a *human driver*	System	System	System	**All driving modes**

Fig. 2.9: Levels of driving automation as proposed by the Society of Automotive Engineers (committee, 2018).

The status of the relation between an artificial system and a potential user is the core or the proposed hierarchy that rules the different levels proposed for automated driving applications (Fig. 2.9). The levels distinguish two groups according to who monitors the driving environment: human or automated system. It also introduces the idea of "request to intervene" that instantiates various precedence levels between user and automated system.

What should be retained from the short discussion above is that autonomy does not mean that an APES should not have relationships with other agents or users, but that the nature of independence, seen as a condition to autonomy, is determined by the *type* of relationship they have. We will define in a further section what are those relationships by discussing the idea perception as a *service*.

Learning and autonomy

One possibility to avoid the complexity of explicitly modeling adaptive systems is to implement this activity as a *learning* phase, i.e. let actual experience or practice shape their structure and values. This is one of the major trends in contemporary perceptual and artificial intelligence systems, relying on the fact that a data driven approach is able to compensate — partly — for the complexity or deficiency of formal modeling.

Learning can be thought as a way to adapt a system to the three off-line types of adaptation: to the operational domain, to the tasks and to the users. But how does it fit in a system claiming a certain level of autonomy?

A first useful distinction to make is to separate autonomy during operation, or during development. Operational autonomy refers to the usual property when speaking of autonomous agents (see previous section page 17). Developmental autonomy introduces the idea that the system has some initiative in designing its own capacity, and can evolve over time: this is where learning can take place.

The classical and mostly used type of learning is data-driven *supervision*: the way it is formulated through an annotated dataset completely defines the task (input and output) and the domain, since data are hypothesized to sample the true underlying distribution. User requirements are usually expressed in the form of performance objectives (error rate, computational time, memory usage, etc.) In this framework, autonomy can be found in the way the algorithm instantiating the prediction, the optimization scheme and the data management strategy are chosen. This is a rather weak type of developmental autonomy, where choices are not ruled by objectives provided by the APES itself but by external prescribers.

Several varieties of learning belong to the category of supervised learning: classification, regression, imitation learning. Reinforcement learning is a weaker way to specify a task — it does not require the exact output, only the cost of making errors — but also requires an external prescriber to tell if there have been an error. The standard supervised scheme may be complemented by other strategies to exploit other sources of data: domain adaptation (Patel et al., 2015; Csurka, 2017; Wang and Deng, 2018), transfer learning (Pan and Yang, 2009; Weiss et al., 2016; Tan et al., 2018) or self-supervised learning (Jing and Tian, 2019) etc. can be used to adapt to larger domains for instance, but all aim at fulfilling the same task.

However, APES's may need a higher level of autonomy especially if versatility is one of their objectives. An important dimension of autonomy, therefore, is self-development, i.e. the capacity to improve its skills, in quality and quantity, by dynamically managing all kinds of available resources. Several learning schemes have been proposed to address this long-run problem — few shot, zero-shot, continual, meta learning — and exploit various types of information or data to add new perceptual capabilities (new class, domain, question etc.). One difficulty when increasing the repertoire of skills is to achieve it without damaging the old ones, i.e. with no regression, a property which requires a certain degree of autonomy for an APES to organize its resources.

We will examine in the next chapter the central role of learning in contemporary perceptual system design.

Perception as a service

As we have seen, one specificity of APES is to consider jointly the perceptual system capacity and the user needs. Introducing an idea of autonomy in perception requires the description in more details of the nature of their relationship.

A complementary way to consider the relationship user-APES is by introducing the idea of *delegation*, i.e. "the act of empowering to act for another". The APES acts on behalf of the user to produce perceptual outputs, and becomes *responsible* of their quality. Conversely, delegation is only possible if the user has *confidence* in the APES skills to fulfill its perceptual or informational needs, i.e. if the APES is considered reliable. We will examine in this section how considering perception as a service is one way to formalize user and APES interaction.

The idea of service has been extensively applied in several computer science applications. One of the most notable formal models is called Service Oriented Architecture (SOA) [12]: it defines the roles of service providers and consumers, and the means of establishing their interactions, trough the shared publication of a catalog of available services and the definition of contracting protocols.

If we pursue the idea of service, the user plays the role of a *client* issuing requests to the *server*, i.e. the APES, according to a ruled *protocol*. Before actually operating the service, a *contract* specifying the service must be set and agreed between the two parties.

When establishing a service, three phases should be considered:

Contracting In this preliminary phase, the client and server agree on what is the expected output and from what material, its quality, but also how should abnormal situations be detected and handled.

Operation This is where the actual perception takes place from sensory inputs. It may also allow some kind of monitoring from the client in order to verify that the process works as expected, depending on what has been agreed during the contracting phase.

Delivery When the server considers that the queried perceptual outputs are available according to what has been contracted, it delivers it to the client who may accept or reject it, and potentially ask for adjustments.

The contracting phase is the most critical one since it defines what can be achieved for the user-client by the APES-server. It relies on prediction of what could happen, and ideally anticipates every critical situation.

[12] https://publications.opengroup.org/standards/soa

The difficulty of drafting an agreed contract between APES and user — which can be a complex and dynamic procedure — is to match two points of view. Tables 2.1 and 2.2 states the various aspects that the contract should address to guarantee a clear and reliable relation between the two parties.

Tab. 2.1: Questions to be answered from server side during contracting phase

• Who issues the query?
• Who is the recipient of the answer?
• What is the nature of the information requested?
• With what quality requirement? How soon? At what cost?
• Can I predict the quality of my output given the context?
• How can I prove that I will be able to answer the query?
• Can I afford to answer it alone?
• If no, what other quality / deadline can I suggest?
• How to agree with the client on the quality / deadline?
• Do I need external services to answer the query?
• Is there conflict with another query being processed?
• How learn from the client what he wants?
• How justify that my answer is efficiently/faithfully produced?
• How express the uncertainty or confidence about results provided?

Basically, what the server has to address is the possibility of providing a reliable and useful output to the client, using its own resources. An important dimension is the capacity of the sever to express the degree of contract fulfillment (justification, uncertainty representation, etc.) so as to build a confident relationship with the client.

From the client side, the main challenge is to define in a precise way what it expects from the service, to agree on what it can potentially offer, and to commit to accept it. What makes a system autonomous is therefore not unconstrained freedom but a capacity of making choices according to its own principles or rules. When formalizing perception as a service, self-governance is achieved if the relationships that tie the various agents can be *explicitly* defined, declared, shared and revoked if needed, i.e. if a contract can be *explicitly* established between the parties and reliably operated. Intelligibility of contracting and operating is therefore a required property of perceptual system to become truly autonomous. We will discuss in more details this issue in chapter 4.

Tab. 2.2: Questions to be answered from client side during contracting phase

- What level of quality / delay can I ask?
- How express unambiguously the expected performance quality?
- What level of delegation can I authorize? What should I monitor?
- What recommendations can I make about the way the query should be processed?
- What useful information can I send to help and improve query processing?
- What quality of service can I expect?
- Can I get a particular quality of service?
- If the answer is not of sufficient quality, is there a way to claim?
- Can I teach the server what I need?
- What action can I make if I think something goes wrong when processing the query?

2.2 Natural APES

The disparity of definitions presented in the previous section (page 6) reveals that the nature and the role of perception has a long history of research and debates in philosophy (Matthen, 2015) psychology (Goldstein, 2010) and neuroscience (Banich and Compton, 2018).

In this section, we present – very broadly – several issues that may shed light on several aspects of APES, with the conviction that natural science or philosophy have something to tell to engineering. We do not however assume that the underlying project of APES studies is to *Build machines that learn and think like people* (Lake et al., 2017) but, when it seems profitable, acclimate and exploit findings from cognitive science that could inspire more efficient design. It may also happen that the formalism and demonstration required to fulfill engineering objectives give some insight and partial justification of natural science hypotheses.

Cognitive science will be considered as the conjunction of three areas of study, although the frontiers between them may not be that strict:

Psychology provides validated hypotheses and models on the nature of perception from a behavioral perspective. It is often divided in several subfields of study such as memory, cognitive disorders, illusions, attention, visual search, etc.

Neuroscience studies the neuronal substratum of perception in the brain, by identifying the anatomical or functional structures and their relations at various levels of analysis.

Philosophy asks fundamental questions on the content of perception, if any, its purpose, its level of truth. It is discussed for instance whether perception should be considered only as a an openness to or as an awareness of the world, as a direct or indirect source of representation of that world, as a teleological or as generic activity, as a source of knowledge or as an assessment of the behavioral possibilities a stimulus affords.

We will briefly identifies several issues and results in those fields that may be of interest for the study and design of APES's.

Neural organization

Formal neural networks, which are now unavoidable in artificial intelligence, have been originally inspired by biological models: they mostly exploit variations of a linear combination+activation pattern ($x_i(t+1) = \phi\left[\sum_j w_{ij}x_j(t) - w_0\right]$) as proposed by McCulloch and Pitts (McCulloch and Pitts, 1943) to model the activity of a single neuron when stimulated by other neural inputs. The simplicity of this model has allowed the usage of mathematical tools to control complex architectures, and study the computational capacity of networks.

However, this model is also well known to be a caricature of natural neuron behavior: it may lead to misunderstanding of brain behavior, and perhaps may hide interesting findings that could inspire new efficient formal models.

In this section, I will briefly point out several issues regarding possible inspiration of neuroscience for APES design.

Brain structures

A first feature that is disregarded as a source of inspiration for perceptual system models is the actual organization of brain structures. Most artificial neural networks proposed in the artificial intelligence literature have a rather homogeneous structure and follow a feed-forward layer based pattern, sometimes combined with skip-connections between layers.

This is to be contrasted with brain organization, where subcortical modules are combined with specific cortical areas in a more loopy architecture (see Fig. 2.6) or when the perceptual system is extended to involve eye motor coordination (Fig. 2.10 and 2.1).

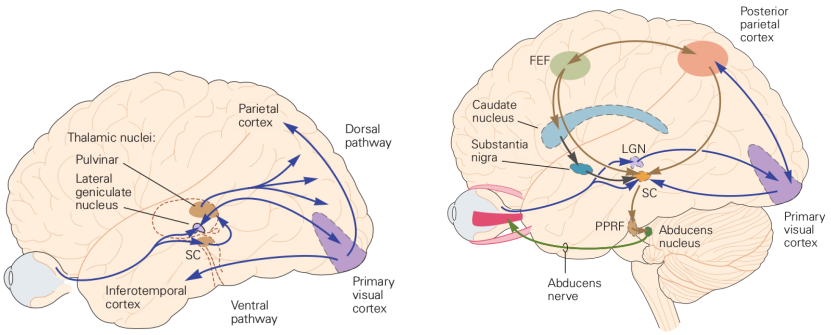


Fig. 2.10: Cortical and subcortical structures involved in visual processing (left) and eye control (right). From (Kandel et al., 2013, chapter 25).

A difficulty however when trying to get inspiration from brain architecture is the precise knowledge about the role of the various structural components: the fact that several regions exchange neural activity does not tell how and for what purpose they do so. Broad functional areas have been identified and have been used to propose an account of how the brain connects to its environment to create a mental world, using for instance imagery (positron emission tomography (PET) and functional magnetic resonance imaging (fMRI)) to justify hypotheses (Frith, 2007; Gazzaniga and Ivry, 2013). The precise connectivity and dynamics of neural activity is however largely unknown.

Sensory system

Natural and artificial sensors may be rather different: the physical principles of detection, of course, but also the form of their outputs. For instance, an eye is not producing images, i.e. arrays of measures regularly indexed by a location in the sensor plane, but streams of activity transmitted through action potentials by the million of fibers of the optic nerve to the brain.

Massive parallel filtering occurs at very early stages of perception, and encodes physical phenomena in behaviorally relevant form (Wässle, 2004). In vision, the retina adapts to illumination and motion changes, for instance, and is able to compress by a factor of 100 the 10^8 rods and cones that transform light in electrical potential.

An important difference of the retina compared to cameras is the spatial variation of photo-receptor density (Fig. 2.11), allowing various types of sensitivity: the central region, the fovea, contains a higher density of cones that have responses that depend on light wavelength, and a peripheral region with a larger density of rods, that have large wavelength response and are much more sensitive to light.

Another key feature of the retina is the amount of computation implemented close to the photo-receptors, and that “whereas the conventional wisdom treats the eye as a simple prefilter for visual images, it now appears that the retina solves a diverse set of

specific tasks and provides the results explicitly to downstream brain areas"(Gollisch and Meister, 2010). The fact that the functional bricks of the retina and the brain are all neurons is another arguments against drawing a fixed boundary between pure perception and cognition.

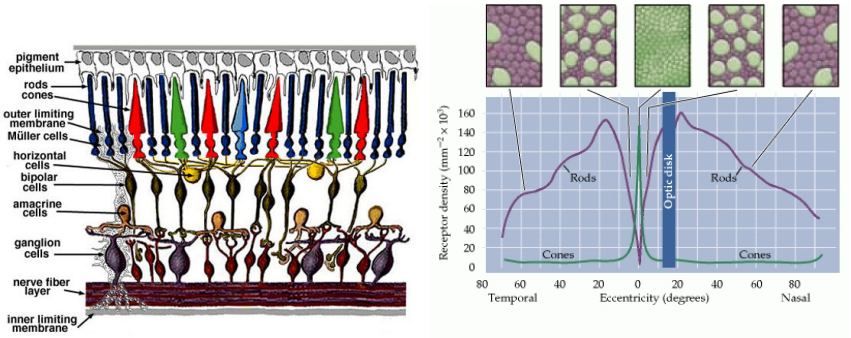


Fig. 2.11: Anatomy of the retina, with the five types of cells: rod, cone, ganglion, bipolar and amacrin. Left: The various cells that shape input light in electrical signals to the brain (photo-receptors and ganglion cells). Right: The varying densities of rods and cones in the retina.

As a recent review article on the functional organization of the retina states (Baden et al., 2018), "today the retina is amongst the best understood complex neuronal tissues of the vertebrate brain." The classical account of retina circuits have revealed that " neurons specifically responding a certain visual feature are usually complemented by a set of neurons that are suppressed by the very same feature" (Baden et al., 2018), and form ON and OFF cell populations of ganglion cells, the neurons that connect the retina to the brain.

However, several aspects are still unclear, typically what is "the neural code of the retina" (Meister and Berry, 1999). In 2012, (Masland, 2012) estimated that "at least half of the encodings sent to the brain (ganglion cell response selectivities) remain to be discovered". (Wienbar and Schwartz, 2018) propose a recent review about the structure of ganglion cell receptive fields, i.e. the spatial locations of the field of view that influence cell activity, and demonstrate their huge static and dynamic variety.

Another unanswered question is the spatio-temporal form of the information conveyed to the brain. For instance, several studies argue that the synchronized activities of output ganglion cells contain information that complement individual cell trains of spikes (Shlens et al., 2008). (Gollisch and Meister, 2008) "report that certain retinal ganglion cells encode the spatial structure of a briefly presented image in the relative timing of their first spikes."

The complexity of natural sensors, even the most well known such as the mammalian retina, has not been translated yet into practical and efficient formal models for applications. Taking a kind of opposite inspiration, several studies have tried to show that current artificial intelligence approaches are in some way competitive with natural sensors: (McIntosh et al., 2016) use CNN to model retinal response to

natural image sequence, (Parthasarathy et al., 2017) propose a deep network model for decoding natural images from the spiking activity of large populations of retinal ganglion cells. (Lindsay, 2018) discusses in what sense deep neural networks may be relevant natural vision models.

Neuron types and connectivity

The basic structural elements in the brain are neurons and glial cells, this last category having no functional role for cognition. Here we only address cortical organization, with the hypothesis that the cortex is the brain structure responsible for cognition — a disputable hypothesis.

We mainly discuss the difference between artificial and biological neural networks structures.

A first difference is the functional nature of neurons. In the traditional McCullough & Pitts model, efferent synapses (i.e. weights) may be either excitatory (i.e. positive) or inhibitory (i.e. negative). This is in contradiction with biological findings for which neurons are either globally excitatory or inhibitory due to the nature of the synaptic neurotransmitters. Excitatory synapses are also much more frequent that the inhibitory ones (20 to 25 %) (DeFelipe et al., 2002). This high discrepancy between excitatory and inhibitory neuron in number, structure and distribution in the cortex suggests different roles: informational association for excitatory neurons, and regulation for inhibitory ones.

The fine connectivity structure and range of individual neurons is also much more diverse than what is commonly used in DNN's. Fig. 2.13 shows several types of ganglion cells from the retina (its outputs), with various input dendritic tree shapes and locations. The cortex contains also many different families of neurons, distinguishing top-level excitatory and inter-neurons. Another widespread feature in the brain but not in DNN's is the presence of inner layer lateral connections (Thomson and Lamy, 2007; Harris and Mrsic-Flogel, 2013) (Fig. 2.12), although recurrent network models such as LSTM can be interpreted as instantiating lateral connections on its inner state. However, "despite extensive research, we still have only a rudimentary understanding of the diverse classes of cortical excitatory and inhibitory neurons"(Harris and Mrsic-Flogel, 2013).

A second difference is the nature of connectivity. Most contemporary DNN's are structured as chains of feed-forward layers (ResNet ian extreme case that may contain until 151 layers (He et al., 2016)), with several exceptions such as Dense Net (Huang et al., 2017a) or U-Net (Ronneberger et al., 2015) that contain skip connections between lower and higher levels, combining mostly convolution and fully connected layers.

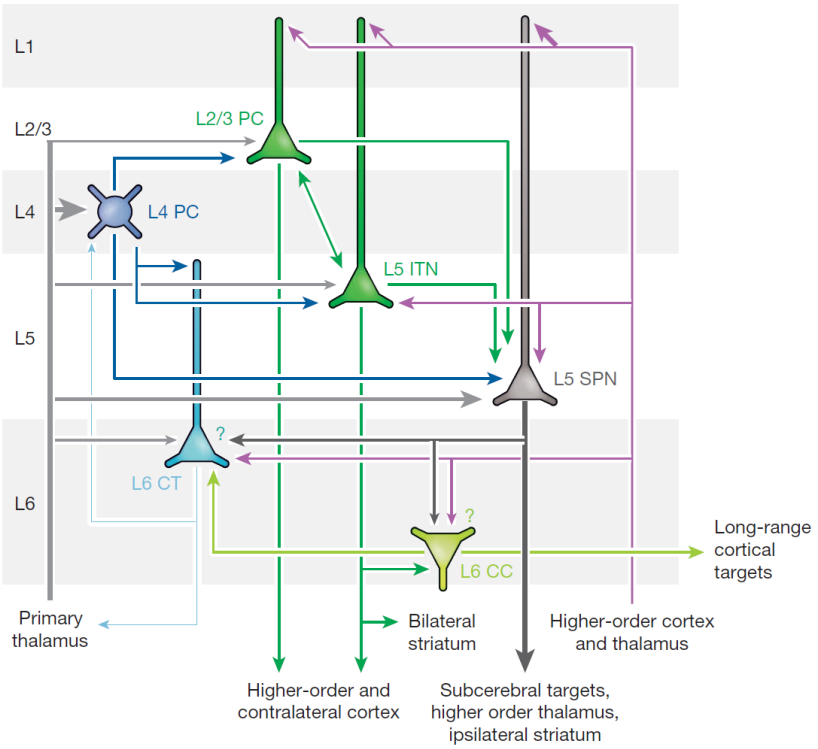


Fig. 2.12: Current understanding of the connectivity between the major principal cell classes of the sensory cortex. Line thickness represents the strength of a pathway; question marks indicate connections that appear likely but have not yet been directly demonstrated. From (Harris and Mrsic-Flogel, 2013).

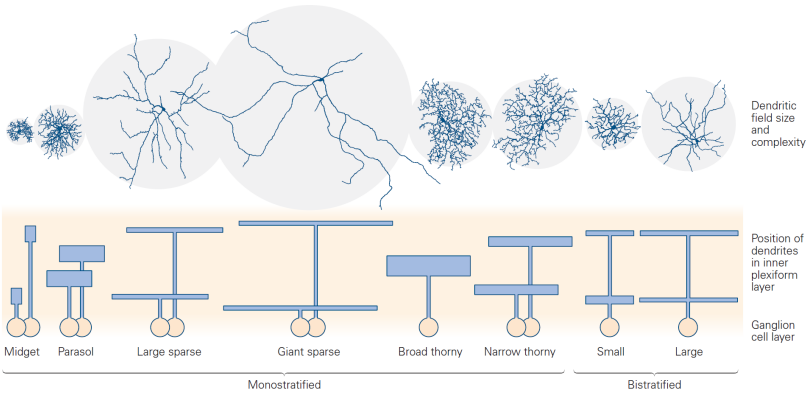


Fig. 2.13: The retina contains more than 13 types of ganglion cells on the basis of their dendritic shape and depth of position in the retina. From (Kandel et al., 2013, chapter 4).

By contrast, the diameter of the cortical graph, i.e. the maximal graph distance between two neurons, is between 2 to 3 [13], meaning that two neurons randomly picked can be connected by at most 1 or 2 intermediate relays. At a larger scale of analysis (cortical areas), connectivity is also structurally and functionally denser than in most artificial DNN's (Bullmore and Sporns, 2009) (see also Fig.2.1).

Neural information

How the brain encodes information to perceive and to act in the world is still not clearly understood. The unified nature of brain constituents, neurons and glial cells, that ensure an amazing continuity between sensors, actuators and cognitive functions hosted in the cortex makes logical, structural and functional analysis difficult.

What the brain really encodes, and for what purpose is still highly debated. What can be stated without too much doubt is that information is distributed in the whole network, that certain subnetworks or areas play specific functional roles, and that dynamical electrical and chemical phenomena are jointly used as support for transmitting or holding information.

Basically, four different feature candidates have been investigated to devise a neural code (Wikipedia, 2019): average firing rate, temporal, population and sparse coding.

The firing rate is the most usual coding scheme proposed in formal models: the density of spikes by time unit measures the intensity level of other afferent neuron activity, and the global input activity is the sum of all firing rates weighted by synaptic efficacy. (Rolls and Treves, 2011) argue after a quantitative information theoretic analyses of neural encoding, particularly in the primate visual, olfactory, taste, hippocampal, and orbito frontal cortex, that most of the information is encoded by the firing rates, and that "a little additional information is available in temporal encoding involving stimulus-dependent synchronization of different neurons, or the timing of spikes within the spike train of a single neuron".

However, other studies argue that a lot of information is encoded in temporal features of trans of spikes: (Panzeri et al., 2010) state that "temporal multiplexing could be a key strategy used by the brain to form an information-rich and stable representation of the environment". (Pillow et al., 2008) propose that time-to-first-spike is the information carrier. (Rullen and Thorpe, 2001) argue that the temporal structure of

[13] This can be computed approximately from the degree of the graph and its size. The total number of synapses in human brain is 1.64×10^{14}, with an average number of synapses per neuron of 6.93×10^3 (Tang et al., 2001), with variations for 1 to 6 between cortical layers I and IV for instance (DeFelipe et al., 2002). The number of cortical neurons is estimated to 16.34×10^9 neurons, and 86.06×10^9 in the whole brain (Azevedo et al., 2009). This gives an approximate diameter = $\log(\sharp\text{neurons})/\log(\sharp\text{synapses per neuron})$ of 2.65.

the spike train is required for fast transmission of detected events from the retina to the cortex.

Different parts of the brain may encode different features, and for various objectives. In this spirit, (Kumar et al., 2010) try to reconcile proponents of firing rate vs temporal coding and propose that "rate and synchrony propagation represent, in fact, two extremes of a 'continuum' defined by the parameters of the feedforward architecture. A particular class of networks may be more suitable for rate propagation, but it can be systematically altered to a network that preferentially propagates synchrony".

Population coding has been a longstanding hypothesis but until recently rather hard to experimentally study due to the difficulty of recording multiple neuron joint activity. Most of the work have been theoretical (Cohen and Kohn, 2011) (Brette, 2012) and with rather small size models. (Gardella et al., 2019) in a recent review of neural population coding and segments models synchronous vs. temporal correlation based approaches.

A fundamental question however is to clarify what the "neural code" refers to. A first proposed segmentation has been to make the hypothetical neuron code as an intermediate correlate of stimulus and behavior (Johnson, 2000). This purely stimulus/response behavioral analysis however makes difficult the understanding of latent mental states which may not directly depend on inputs (Northoff, 2013). More global modulatory phenomena may also encode other type of information, such as the effect of brain stem on attention and arousal (Kandel et al., 2013, Chapter 25).

As a final statement of this section about a possible collaboration between formal mathematical models and neuroscience, "although decades have passed since perceptrons and associative memory networks were invented, it is still unclear how well these models explain visual perception and the storage and recall of memories" (Kandel et al., 2013, p. 1599). The story is clearly not ended: (Kietzmann et al., 2018) present a recent review of how deep network models have been used in computational neuroscience. (Barrett et al., 2019) explore opportunities for synergy between computational neuroscience and artificial intelligence. Regarding APES design, natural neural networks remain an interesting, and under exploited, source of inspiration.

Predictive coding

Principles

Predictive coding (Rao and Ballard, 1999), also often referred under the expressions *Bayesian brain* (Knill and Pouget, 2004; Friston, 2012; Seth, 2015) or *Free-energy principle* (Friston, 2010), is a rather recent field of thought proposing a computational unifying view of behavioral and neural phenomena.

The idea of predictive coding originates in neuroscience (Feldman and Friston, 2010; Bastos et al., 2012), and opposes "classical theories of sensory processing [that] view the brain as a passive, stimulus-driven device" and "emphasize the constructive nature of perception, viewing it as an active and highly selective process [for which] there is ample evidence that the processing of stimuli is controlled by top-down influences that strongly shape the intrinsic dynamics of thalamocortical networks and constantly create predictions about forthcoming sensory events." (Engel et al., 2001)

The main principles organizing the predictive coding (PP) way of thinking are the following (as summarized in (Wiese and Metzinger, 2017)):

1. **Top-down Processing**: Computation in the brain crucially involves an interplay between top-down and bottom-up processing, and PP emphasizes the relative weighting of top-down and bottom-up signals in both perception and action.
2. **Statistical Estimation**: PP involves computing estimates of random variables. Estimates can be regarded as statistical hypotheses which can serve to explain sensory signals.
3. **Hierarchical Processing**: PP deploys hierarchically organized estimators (which track features at different spatial and temporal scales). [14]
4. **Prediction**: PP exploits the fact that many of the relevant random variables in the hierarchy are predictive of each other.
5. **Prediction Error Minimization (PEM)**: PP involves computing prediction errors; these prediction error terms have to be weighted by precision estimates, and a central goal of PP is to minimize precision-weighted prediction errors.
6. **Bayesian Inference**: PP accords with the norms of Bayesian inference: over the long term, prediction error minimization in the hierarchical model will approximate exact Bayesian inference.

[14] "Each area of the cortex performs a specific causal inference, and the passage of error messages between areas allows them to be updated. The layered architecture of the cortex corresponds to a natural distribution of the calculation. Descending connections implement the predictive model, that is, the prediction of the n-level signals, based on the representations inferred at the n + 1 level. Conversely, the ascending connections, coming from the upper layers of the cortex, transmit the prediction error, that is the difference between the received input and its prediction. The error signal is used to update the top-level model to prevent this error from occurring again in the future." From Cours de Stanislas Dehaene sur le cerveau prédictif. Original text: "*Chaque aire du cortex réaliserait une inférence causale spécifique, et le passage de messages d'erreurs entre aires permettrait leur mise à jour. L'architecture en couches du cortex correspondrait à une répartition naturelle du calcul. Les connections descendantes implémenteraient le modèle prédictif (forward model), c'est-à-dire la prédiction des signaux du niveau n, sur la base des représentations inférées au niveau n+1. Inversement, les connections ascendantes, issues des couches supérieures du cortex, transmettraient l'erreur de prédiction, c'est-à-dire la différence entre l'entrée reçue et sa prédiction. Le signal d'erreur est utilisé pour mettre à jour le modèle de niveau supérieur, afin d'éviter que cette erreur ne se reproduise à l'avenir.*"

7. **Predictive Control**: PP is action-oriented in the sense that the organism can act to change its sensory input to fit with its predictions and thereby minimize prediction error; among other benefits, this enables the organism to regulate its vital parameters (like levels of blood oxygenation, blood sugar, etc.).
8. **Environmental Seclusion**: The organism does not have direct access to the states of its environment and body, but infers them (by inferring the hidden causes of interoceptive and exteroceptive sensory signals). Although this is a basic feature of some philosophical accounts of PP, it is controversial.
9. **The Ideomotor Principle**: There are "ideomotor" estimates; computing them underpins both perception and action, because they encode changes in the world which are registered by perception and can be brought about by action.
10. **Attention and Precision**: Attention can be described as the process of optimizing precision estimates.
11. **Hypothesis-Testing**: The computational processes underlying perception, cognition, and action can usefully be described as hypothesis-testing (or the process of accumulating evidence for the internal model). Conceptually, we can distinguish between passive and active hypothesis-testing (and one might try to match active hypothesis-testing with action, and passive hypothesis-testing with perception). It may however turn out that all hypothesis-testing in the brain (if it makes sense to say that) is active hypothesis-testing.
12. **The Free Energy Principle**: It says that any self organizing system must maximize the evidence for its own existence, which means it must minimize its free energy using a model of its world, which on most PP accounts would amount to the long-term average of prediction error.

Promises and limitations

The predictive coding approach has been largely investigated and criticized at a philosophical level (Clark, 2013; Clark, 2015b; Hohwy, 2013; Hohwy, 2016). Several claim that predictive coding is "a genuine departure from many of our previous ways of thinking about perception, cognition, and the human cognitive architecture" (Clark, 2013) while others show that it has connections to other traditions (Kant (Swanson, 2016), German neo-Kantians of 19th century and Husserl's phenomenology (Zahavi, 2018)). Predictive coding is also claimed to be a modern version of Helmholtz *unconscious inference* principle who was among the first to apply the idea of Bayesian inference to sensory perception (Helmholtz and Southall, 1924).

Predictive coding is attractive for the study of APES for various reasons:

- It integrates perception, cognition and action under a unique regulatory principle (feature 5).

- It requires local and global interplay between bottom-up and top-down information flows (feature 1). Fig. 2.14 is an example of global integration of various flows in the brain.
- It allows the presence of internal "mental" models that are dynamically exploited as hierarchical hypothesis generators (features 4 and 11).
- It emphasizes the active nature of cognition, hence perception (features 7 and 11) (Donnarumma et al., 2017).
- It justifies attention as an optimal control principle (feature 10) (Feldman and Friston, 2010).

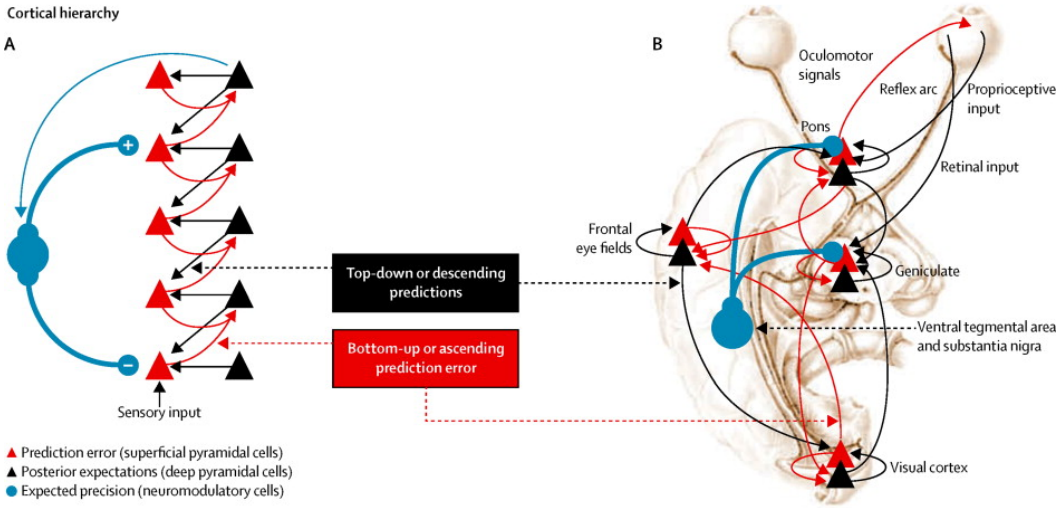


Fig. 2.14: Integration of perception and cognition as a single network. "Neuronal activity encodes expectations about the causes of sensory input, and these expectations minimize prediction error. Minimization relies on recurrent neuronal interactions between different levels of the cortical hierarchy." From (Friston et al., 2014).

One important dimension of the predictive coding approach is that the problematic question of perceptual boundaries, as discussed page 15, is somehow (dis)solved in a global sensory predictive objective, making the whole cognitive activity directed towards perceptual prediction. Predictive coding, in a way, reverses the way of taking the problem: cognition becomes an activity essentially devoted to perception prediction. Perception prediction is not a means, it is the main motor of cognition, its fundamental principle.

Predictive coding has also been proposed as a way to solve the cognitive penetrability of perception problem i.e. the idea that top-down processes may, or may not, shape perceptual processes (Zeimbekis and Raftopoulos, 2015). (Lupyan, 2015) states "that the controversy surrounding cognitive penetrability of perception — the idea that perceptual processes are influenced by non-perceptual states – vanishes" when adopting a predictive coding interpretation of cognition and concludes that "penetrable perceptual systems are simply better and smarter in fulfilling their function of guiding behavior in a world that is at once changing and predictable." Predictive

coding makes perception penetrable by principle, but at a very local scale through top-down prediction connections (see Fig. 2.14).

This unification is conceptually appealing, but has however several drawbacks when trying to apply it in practice for engineering purposes or for scientific investigation: the fact that every element or component of the cognitive system may contribute to the overall predictive objective, although the predictive coding approach has both local (Bastos et al., 2012) and global (Friston, 2010) instantiation, makes analysis and control difficult.

A first response to that problem is to restore the idea of modularity of cognition but in a renewed way. (Drayson, 2017) argues that the continuity claim — the fact that there are no clear boundaries between cognitive processes and non-cognitive (especially perceptual) processes — and the non-isolation claim — the fact that "no part of perceptual processing is informationally isolated from higher-level cognitive processing" — do not necessarily entail that predictive architecture is not modular. She proposes "to understand modularity as a flexible and dynamic feature of architectures, and to appreciate that predictive architectures are modular architecture", perhaps not in a strict information encapsulated way. The question however remains to figure out how a predictive architecture is able to *dynamically* restrict and modularly organize information processing.

Another problem with the predictive coding approach is its solipsism. As (Frith, 2007, p. 132) highlights: "My Perception Is Not of the World, But of My Brain's Model of the World" (feature 8). In a way, a predictive coding based system does not externalize anything, except if it is understood as a means to optimize its predictive error, for instance by acting so that the resulting sensory data is changed. The value or perception and action is not their useful contribution to other tasks — as sign or measure producer — but their ability to internally anticipate sensory data and features.

Perceptual experience

Perception and especially vision is a heavily studied topic in experimental psychology. The fact that it is a dynamic, constructive, adaptive and selective process is not really questioned[15].

In this section, I will focus on specific features that should be considered for APES design.

[15] In a general introduction about perception, (Pomerantz, 2003) states the " following eight facts [that] cover the basics of what is known and widely accepted about perception today": it is limited, selective, refers to the distal stimulus, not the proximal stimulus, requires time, is not entirely veridical, requires memory, requires internal representations, is influenced by context.

Perceptual organization

A first admitted feature of perception is the fact that the external world, given by senses, appears readily *organized*, i.e. is shown to the mind with structure, shapes, objects, layout, etc. One account of this phenomenological experience is the Gestalt school which hypothesizes that visual experience follows principles of proximity, similarity, symmetry, etc. so as to make "the whole something else than the sum of its parts" (Koffka, 2013). One notable phenomenon of "wholeness" in perceptual experience is the figure/ground separation that is manifest in famous examples of object perception instability (e.g. Rubin's vase/face drawing (Wagemans, 2015, section 4)). In these examples, the foreground is perceived globally as a segmented *salient* shape, and not as the mere aggregation of independent elements.

An extension of the idea that the world appears spontaneously meaningful is by introducing the idea of *affordance*, i.e. to assume that "to perceive [surfaces] is to perceive what they afford, [implying that] values and meanings of things in the environment can be directly perceived" (Gibson, 2019, pg. 119). "The affordance of some thing does not change as the need of the observer changes. The observer may or may not perceive or attend to the affordance, according to his needs, but the affordance, being invariant, is always there to be perceived" (Gibson, 2019, pg. 130).

These two related approaches of perceptual experience — Gestalt and affordance — share the fundamental idea that perception, at some point, is ruled by invariants that are accessible, just waiting to be used for various tasks. This view is in agreement with the classical symbolic AI approach that considers perception as a large database of meaningful percepts filled by some automatic cognitively impenetrable process (Firestone and Scholl, 2016; Pylyshyn, 1999).

However, specifying these invariants and designing ways to build them is a difficult task. The search for invariants has been fully studied for instance in geometrical vision (Mundy, 2006), but has been postponed by lack of robustness and expressiveness for real world applications. Indeed, it seems difficult to imagine invariants independently of any purposive objective function, without asking what will the invariant be used for or what should be the good trade-off between specificity and invariance (the only generic invariant is the constant function (Burns et al., 1992)). Deep learning models offer a new way to search for invariance as they can be interpreted as implicitly searching genericity in the first steps of network layers initially optimized for classification and used as all-purpose image features for other tasks.

Attention

Perceptual organization addressed the question of an input sensory structure, and hypothesized that raw sensory data were shaped by automated processes offering

ready to use informative primitives or invariants. Although these first building blocks encode sensory data for better "ecological" efficiency, the compression step provided by perceptual organization is, hopefully, under-constrained: there are many ways to organize the perceptual stream to adapt to the possible tasks that rely on it. This is where *attention* plays a central role as a selective and purposive mechanism.

There is a very large body of evidence that attention is a key feature of cognition either from neuroscience (Gazzaniga and Ivry, 2013, chapter 7), experimental psychology (Nobre, 2014; Carrasco, 2011) or philosophy (Mole, 2017; Ganeri, 2017) perspectives.

The role of attention can be considered as a global physiological state — it is usually referred to *arousal* in this case — or as a way to *select* cognitive resources, actions and inputs. This last selective dimension is especially relevant for perceptual systems, and is commonly justified as a way to deal with the limited processing or memory capacity of the brain (Buschman and Kastner, 2015).

Attention is fundamentally a cognitive high-level conditioning perceptual process ruled by capacity constraints. The influence of cognition to perceptual behavior has been spectacularly demonstrated by the variation of fixation point and saccade scanpath when asked to answer different types of questions about the visual content of a picture (Fig. 2.15) (Tatler et al., 2010). The nature of the scanpath can be interpreted as a fingerprint of the underlying task asked to the subject (Borji and Itti, 2014), although not a clear index (Greene et al., 2012).

Fig. 2.15: Cognitive conditioning of eye saccades. Center: subject asked to freely examine the frame. Right: the subject is asked to estimate the material circumstances of the family (Yarbus, 1967). See also (Tatler et al., 2010) for a more recent reproduction of the original Yarbus experiment.

Attention is an essential and ubiquitous feature of perceptual experience. Fig. 2.16, for instance, displays the various areas in the brain that are claimed to play a role in visual attention. As the figure shows, perceptual attention involves cortical and sub-cortical areas, and exploits various targets of selection, besides gaze control, such as memory and perceptual features (Gillebert and Humphreys, 2013).

The generality of attentional phenomena in cognitive activity makes the clear specification of the object of study difficult: it is common for instance to make a distinction between overt and covert attention, or to organize the various forms based on the

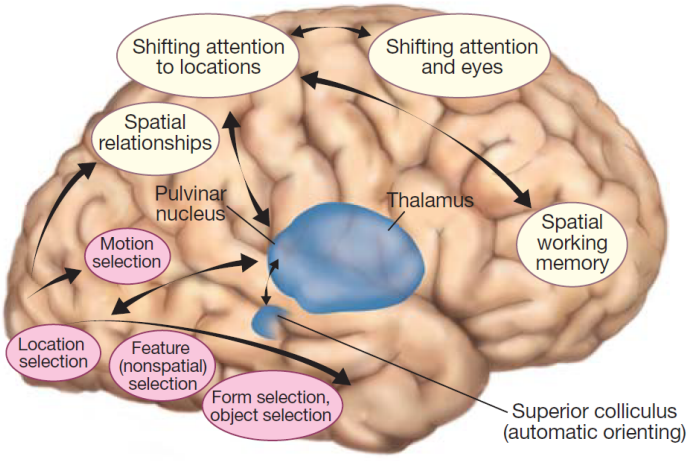


Fig. 2.16: Cortical and subcortical regions involved in visual attention. From (Gazzaniga and Ivry, 2013, chapter 7). See also Fig. 2.1 for a more specialized diagram about saccade generation.

internal/external location of attention targets (Chun et al., 2011). Attention is also either considered as a selection, able to discard irrelevant resources (the most standard approach), as a modulation weighting features and information sources or as a way of regulating priority process (Watzl, 2017).

Attention is therefore more than a simple perceptual system feature, but an essential dimension of cognitive activity and even of consciousness (Cohen et al., 2012; Watzl, 2017). Here again we find that it appears difficult to isolate perception from cognition due to the functional loopy structure of information flows that connect the various cognitive components involved in the production of perceptual outputs.

The consequence for APES design is 1/ that attention should be functionally exploited in models 2/ but that the variety and complexity of attentional features make them too difficult to be apprehended globally. What I propose to simplify the picture is to consider attention as a repertoire of means that efficiently control resources to complete cognitive tasks, with three different functional aspects: selection, modulation, and priorizing.

Attention — at least of a certain type — is now widely used in deep learning models as a generic computational principle (Vaswani et al., 2017). Perceptual attention computational models have mainly focused on two problems: salient object detection (Borji et al., 2014a; Borji et al., 2015) and gaze prediction (the "where to look next" problem)(Borji et al., 2014b; Bylinskii et al., 2016) and used for several perceptual applications (Nguyen et al., 2018).

We will give a more detailed discussion on formal models in the following chapter.

External influence to perception

We have seen in the previous sections two fundamental features of perceptual experience: it deals with an organized field of sensory inputs, and the efficient exploitation of this field requires various forms of resource management that can be associated with the idea of attention, a conditioning process that acts as selection, modulation or priorizing.

Another spectacular phenomenon that reveals an external — negative — influence on perceptual experience is the failure to detect several events that "normally" should have been noticed. Two major forms of failure have been thoroughly investigated: *inattentional blindness* (Mack and Rock, 1998) ("observers generally do not see what they are looking directly at when they are attending to something else"[16] (Mack, 2003)) and *change blindness* (Rensink et al., 1997; Simons and Rensink, 2005) (observer do not perceive slight changes in sensory inputs when another perturbing event is inserted) [17]. What those phenomena reveal is the difficulty of making perceptual behavior independent from an underlying master task in which the subject is engaged, impairing the capacity of the subject to become aware of signals that are expected to be otherwise noticeable. It is not clear however whether perception badly selects informative sources and fails, or if the role assigned to perception by the main task causes blindness.

Task conditioning is not the only influential actor: emotion (Zadra and Clore, 2011) or motivation (Engelmann et al., 2009) may globally modulate perception, multiple sensory modalities may interact (Shams and Kim, 2010), working memory state and visual attention are tightly interdependent (Chun, 2011; Gazzaley and Nobre, 2012). A larger list of global influences on perception can be found in (Firestone and Scholl, 2016).

However, findings form experimental psychology are difficult to transfer to APES design, for several reasons:

- Perceptual experience features are indirectly inferred from global behavior analysis, using objective physiological measurement or by collecting subject outputs (action, reaction, wording). They can't be used to give precise insights of what's in the perception box.
- The fact that perception, when experienced, receives potentially many sorts of influence is an obstacle for its study in isolation — we are faced once again with the problem of perception/cognition separation.
- Perceptual experience is often understood in connection with conscious awareness, a global mental state — perceptual blindness phenomena, for instance, are revealed because subjects become unaware of meaningful events. Here

[16] One famous example is the invisible gorilla illusion http://www.theinvisiblegorilla.com/videos.html.
[17] See http://nivea.psycho.univ-paris5.fr/#CB for examples.

again, how perception contributes to awareness, from what mechanical processes, is hard to devise.

In a previous section, we proposed to model *perception as a service* to try to solve the conceptual difficulty brought by drawing fixed functional boundaries with cognition (see page 27). Using this formalism, both parts have some knowledge of what they expect from each other. The cognitive conditioning of perception is partly ensured by the contracting phase that determines output requirements and defines the input elements or priors that may be necessary to complete the perceptual task. Those inputs may be misleading for several perceptual tasks (for instance when asking for the presence of gorilla after having asked to count passes in the "invisible gorilla" experiment), but not for the main current task. Decomposing perceptual experience in this framework could be a way to clarify a potential role of perception in an extended and more versatile way.

Philosophy of perception

Perception, as the spontaneous and unique window and knowledge source of the world, has been the concern of many philosophers (Matthen, 2015, part 1), and continues to stimulate proposals and discussions (see (Fish, 2010) for a presentation of recent proposals mostly from analytic philosophy tradition). Two main directions have been investigated. The first one addresses metaphysical and epistemological questions and discusses in what terms and conditions perceptual experience is able to veridically account for the external world (Lyons, 2017; Crane and French, 2017). The second one is more concerned with describing the experience of perceiving itself, its phenomenology, and its relation to consciousness and mental states.

This section will of course not present the whole corpus of philosophical investigations and proposals on this topic, but select several ideas that may be useful to guide APES design, with the underlying and arguable assumption that natural perception — the object of philosophical inquiries — is the ultimate reference of artificial systems.

The problem of representation

The common assumption in science and engineering is realism, i.e. the fact that perception has to say something about the outer world. Perception can be used as a measure, interpretation or description of the world. Once it is faithfully represented, inference from the perceptual content can consequentially be made for various tasks. Philosophical investigations are mostly motivated by questioning this *naive* realism assuming that perception accounts for the world as it is.

A critical epistemological issue, therefore, is the nature, and the existence, of perceptual representations, or more generally of perceptual content (Brogaard, 2014;

Siegel, 2016). Indeed, it is customary to make a distinction between two types of perception — direct and indirect — that assume the existence of a mediation between the subject/perceiver and the world, or not. The mediation may take the form of what has been called sense-data (Huemer, 2016), that are mental entities, ideas or objects possessing the properties or attributes that appear to us.

Thinking of perception this way, i.e. through the mediation of a representation, has several conceptual advantages: it is an efficient formal recipient for references to the world, and can host veridical status — a representation can be true or erroneous. It can also explain in a simple way illusions or hallucinations as flaws in the representational process (Fish, 2009), phenomena sometimes presented as being *the problem of perception*. A correlated issue is to make clear what it means to be true or false for perception, what is the relation between perception and knowledge.

However, as (Nanay, 2015) states, the fact that perceptual representations exist has been questioned: "Twenty years ago, the vast majority of philosophers of perception would have agreed that they do, but this is no longer so." Instead, several philosophers claim that "conscious perceptual experience is neither reducible to nor explicable in terms of representational states or content." (Locatelli and Wilson, 2017)

Two main strategies have been proposed to eliminate representation from perceptual experience: 1/ Avoid the use of representation as an intermediate structure between world and mind when describing or explaining perceptual experience. 2/ Extend the object of investigation from mere perception to a larger set of experiences.

The first type of approaches claims that there is no need of a mediation to have an experience of the world: perceptual experience is perceiving objects, not a representation of mental objects. One proposal to account for the direct experience of the world is by introducing the idea of *intentionality*, the fact that subjective experience takes the form of an object that is, by nature, in relation with something else. The term intentionality has a long history in philosophy and many variations. It is sometimes simply presented as *aboutness* of mental states, allowing the the existence of a content, which makes it a version of representationalism. Another tradition endows intentionality with an ontological role, and asserts that perception is just what is experienced — we directly see a "red tomato" as an object of the world, we do not see an internal representation of an object which has been causally constructed from the world with 'redness' and 'tomatoness' features. (Drummond, 2012) discusses the various forms of intentionality — representationalist and presentationalist — and their limitations. We will present a more detailed discussion of this question of intentionality in the next section about phenomenology.

A second type of approaches proposes to eliminate the central role of representation by ontologically intricating perception with the whole subjective experience of the

world, and especially with body experience, leading to an *embodied* or *enactive* perception (Varela et al., 1991; Gangopadhyay and Kiverstein, 2009; Wilson and Foglia, 2017; Shapiro, 2014; Gallagher, 2017). Two important aspects introduced in this trend are the essential dimensions of *temporality* and *action* in perceptual experience: no perception if it is not temporally constructed from outer or inner actions, those actions being real — they have an influence on the world or the body — or virtual — they simulate their impact.

The importance of action in perception leads to the potential role of anticipation or prediction capacity that will give its horizon to experience. Prediction is obviously a central concern of the Predictive Coding approach (see page 36) which unifies experience under a common regulatory principle where the brain constructs internal models of the possible causes of sensory inputs to internally minimize the weighted prediction error on these inputs. (Clark, 2012; Clark, 2015a) defend predictive coding as a way to conciliate representionalism and anti-representationalim such as the enactivist trend. In the same spirit, Madary proposes that "visual perception is an ongoing process of anticipation and fulfillment. In short, perception is best understood as an ongoing cycle; but instead of a cycle of action and perception, it is better understood using the more general framework of anticipation and fulfillment."[18]

Phenomenology: perception and consciousness

Perception can be studied as a faculty, with the role of providing correlates of external objects and states of affairs, making the role of a hypothetical content a central issue, or as a whole conscious experience. What phenomenology proposes is to unify these two different views under the primacy of phenomenon from which perceptual content can be derived as a feature.

Phenomenology (Gallagher and Schmicking, 2010; Zahavi, 2012) and perception are terms that have been associated in many philosophical studies since the seminal works of Brentano (Brentano, 2014), Husserl (Husserl, 1931; Husserl, 1997) and Merleau-Ponty (Merleau-Ponty, 1945). This way of thinking starts from the nature of perceptual experience as it appears to consciousness in order to describe, classify and identify its common and invariant structures and mechanisms. It therefore takes the problem of subjective experience seriously, not as a conceptual chimera, and puts the question of consciousness as an object of investigation.

A key concept of Phenomenology[19] is intentionality — we have already introduced this term in the previous section — that characterizes the mind "'as a whole' rather than to particular mental events or states [...]. Mind as such is intentional. Mind as such transcends itself towards the world and relates itself to the existent world, and every instance of 'minding' the world participates in this relation, albeit, [...]

[18] Cf. Madary, 2016, p. 9.
[19] In the following, Phenomenology with a capital letter will refer to the philosophical movement.

in different ways."[20] Roughly summarized, the world exists to the mind because consciousness is, *in essence*, intentional.

Siewert provides a summarized account of phenomenological insights about perception: "Beginning with Edmund Husserl, the intentionality of perception is investigated by asking: how can experience, itself in near constant flux, nonetheless be of stable objects, so that meaning and knowledge might be possible for us? The key to answering this question, he proposes, is to see perceptual consciousness as dynamic and prospective — a process wherein the needed constancies are achieved via the successful anticipation of further experience through movement and direction of attention."[21] As Noë states, the main phenomenological question of vision, and also of other senses, is that "as a result of saccadic suppression, the data made available to the retina takes the form of a succession of alternating snapshots and grey-outs. How, on the basis of this fragmented and discontinuous information, are we able to enjoy the impression of seamless consciousness of an environment that is detailed, continuous, complex and high-resolution?"[22]

Presenting the (too) complex, sometimes esoteric and technical ensemble of concepts and words that have been developed by the founder of the Phenomenology, Edmund Husserl, with the objective of offering a *method* (Schmicking, 2010) able to faithfully describe conscious life as it appears, is clearly beyond the scope of this short section. One can however remember from the previous paragraph two central ideas: the dynamic and attentional dimension of consciousness and the participation of knowledge on perception through anticipations. The enactive or embodied approaches of perception mentioned in the previous section can be interpreted as a modern extension to Phenomenology, at least of several of its original intuitions about kynaesthesic perception (Husserl, 1997; Drummond, 1979).

The question of the relation between cognition and perception, a recurrent problem as we have seen in previous sections, is somehow solved by Phenomenology by eliminating the duality between pre-existing mind and world, or better, by describing how the world exists *transcendentally* for a mind without any metaphysical hypothesis.

Invoking consciousness when addressing engineering or scientific issues is always risky: consciousness is never far from affect, feelings, non-conceptual experience[23], and to moral questions such as the nature, condition and even existence of free will. Critics and proponents of reductionism — the hypothesis that mind can be reduced to physical or neural activity — abound in philosophical literature. For instance, one of the most recent and commented philosophers who have been critical about reductionism, Markus Gabriel (Gabriel, 2017), revisits recent philosophical

[20] Cf. Drummond, 2012, p. 125.
[21] Cf. Siewert, 2015, p. 194.
[22] Noë, 2002.
[23] Several philosophers have proposed the word *qualia* to refer to subjective experience, a concept heavily criticized by Daniel Dennett as an ill defined object (Dennett, 2017).

proposals that have accompanied the development of neuroscience and promotes a "neo-existantialist" anti-scientist approach of consciousness that could allow a better account of freedom.

However, the phenomenological approach is appealing for scientists and engineers, since it relies on *experimenting*, although at the first person level, and has led to an objective of "naturalization" (Petitot et al., 1999; Gallagher, 2012), that has been followed either as "an extension of natural science" or, more modestly, "as a meaningful and productive exchange with empirical science"[24]. This was this last option I tried to follow in my PhD thesis (Herbin, 1997a) where one of the objectives was to investigate a phenomenological description of visual recognition, taken as the simplest, although already complex, simultaneously cognitive and perceptual experience.

Phenomenology of visual recognition

Two chapters in my PhD thesis (Herbin, 1997a) have been inspired by Phenomenology.

The first one presented the literature in computer vision available at that time (i.e. before the "Deep Learning Era") under two points of view: vision for recognition, and recognition by vision, and tried to examine how intentionality could be implemented along three dimensions: integration of the sensory and representative point of view, computational and not simply algorithmic definition of a dynamic vision and declared specification of a semantic and perceptual context. This chapter was also a general justification of the idea that recognition should be active by nature — in operation and specification — introducing the formal models developed in the subsequent chapters.

The last chapter was a more general questioning of visual recognition divided in two parts: a critical discussion of recognition seen as a matching between perception and cognition representations, and a tentative study of a phenomenological account of visual recognition, a cognitive capacity characterized by the presence of an objective exterior — a world of objects, the experience of knowledge, and the inscription in a temporality manifested by the suffix re-. The study discussed the work of two philosophers: Gilles Deleuze and Edmund Husserl. The first philosopher helped us develop a critical perspective of the dogmatic way of thinking that reduces recognition to the normative manipulation of concepts and to the reduction of recognition to the experience of the same. The second philosopher allowed us to discuss a phenomenological account of perceptual recognition, and to identify several limitations of the idea of fulfillment, a key feature of perceptual intention, when instantiated with recognition. The confrontation of these two philoso-

[24] Cf. Zahavi, 2010, p. 14.

phers resulted in the proposition that the fundamental features of cognitive act such as visual recognition, its essence, has the form of a circumscribed multiplicity and the nature of a virtual alterity. In a less fancy formulation, one can say that visual recognition is both *about* an external world, but that the set of indeterminations of this world *among* which to choose, what it is not, is defined in the recognition act itself as one of its features.

2.3 APES study and design

The discussion above mixed considerations from various perspectives (artificial intelligence and cognitive science) in order to identify the specificity of APES, i.e. consider perception as an autonomous system, and has often assumed that natural perception is the model by which being inspired.

This section summarizes the identified properties that artificial APES should implement and discusses several challenges that should be addressed in order to satisfy them.

APES properties

The properties that an autonomous perceptual system should verify can be organized according to its nature, to what is expected from it and to the relation with its environment (world and client).

expressiveness

- The role of APES is to express two types of perceptual objects: measures and signs referring to features, properties or attributes of the world.
- Perceptual objects have one or several recipients that are engaged in a specific task for whom they are meaningful.
- The value of signs or measures is related to recipient needs, i.e. to their potential contribution to the task.

agency

- Perception is a constructive process that involves complex mechanisms subsuming the passive functional input/process/output chain and is therefore better understood as an active dynamical system: an agent.
- As such, an APES adapts to the situation with different time scales: A/ by modifying its functional structure to comply with a contracted specification, B/ by dynamically adjusting its resources to fulfill its specified objectives,

typically by selective attention, or C/ by learning how to satisfy output quality requirements.

cognitiveness

- The production of perceptual outputs makes use of skills such as memory, decision, knowledge, planning, reasoning that are usually considered as cognitive: the separation between perception and cognition is functionally mutable.
- The distinction between cognition and perception can be formally instantiated as a flexible client/server architecture, where the relation between the two parts is contractually specified.

trustworthiness

- Perceptual products may be qualified and justified to the recipient for efficient and safe exploitation.
- Reliability of perception is acquired when contracted requirements between the perceptual system and the user/client are satisfied. This implies that the perceptual agent assumes explicit responsibility for what it produces.

Challenges

The list of desirable properties stated above is abstract and requires work to be fully satisfied. This section discusses broad challenges or issues that may organize research actions. After all, we still do not know how to design reliable, efficient, versatile and safe artificial perceptual systems.

Systemic complexity management

A fundamental problem when considering perception as system is the need to master the complexity of its structure and behavior. The examples of Fig. 2.2, 2.1, 2.4 and 2.3 show intertwined patterns of functional relations, not to mention the dynamics of their inner states.

Complexity management is a very old scientific topic[a]. The engineering of large and complex systems has also given rise to several studies (Dominique Luzeaux, 2011). However, applying those results to perception itself has not been really addressed.

Thinking of a system as a set of interconnected *modules* with dedicated roles is a spontaneous approach. Modularity, as a design principle, decomposes a complex task in small, simple and controllable pieces, with local requirements, and assembles them to complete the final task. When applied to perception, however, modularity happens to be too rigid, not robust to hazards and, at

the end, rather inefficient because resources are often multiplied instead of being shared.

One of the reasons of the recent successes of deep network approaches is to relax in a certain way the fine local control of the modularity approach, and to solve problems by distributing their resolution between small calculation units — the neurons — without prescribed roles, using *end-to-end* learning strategies. The price to pay, as is well known, is a lack of intelligibility of the resulting perceptual process, which is a clear obstacle to trustworthiness, but also to adaptability: all possible situations must be represented in the learning data.

There is therefore a need for better formal tools and models that, while being empirically efficient, allow sharing of calculation, flexibility, adaptability, compositionality and global learning.

Recent approaches such as memory networks (Weston et al., 2014), neural module networks (Andreas et al., 2016), Neural Turing Machines (Graves et al., 2014) or capsule nets (Sabour et al., 2017) are tentative answers. They still require however developments to be competitively compared with raw opaque deep learning approaches.

[a] See for instance the nice interactive map displaying the various trends and disciplines associated with "complexity sciences" http://www.art-sciencefactory.com/complexity-map_feb09.html

Model and evaluation of perception as sign production

We have proposed the idea that the output of perception may take the form of signs, i.e. objects that combine three components: the sign itself, what it refers to, and its recipient — the *interpretant*. This proposition was presented as a way to solve the difficulty raised by defining perception as a univocal and universal description of the world taking the form of a representation that eludes what it should be used for.

Introducing a third actor — the recipient or *interpretant* — can be seen as a kind of trick that delays the precise specification of perceptual products: their value depend on the nature and needs of the interpretant, which makes it part of the model itself.

One possibility of describing its role in the model is by instantiating it as a requirement or constraint for sign generation.

This leads to questions related to the status of *semiosis*, the process that actually builds the signs — in our case perception: When will it be said to succeed or to fail? How decide if a sign is right or wrong? In what

sense? How define ground truth, and metrics, for an object that has a triadic nature (reference/sign vehicle/interpretant)? If a sign is bad, is it because it does not meet the requirements, or is it because the requirements are bad themselves?

Implementing the principle of perception as sign production as an engineering project addresses therefore two problems: How define practical formal models that explicitly consider the user/recipient as a feature of the perceptual process? How evaluate them?

Versatility of perception

As we have seen, it seems difficult to define fixed and inalienable limits to perception, especially when it has to be opposed to cognition. It has been proposed to bypass this aporia by endowing perception with autonomy and bind it to the cognitive task to which it contributes as a service, thereby eliminating the existence of a "pure" ateleological perception module.

One of the main reasons to avoid a rigid interface between cognition and perception is the fact that there are multiple types of perceptual outputs depending on what they should be used for: in other words, perception is expected to be versatile, multi-purpose. The next question is to design models able to efficiently implement this versatility property.

Of course, an efficient versatile perceptual system should not consider each task independently and should factorize its resources. Multiple task learning (as exemplified for instance Fig. 2.8 pg. 21) is one answer to this question, but is mainly used as as a way to improve and regularize inner neural representations by sharing objectives.

More generally, what is needed is a model able to adapt to a variety of tasks with limited means. Two complementary research directions are possible to address this question of resource economy: either functionally using ideas such as compositionality and module re-usability, or dynamically through selective attention, scheduling, planning, or information flow control (sequential vs. parallel).

With the rather recent and drastic improvement of perceptual capacities of artificial systems, the question of versatility can now become one item in the research agenda. Typical issues are the following:

- *optimality:* How define global costs or principles able to arbitrate between conflicting tasks? How specify them?

- *evaluation:* How define metrics measuring the general adequacy of a multi-purpose system?
- *incrementability:* How increase the repertoire of tasks, or the integration of an existing (off-the-shelf) function? How ensure that a new task will increase perceptual capacities and not damage the system?

Specification of perceptual service

The idea of considering perception as a service was a way to make more flexible its relation to the recipient of its production as expressed by the lists of questions of Tab. 2.1 and Tab. 2.2. They need however to be more quantitatively and formally specified to be exploited in an artificial model.

Those questions address various dimensions of the relation between perceptual system and user/client as a whole process: the nature of the perceptual products, their target quality through explicit requirements, but also describe and anticipate the different ways of reacting, from both parts, when those requirements cannot be satisfied.

This dynamic dimension of the client/server relation is an essential feature that gives its value to perception by making it functionally trustworthy. It is also a key ingredient able to define flexible interface between cognition and "pure" perception by specifying, as long as the service is being active, what level of *impenetrable* processing is delegated to the server. However, before developing processes and algorithms implementing the service, it is necessary to define a formalism capable of modeling this flexibility.

Representation elimination

The existence and role of a mental representation has been actively discussed in cognitive science and philosophy, as has been briefly related in the previous section. One of the extreme option was to invoke consciousness as a starting point from which derive perception as a participant to mental experience. Consciousness, in its dynamical account by Phenomenology — at least in the Husserlian tradition — can be presented as a way to avoid the prerequisite of making world representation the main objective of perception.

Questioning the role of representation from an artificial intelligence perspective translates into a more practical interrogation: Can we really do without perceptual representation?

For engineers or computer scientists involved in artificial perception design, this question can at first be rather irrelevant, since one major objective is to imagine algorithms able to build representations from sensory data that

will be exploited in various tasks. During the design process, engineers need to know what is the system actually doing, measure its performance, and optimize it. Exhibiting a representation is a handy output that can be used as a behavioral gauge to reveal in more details how the perceptual process is functioning — bad or good.

There have been several attempts to get rid of representations, mostly inspired by an embodied cognition approach posing sensory-motor loops as the key structuring element so as to produce "intelligence without representation" (Brooks, 1991; Di Paolo et al., 2017). Realizations have been spectacular — evolving robots designed from limited principles — but limited in scope and demonstrated cognitive capacities.

The question remains to be able to design systems whose behavior can be anticipated, if not controllable — an engineering objective — without introducing representation as an intermediate feature. Predictive coding (see presentation page 2.2) may offer a more cognitive alternative to the sensorimotor approach that reduces the role of perception as a means to act.

Biologically plausible *and* efficient models

Artificial intelligence is very far from providing systems as versatile, adaptive, integrative and resilient as the brain, in other words, as *globally* efficient.

Nature has been therefore a constant inspiration for artificial perceptual systems but very few models have given rise to efficient approaches from an engineering point of view, except maybe for low level computer vision (Medathati et al., 2016). However, the popular Deep Neural Networks, although originally inspired by neuron models, have very few in common with brain components, structure and behavior, although connections between the two worlds of artificial and natural neural networks continue to be drawn (Marblestone et al., 2016; Hassabis et al., 2017).

Brain behavioral studies have of course given rise to many formal proposals: however the goal of such models is primarily the verification of scientific hypotheses, not the reproduction of behaviors themselves, or if so, at a simplified scale with the same objective of scientific fact assessment. Regarding vision — the most studied perceptual modality — models have addressed various scales of brain behavior and structures, from fine neural population dynamics (Acebrón et al., 2005), to models for saccade generation dynamics (Girard and Berthoz, 2005), and global description of visual process integration in the brain (Bullier, 2001). When more cognitive capacities are involved, that

require more abstract interpretation, models tend generally to get farther away from biological observation.

One potential approach that may produce a fruitful integration of artificial intelligence and neuroscience objectives is predictive coding (see pg. 36). Several formal models have been proposed to implement its principles at the neural level(Bastos et al., 2012) (Spratling, 2012) (Spratling, 2016) and for artificial perception applications: (Lotter et al., 2017) for instance describe deep network model inspired by predictive coding model to predict future frames in a video sequence; (Spratling, 2017; Wen et al., 2018) use predictive coding inspired models for object recognition tasks.

However, as (Cox and Dean, 2014) concludes, "seizing the opportunity [of recent successes in machine learning and recent advances in neuroscience technology] will require effort and a cultural shift, as the two fields often have very different goals and approaches".

Self-assessment

It has been proposed in the previous section that perceptual autonomy implied a capacity of being responsible for the quality of the perceptual products, or more precisely that an APES should be able to asses by itself that it can satisfy, or not, the agreed requirements.

A series of questions requires further investigation to instantiate this capacity: What is the nature of this assessment? How can it be built? In what way express it? Can it be trusted?

There are at least two possible types of contents that can be expressed as self-assessment: the quality of perceptual production, and the way it has been produced.

A first issue is therefore to clarify in the first place what should be considered a *good* perceptual output. This problem can be addressed either as truth, that qualifies the faithfulness of the reference to the outer world, or as usefulness, that relates to its *value* for the user.

From a physical view of perception as a measure of the world, acknowledging that perception is inherently noisy, ambiguous, approximate, untrue, is a standard hypothesis: engineering has a long history of formal tool development about error modeling and exploitation. In Artificial Intelligence, uncertainty happens to be more controversial and debated [a], maybe because the distinction between truth and usefulness is unclear.

The capacity of self-assessing usefulness of perception implies that the perceptual system must have integrated in some way what the recipient will do with its production. This role is played by the requirements that have been agreed between the APES and the client/user, meaning that those requirements must be expressed in a form that allows practical verification or checking. We will come back to this question in more details in chapter 5.

A second type of self-assessed content is a statement about the way the APES has built its perceptual outputs: this statement can be used to get confidence about APES behavior or as a justification of the perceptual production quality. Generating such content that makes APES more intelligible will be discussed in chapter 4.

Self-awareness has been rather recently introduced to characterize computing systems able to address their own behavior as humans do (Lewis, 2017; Lewis et al., 2016) [b] Self-assessment can be considered as a fundamental functionality of self-awareness. An interesting direction could be to study how the ideas and concepts developed in this field, mainly targeting cyber-physical objects (IoT) or data centers, can be transferred and adapted to perception.

[a] Cf. the annual conference about Uncertainty in Artificial Intelligence http://www.auai.org/

[b] "Self-aware computing systems are computing systems that: 1. *learn models* capturing *knowledge* about themselves and their environment (such as their structure, design, state, possible actions, and runtime behavior) on an ongoing basis and 2. *reason* using the models (e.g., predict, analyze, consider, and plan) enabling them to *act* based on their knowledge and reasoning (e.g., explore, explain, report, suggest, self-adapt, or impact their environment) in accordance with *higher-level goals*, which may also be subject to change." (Kounev et al., 2017)

Topics of investigation

The goal of this introductory chapter was to propose a research angle that does not restrict perception to a generic provider of information about the world, without any specificity or purpose "in mind". The idea of APES suggests that the right way to study perception is to consider it simultaneously *as* a system and *in* a system.

The next chapters will discuss in more details three different issues to be studied in order to address this approach:

Operation & development How are organized the various systemic components and their dynamic interactions? What let contextually adapted and what should be generic? What can be analytically modeled and what needs to be empirically learned?

Intelligibility What is the best way to monitor the internal states of the system, their dependencies or their correlations? What *signs* should be

generated and outsourced to inform of its behavior — good or bad?

Safety What is expected from the perceptual system? How define a good behavior? How improve predictability and robustness to perturbations, or conversely, how detect instabilities in perceptual systems?

3 OPERATION & DEVELOPMENT — The life of APES

The previous chapter examined in what sense an idea of autonomy could be associated with perceptual systems and discussed its conceptual implication. The purpose of this chapter is to examine how to model and design such systems, with an emphasis on their temporal evolution at two scales:

Operation: How does an autonomous perceptual system "function" and adapt to the situation? What is the dynamical structure of calculations, actions, decisions, etc. that produce perceptual outputs?

Development: What can be formally prescribed, what needs to be empirically learned? How make it evolve and improve?

Those two scales are of course inter-dependent: a functional structure is worth if it can be developed and implemented, a good development relies on the potentialities and expressiveness of the functional structure.

3.1 Operation

Patterns of operation

In the previous chapter, we presented adaptivity as an answer to the control and design of multiple task systems; we proposed that equipping perceptual system with autonomy is a direction towards efficient design for adaptivity, and that learning is a capacity that allows autonomy in the system development process. We have also insisted that a key feature of an APES is the nature of its relationship with users, the recipients and consumers of the signs or measures it produces. The central role of the relationship with users has led to the definition of perception as a service, i.e. a process capable of providing guaranteed answers to given tasks or questions, ruled by an agreed contract.

The emphasis on the place of the final user in APES design can be symbolized by an evolution of the functional patterns that connect the system to its outside. Fig. 3.1 depicts five classes, indexed by an iconic letter roughly representing the connectivity shape:

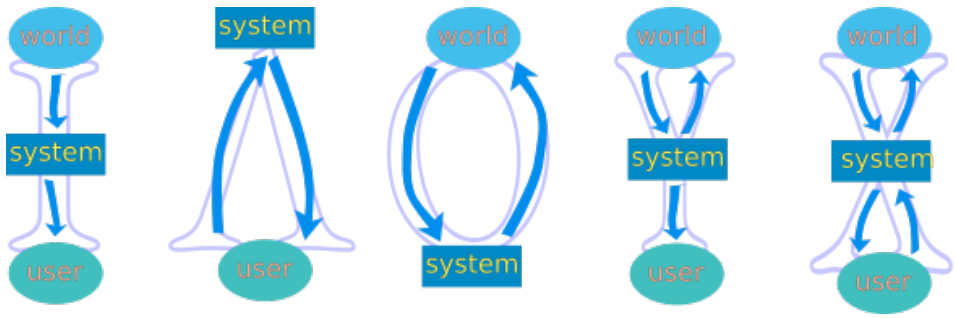


Fig. 3.1: Functional patterns of perceptual systems.

I This is the classical filtering structure that univocally receives world features as input and sends output to an unknown user which has no ability to interact with the system.

Λ In this pattern, the user asks a system to complete a task for him. This pattern can describe distant calculation when the user needs remote resources, for instance, but all the information needed to answer the request are hosted by the system.

O This is the typical interactive connectivity scheme of an autonomous agent where the system receives inputs from the environment and issues actions. This pattern describes best robotics approaches when the system is alone and do not interact with other agents or users.

Y In this evolution of the previous pattern, the system interacts with the world in an autonomous way, and outputs its perception products towards the user but do not receive any feed-back from it. The system knows what to do, and how to interact with the world to provide useful outputs.

X In this last pattern – the richest – the system has to deal with two distinct entities: a world that is the source of input and recipient of actions, and a user that interactively asks for information, skills or resources, for a service. This pattern is the more complex one since it involves two feed-back loops that may conflict. However, it allows seamless and continuous adaptivity and, potentially, better quality of service.

The **I** pattern, although the most practiced in image processing or computer vision, is a filter that maps sensory data to a given output space meaningful to the final user, and does not involve any possible interaction with the source of data: it cannot be considered functionally autonomous. The **Λ** pattern may describe interactions with the user in an autonomous way, but does not produce any output (sign or measure) that relates to a contingent world, losing its perceptual objective. The **O** pattern

interacts with the world, but does not externalize any output towards a potential user.

From the collection of functional patterns, only two involve autonomy and perception – the features specific to an APES approach: the **Y** and **X** patterns. In the following, we will examine how they have been addressed in the literature.

Active perception: the **Y** pattern

Dynamically interacting with the world to produce relevant perceptual outputs has a long tradition in artificial intelligence. The idea that perception has to deal with a temporal stream of sensory data is a rather natural question. The specificity of an *active* perception approach is to control, or at least influence in some way, information sources and other processing parameters to complete the current tasks and output their results. Interaction with the user is only final, and one-way. The asymmetry between a dynamic interaction with the world, and a static relation with the user is represented by the **Y** pattern.

There are mainly two reasons to study an active perception approach: better adaptativity under limited but controllable resources – a perceptual system cannot handle all the aspects of the world due to its limited processing capacity – and incompleteness of readily available perceptual features – the world does not reveal right away its nature. Both reasons justify to *temporally unfold* perception, to make it a true dynamical process.

The fact that perception is the consequence of action is a very old idea, especially in the robotics domain. (Bajcsy et al., 2018) presents a large retrospective view from a perceptual-motor loop perspective by several of the first researchers in this area. "[Their] main argument is that despite the recent successes in robotics, artificial intelligence and computer vision, a complete artificial agent necessarily must include active perception. The reason follows directly from the [following] definition [...]: An agent is an active perceiver if it knows why it wishes to sense, and then chooses what to perceive, and determines how, when and where to achieve that perception." Fig. 3.2.

Their presentation of active perception emphasizes physical action on the perceptual system and is therefore close to an embodied perception approach, "where an agent (animal, robot, human, camera mount) changes position in order to improve the view of a specific object and/or where the agent uses movement in order to perceive the environment (e.g. for obstacle avoidance)".[1]

Another dimension of active perception is also discussed in this recent review, but less explicitly, with the claim that "The essence of active perception is to set up a

[1] https://en.wikipedia.org/wiki/Active_perception

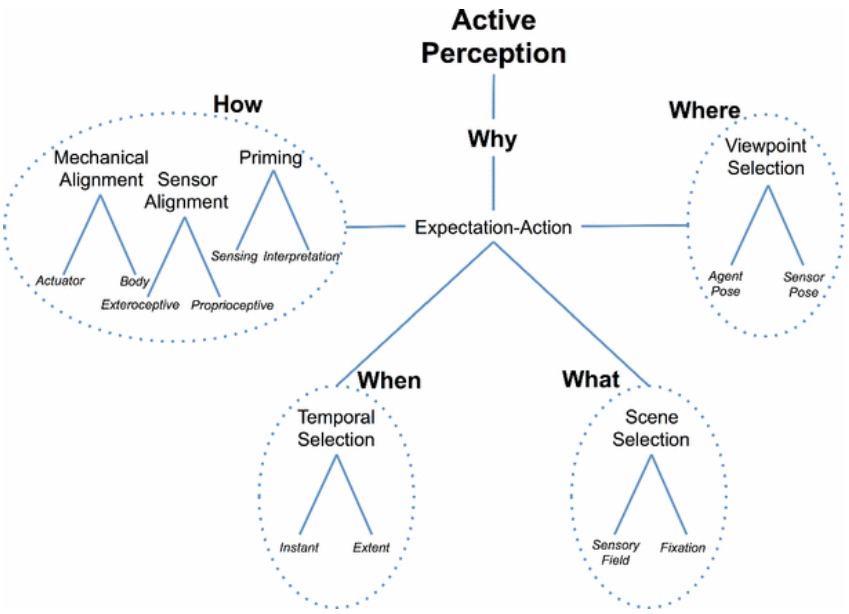


Fig. 3.2: Active vision issues from a robotics perspective (from (Bajcsy et al., 2018)).

goal based on some current belief about the world and to put in motion the actions that may achieve it" (Bajcsy et al., 2018). This statement reveals the informational concern of active perception with the introduction of belief about world states as a fundamental feature.

The activity of perception has a dual nature: actions can be physical (as illustrated in Fig. 3.2), but can also be algorithmic by recruiting the set of calculations adapted to the current situation or perceptual state. From a formal perspective, selecting a point of view, or choosing a feature extractor to apply on the data are both actions that produce new data potentially contributing to task completion (Kragic, 2018).

We propose to abstract the role of an action as control, i.e. any *output* that has an impact on the perceptual system input, the sensory data, and not restricted to physical modifications. This extended definition unifies actions as – controllable – sensory data generation, either through physical or algorithmic means.

An underlying assumption of an active perception approach is that the contour of the system is given: its formal inside/outside boundaries are known and fixed. Actions do not have an ontological impact on the system, and do not modify the nature or the structure of the interaction with its outside – world or user. This assumption maybe problematic if a developmental approach of perception is adopted (Ivaldi et al., 2013), for instance, if the repertoire of skills and actions evolves during system lifetime. We will come back to this question in a further section (pg. 89).

The active perception approach is not limited however to controllable sensory data generation, or more precisely this capacity entails or implies several other properties or features. The following list defines the features of an active perception system:

Dynamic Time inscription is of course a mandatory feature, but perception as an activity is also characterized by a *behavior* with internal state dynamics.

Sequential As a consequence of its dynamic property, active perception recursively updates its internal states from sensory data and generates its products from a sequence or a stream of sensory data and actions.

Active The system outputs a signal that potentially modifies sensory input data. This modification is expected to be predictable in some way, typically thanks to a dynamical or a probabilistic model.

Receptive As perceptual, the system is open to influence. One important dimension of an active system is a capacity to modulate the nature and the form of receptivity, for instance through *attention*.

Incomplete Sensory data only partially reveal the world: perception is always in progress, unfinished, sketchy. The role of an active system is to state when to stop and decide that it is able to deliver relevant outputs.

Cognitive Perception may involve static structures such as models and knowledge, and dynamic components such as memory.

Teleological The purpose of action generation is to, ultimately, improve the level of task completion or user satisfaction. The repertoire of purposes may be itself variable to allow versatility of the system.

Optimal Task completion can be expressed as a global reward or cost: informational uncertainty about states of the world, energy, duration, memory storage etc. Ideally, the role of activity is to optimize this cost/reward with final or anytime objectives, or at least to control it.

Intentional The goal of perception is to produce *references* to objects or properties external to the system in the form of measures or signs.

Self-aware The system should have an estimate or a knowledge of how far it is from its objectives. It may take the form of a *belief* about states of the world.

Models of active perception

The expected properties of an active perception system can be expressed formally, with various levels of complexity and under different kinds of hypotheses and knowledge. The basic ingredients in an active perception model make use of three different types of formal objects: variables, dynamic models and inference process.

We present briefly in this section but in a rather general way their nature and organization.

A first step when modeling active perception is to identify the dynamic variables that stand for the system and the world features. An active system, characterized by an internal state $\mathbf{z}_t$, is able to emit actions a_t towards the world to modify its variant part $\mathbf{w}_t$. The world may be also characterized by invariant features h such as type, category, attribute, etc. that may have a global conditioning influence on world behavior (a pedestrian doesn't move the same way as a car for instance) and can be considered as a latent variable. The system is perceptual, and have access to sensory data $\mathbf{x}_t$, the raw source of information used to produce perceptual outputs $\mathbf{y}_t$, often a partial estimate either of world state $\mathbf{w}_t$ or of its invariant part h [2]. It is usual to consider those variables as random.

The internal perceptual system state $\mathbf{z}_t$ may be complex. It can encode the way sensory data is processed (extracted features, applied algorithm, etc.), a local memory that summarizes in a synthetic and informative form past sensory data, action and world state estimates, proprioceptive pose or location of the perceptual system in local coordinates, computing features (architecture, memory load, energy), indicators of task completion (belief, duration), etc. The nature of this internal state and the ways it is exploited differentiate the various models proposed in the literature.

The history of inner states, actions and observations is noted $\mathbf{\Phi}_t$ and collects all the contingent elements that are accessible by the perceptual system and are used as input data for decision or inference.

Tab. 3.1 summarizes the variables that can be exploited for active recognition models.

Tab. 3.1: List of variables potentially used in active perception models.

Variables	
t	Time
$\mathbf{x}_t$	Sensory data
a_t	Action
$\mathbf{w}_t$	World state (evolving part)
h	World state (invariant part)
$\mathbf{z}_t$	Internal state
$\mathbf{y}_t$	Perceptual output
$\mathbf{\Phi}_t = [\mathbf{x}_{1:t}, a_{1:t}, \mathbf{z}_{0:t}]$	History of past observations, inner states and actions

The objective of a perception system is to infer a user-relevant output $\mathbf{y}_t$ from input data. The *active* dimension says that input data is the combination of sensor outputs,

[2] In these notations, we follow the convention used in pattern recognition where $\mathbf{x}$ is the sensory input used to predict $\mathbf{y}$. In control theory, $\mathbf{y}$ is usually taken to be the measurement whereas $\mathbf{x}$ is the state.

actions and inner system state, $\mathbf{\Phi}_t$. Formally, perceptual inference follows an **I** pattern (Fig. 3.1) but with a complex temporal input: $\mathbf{\Phi}_t \rightarrow \mathbf{y}_t$. Here, the $i \rightarrow o$ symbolic expression simply means that the variable o on the right side has some causal relation with the variable i on the left side. This relation may be functional or random, i.e. one can make the hypothesis that posterior ($\Pr[o|i]$), likelihood ($\Pr[i|o]$) and joint distribution ($\Pr[i, o]$) can be defined. It is usual that inferences exploiting this type of random relation may come with an uncertainty representation (score, probability, correlation matrix, belief, etc.).

Classical types of perceptual outputs $\mathbf{y}_t$ are related to world states, and can be divided into two categories:

- World feature prediction: $\mathbf{\Phi}_t \rightarrow \hat{h}_t$. The goal is to provide a description of user-relevant invariant features – attribute, category, shape, location etc. Although the features are expected to be time invariant, their prediction may vary when new *actively* generated data is available.
- World state estimation: $\mathbf{\Phi}_t \rightarrow \hat{\mathbf{w}}_t$. The goal here is to predict several features of the changing world, objects or events, and potentially include the relation of the perceptual system itself with the world, for instance for ego-localization. This type of estimate is usually expected to provide measures of world features.

Tab. 3.2 summarizes the fundamental predictive functions involved in active perception.

Tab. 3.2: The two types of predictions.

Predictions	
$\mathbf{\Phi}_t \rightarrow \hat{h}$	World feature prediction
$\mathbf{\Phi}_t \rightarrow \hat{\mathbf{w}}_t$	World state estimation

What is specific to active perception with this formalization is the fact that the input contains actions, i.e. controllable features. The main question is therefore to find the best way to generate those actions: $\mathbf{\Phi}_{t-1} \rightarrow a_t$, i.e. to estimate the impact of a possibly random action on current perceptual task completion, the production of perceptual output from the sequence of sensory data and actions: $\mathbf{\Phi}_t \rightarrow \mathbf{y}_t$.

One unifying modeling alternative would have been to consider the perceptual output production as a specific *terminal* action. We keep the distinction between those two types of inference two in order to make more salient the separation between world interaction and perceptual output generation.

Tab. 3.3 summarizes the fundamental decision functions involved in active perception.

The design of the functions of Tab. 3.3 can only be done from modeling assumptions needed to predict the dynamic behavior of key variables when applying a given action:

Tab. 3.3: The two fundamental (random) decision functions of active perception.

Inference & decision	
$\mathbf{\Phi}_t \rightarrow \mathbf{y}_t$	Perceptual output generation
$\mathbf{\Phi}_{t-1} \rightarrow a_t$	Action generation

- The world: acting in the world may modify it or change its relation with the perceptual system. World dynamics depends on its history and on its invariant features. The world itself may be changing without any action on it, for instance if it contains mobiles.
- The perceptual system: its internal state can be modified sequentially typically to update its observation parameters (such as those depicted in Fig. 3.2), or its belief or estimated score measuring current task completion level.
- The sensory data: from a pure perceptual perspective, actions modify the way the perceptual system senses a potentially evolving world. Sensory data changes may result from actions on the world itself, but also from viewing conditions modifications, for instance by changing focal length, image filters, gaze direction, etc.

Tab. 3.4) summarizes the three different variables that may be impacted by an action.

Tab. 3.4: General dynamic models used in active perception.

Dynamic models	
$h, \mathbf{w}_{0:t-1}, a_t \rightarrow \mathbf{w}_t$	World dynamics
$\mathbf{\Phi}_{t-1}, a_t \rightarrow \mathbf{z}_t$	System dynamics
$h, \mathbf{w}_{0:t-1}, \mathbf{\Phi}_{t-1}, a_t \rightarrow \mathbf{x}_t$	Sensor dynamics

The practical implementation of these abstract models have been addressed in the literature with various semantics and simplifying hypotheses. Four different families of approaches have been proposed: utility based action selection, reward based optimal policy design and controlled random sampling. The following sections will describe their main features, concentrating on the question of invariant world feature prediction: $\mathbf{\Phi}_t \rightarrow \mathbf{y}_t = \hat{h}$.

Utility based action selection

The first, and probably dominant, formalization of active perception is to ask what the next "best" action given system history should be, i.e. the most "useful" to complete the task of accurately predicting the invariant feature h. The goal of this action is to generate a new input sensory data $\mathbf{x}_t$ expected to be useful for improving the quality of the estimate $\hat{h}_t$.

Let $\mathcal{U}_t$ denote the function measuring the utility of applying an action a at time t to be chosen from a set $\mathcal{A}_t$ given the history of the system $\mathbf{\Phi}_{t-1}$. Action generation can be rewritten as an optimization problem:

$$a_t = \underset{a \in \mathcal{A}_t}{\operatorname{argmax}}\, \mathcal{U}_t(\mathbf{\Phi}_{t-1}, a) \tag{3.1}$$

Most of the approaches that describe utility functions – not all – are embedded into a probabilistic framework expected to globally catch the contingent noise of sensing devices, but also to control the complexity of modeling and handling high dimensional sensory inputs.

A standard way to handle uncertainty is to make use of a Bayesian framework, i.e. a formal environment where posterior and prior probability distributions of useful variables are accessible and mechanisms to update their values when a new outcome is available.

The classical formulation of predicting world state from sensory data is to introduce a belief function b_t defined as a posterior distribution that plays the role of the system internal state $\mathbf{z}_t$:

$$b_t(h) = \Pr[h|\mathbf{\Phi}_t]$$

and to estimate its value as maximizing this belief

$$\hat{h}_t = \underset{h}{\operatorname{argmax}}\, b_t(h). \tag{3.2}$$

Other randomized estimators can be used instead of the maximum posterior.

The Bayesian framework is also interesting because it describes the internal system state dynamics $\mathbf{z}_t$ with a simple recurrent updating scheme using Bayes inversion formula:

$$\begin{aligned} b_t(h) &= \Pr[h|\mathbf{\Phi}_t] \\ &= \Pr[h|\mathbf{\Phi}_{t-1}, \mathbf{x}_t, a_t] \\ &= \frac{\Pr[\mathbf{x}_t|a_t, \mathbf{\Phi}_{t-1}, h]}{\Pr[\mathbf{x}_t|a_t, \mathbf{\Phi}_{t-1}]} \cdot \Pr[h|\mathbf{\Phi}_{t-1}] \\ &= \frac{\Pr[\mathbf{x}_t|a_t, \mathbf{\Phi}_{t-1}, h]}{\Pr[\mathbf{x}_t|a_t, \mathbf{\Phi}_{t-1}]} \cdot b_{t-1}(h) \end{aligned} \tag{3.3}$$

making posterior (or belief) at time $t-1$ a prior for the posterior at time t.

If the only useful feature is the maximum of eq.(3.2), the updating scheme can even be made simpler by considering that the denominator of eq. (3.3) is only a normalization factor:

$$b_t(h) \propto_h \Pr[\mathbf{x}_t|a_t, \mathbf{\Phi}_{t-1}, h] \cdot b_{t-1}(h) \tag{3.4}$$

where the notation $f(h) \propto_h g(h)$ means that $f(h) = \frac{g(h)}{\sum_{h'} g(h')}$.

Eq. (3.4) is simple but hides a latent complexity: the whole history $\mathbf{\Phi}_{t-1}$. A standard simplifying assumption is to consider that the probability of observing sensory data $\mathbf{x}$ is independent of the history and is only the consequence of the sampling action a_t. The belief updating equation becomes:

$$b_t(h) \propto_h p(\mathbf{x}_t|a_t, h) \cdot b_{t-1}(h) \tag{3.5}$$

where the conditional likelihood $p(\mathbf{x}_t|a_t, h)$ used as a recurrent multiplicative factor is the only needed feature that relates sensory experience to perceptual output in the model. If this likelihood can be reliably and efficiently estimated, for instance when action and sensory data spaces are finite or Gaussian distributed, we have a very compact model. The only remaining question is to find good utility functions able to exploit this model.

Several approaches that make use of information theory concepts have been proposed in the literature to justify the design of utility functions: entropy (Callari and Ferrie, 2001; Arbel and Ferrie, 2001; Defretin et al., 2010; Butko and Movellan, 2010), expected loss of entropy (Paletta and Pinz, 2000; Borotschnig et al., 2000), mutual information (Schiele and Crowley, 1998; Atanasov et al., 2014; Denzler and Brown, 2002), Kullback–Leibler divergence (Laporte and Arbel, 2006), free energy (Friston et al., 2017) to name the most popular ones.

(Croon et al., 2009) empirically compares several of those strategies on a viewpoint selection for 3D object recognition, which is often considered as a prototypical problem for active recognition problem, although online feature selection has also been evaluated (Potthast et al., 2016). There is no clear winner among the proposed approaches in the literature, although the mutual information based action selection seems to perform a little bit better on their experiments. More recently, (Daucé, 2018) provides a theoretical comparison of several utility functions and shows that they can be considered as either innovative – they tend to generate actions that may potentially contradict current prediction state – or conservative – they favor action that are likely to confirm it. This distinction, which can be considered as a version of the exploration/exploitation principle of reinforcement learning (François-Lavet et al., 2018), explains the impact of modeling errors (mainly the estimation of $p(\mathbf{x}|a, h)$) on prediction performance, and argues in favor of strategies mixing several utility functions. A utility function combining an "innovative" predicted entropy and a "conservative" current best class likelihood has been proposed in (Defretin et al., 2010).

Most of the utility functions proposed in the literature rely on the sequential Bayesian updating rule of Eq. (3.5), and on the assumption that it can provide a good estimate of the posterior distribution that can be used for further discrimination. In (Herbin,

2014), I proposed a sequential strategy that rely on a different action selection principle that maximizes the expected number of hypotheses that can be rejected.

Sequential hypothesis rejection strategies (Herbin, 2014)

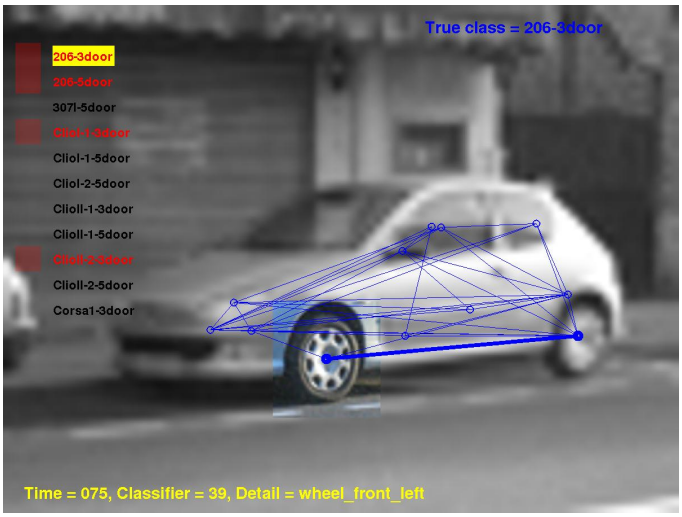


Fig. 3.3: Example of foveated regions during the active recognition process. On the upper left side of the figure, active hypotheses have red color, rejected are in black, best current hypothesis has yellow background. The graph depicts the sequence of focus locations. More examples in the video http://youtu.be/51IbY3A0yC4.

The recognition task is considered as a sequential hypothesis rejection process, starting from a set of possible hypotheses or classes Ω_0 and iteratively reducing the set of active hypotheses by applying a sequence of rejection tests.

The proposed algorithm follows a classical active recognition scheme: at each instant t, the system selects an action able to generate a new piece of information from the environment, combines it with past acquisitions to improve the completion level of the recognition task and updates the set of active hypotheses Ω_t, i.e. hypotheses that are not believed to be false. In this formulation, the internal state $\mathbf{z}_t$ of Tab.3.1 is reduced to Ω_t. The utility function is defined as the average rejection capacity of a test able to discard a given subset of hypotheses.

This type of approach is a contrario, in the sense that it does not try to isolate the most likely hypotheses, but discards iteratively the less likely. It has been validated on a problem of fine grained car recognition where actions consist in focusing on several details of the field of view with high resolution (Fig. 3.3).

Reward based optimal policy

The previous section described a "next best action" strategy incrementally modifying a predictive belief about the invariant part of the world h as internal state, using several types of uncertainty representation (posterior, votes on rejected hypotheses) summarizing the history of active sensory data acquisition. The value of an action was evaluated as the one step uncertainty reduction on this state using various types of utility functions and concepts of uncertainty.

Another common action selection strategy is to extend the idea of uncertainty reduction to a generalized idea of reward: an action is good if it is rewarded a high uncertainty reduction, but also if it is not too costly, risky, time consuming, etc. A reward based formulation allows a better modeling flexibility and expressiveness. The objective of a global action selection strategy – usually called policy in the literature – can then be interpreted as *maximizing* a global reward that characterizes the current task completion level.

Another advantage of a reward-based formulation is a clearer formal separation between modeling and control, making easier the introduction of dynamic models (see Tab. 3.4). This is one of the reasons of the popularity of this family of approaches in robotics, often with an emphasis on controlling physical dynamical systems rather than on expressing perception as a dynamical system.

In its most general form, the reward at time t, r_t, is a random real function of potentially all the variables involved in the active perception process:

- the action a_t: several actions can be more costly than others, or more adapted to complete the task;
- the sensory data $\mathbf{x}_t$: the reward is a quality measure of its information content;
- the internal state $\mathbf{z}_t$: it encodes the state of completion of the current perceptual task;
- the history $\mathbf{\Phi}_{t-1}$: the usefulness of a new experience can be relative to the previous state (Markov hypothesis) or to the whole past experience, for instance to avoid revisiting the same locations;
- the world state history $\mathbf{w}_{0:t}$: prediction quality depends on outer world motion or evolution features and on its complexity;
- the world invariant state h: several features are more rewarding than others, more or less easily detectable for instance.

The objective of an action selection policy is to choose the action that maximizes an expected cumulated or average reward, either with finite or infinite horizon:

$$R_t = \mathrm{E}\left[\sum_{k=1}^{L} \gamma^k r_{t+k+1}\right]$$

where $\gamma < 1$ is a discounted factor that models the impact of long term rewards.

The literature addressing the resolution of this family of problems is huge, since the seminal work of Bellman that defined the Markov Decision Process model (MDP) and Dynamic Programming to solve it (Bertsekas, 1995). Several variations have been proposed, introducing for instance multiple objectives or multi-valued rewards (Roijers et al., 2013), or more complex processes such as Partially Observable Markov Decision Processes (POMDP) (Ross et al., 2008), leading to a whole domain devoted to planning (Ghallab et al., 2004; LaValle, 2006).

The exact resolution of reward-based action selection policies assumes the availability of reliable dynamical and behavioral models (Tab. 3.4). Typically, the most usual model is to make a Markov assumption on world $(h, \mathbf{w}_{t-1}, a_t \rightarrow \mathbf{w}_t)$ and/or internal $(\mathbf{z}_{t-1}, a_t \rightarrow \mathbf{z}_t)$ state dynamics.

When dealing with complex situations, typically with large internal state spaces or high dimensional sensory data, reliable dynamical models are difficult to obtain. One usual proposed solution to somehow to get rid of the modeling step is to directly learn the action selection policy exploiting rewards as reinforcement signals. *Reinforcement learning* (Bertsekas and Tsitsiklis, 1996; Sutton and Barto, 1998) has become a major technique in AI, with remarkable and spectacular successes (Silver et al., 2018; Vinyals et al., 2019), and has evolved to acclimate formal objects such as deep networks (François-Lavet et al., 2018), or new formal optimal settings such as multi objective (Liu et al., 2014), preference (Wirth et al., 2017) or imitation (Osa et al., 2018).

Many of the applications of reward-based policy design have target objectives where perception appeared as a means to complete another task (navigation, localization, visual servoing, path planning, gaming) not the final outcome of the process. The main question regarding APES's is to adapt this framework to perceptual systems, i.e. systems whose main objective is to produce signs about or measures of the world. A few studies have addressed the question of budgeted or controlled feature acquisition for data classification (Karayev et al., 2012; Dulac-Arnold et al., 2013; Weiss and Taskar, 2013; Nan et al., 2015; Huang et al., 2017b; Shim et al., 2018; Janisch et al., 2019) often optimized using reinforcement learning.

Active sequential testing

The utility function based family of active perception approaches puts the emphasis on finding the best next action and makes use of a Bayesian incremental belief updating step relying on the availability of a conditional sensory data likelihood model $p(\mathbf{x}|a, h)$ (Eq.3.5). Their performance depends on the quality of this likelihood that plays the role of a predictor using Bayes inversion formula, which is a powerful tool when the likelihood is reliable, but may lead to unstable or erroneous beliefs

if badly estimated, for instance with high dimensional input sensory data such as images. Another source of instability of a Bayesian approach is the normalization step over the set of possible hypotheses that assumes that all the possible prediction values are known beforehand.

Another active strategy dealing with uncertainty, more "frequentist" in spirit, is to accumulate pieces of evidence in a statistical process that would eventually converge to a good global likelihood estimate of each possible perceptual output.

Sequential decision strategies have a long history in statistics since the early work of Wald (Wald and Wolfowitz, 1948) (see (Naghshvar and Javidi, 2013) for recent developments or (Tartakovsky et al., 2014) for a recent account of sequential analysis). In computer vision, sequential decision processes have been implemented in the form of coarse to fine strategies (Blanchard and Geman, 2005; Fidler et al., 2010; Gangaputra and Geman, 2006) or cascade-like structures (Viola and Jones, 2001) applied to categorical object detection rather than classification. The main objective of a sequential strategy is to control the false alarm rate, and as a secondary objective the variability of object features in the target category. Fewer studies have addressed the question of object classification with a sequential decision approach.

Basically, a sequential testing approach updates the likelihood (or log-likelihood) conditionally to each hypothesis h:

$$\begin{aligned} l_t(h) &= \Pr[\mathbf{\Phi}_t|h] \\ &= \Pr[\mathbf{x}_t, a_t, \dots \mathbf{x}_1, a_1, \mathbf{x}_0|h] \end{aligned} \tag{3.6}$$

and decides when to stop acquiring new data and how to choose the most relevant hypothesis. The main differences with the Bayesian approach is that we work with likelihoods instead of normalized posteriors or beliefs, and that actions are often random, i.e. they are sampled from a probability law, static or adapted to the current likelihood distribution.

The global likelihood of Eq. (3.6) can be written also using a recurrence similar to Eq. (3.4):

$$\begin{aligned} l_t(h) &= \Pr[\mathbf{x}_t, a_t, \mathbf{\Phi}_{t-1}|h] \\ &= \Pr[\mathbf{x}_t, a_t|\mathbf{\Phi}_{t-1}|h] \cdot \Pr[\mathbf{\Phi}_{t-1}|h] \\ &= \Pr[a_t|\mathbf{\Phi}_{t-1}] \cdot \Pr[\mathbf{x}_t|a_t, \mathbf{\Phi}_{t-1}, h] \cdot \Pr[\mathbf{\Phi}_{t-1}|h] \\ &= \mu_t(a_t, \mathbf{\Phi}_{t-1}) \cdot \Pr[\mathbf{x}_t|a_t, \mathbf{\Phi}_{t-1}, h] \cdot l_{t-1}(h) \end{aligned} \tag{3.7}$$

where μ_t is the action law at time t, i.e. the free control of the process that depends on the accumulated observations $\mathbf{\Phi}_{t-1}$.

Several modeling hypotheses must be made to be able to exploit the decomposition of Eq. 3.7 on the functional structure of the action law μ_t – how to summarize

past experience $\mathbf{\Phi}_t$ and how to (randomly) choose the wright action – and on the controlled data acquisition $\Pr[\mathbf{x}_t|a_t, \mathbf{\Phi}_{t-1}, h]$. The final hypothesis prediction can be done using maximum likelihood decision: $h^* = \text{argmax}_h\, l_t(h)$. This likelihood formulation also makes possible the exploitation of statistical results, typically consequences of the law of large numbers. This is what has been done in previous work, under two modelling hypotheses: independent identically distributed (i.i.d.) (Herbin, 2003; Herbin, 2004) and Markov chain (Herbin, 1996; Herbin, 1997b; Herbin, 1998; Herbin, 2002).

Asymptotics of random sampling strategies for object recognition

I.I.D. case

The original idea was to model object of interest as a stationary family of histograms ($\Pr[\mathbf{x}_t|a_t, \mathbf{\Phi}_{t-1}, h] = p(\mathbf{x}|a, h)$), and exploit this model for inference when conditionally sampling the modalities a of each element of the family. Testing is based on accumulating evidences on the discrimination between all-pairs of hypotheses h and h'. For a stationary sampling law $\mu(a)$, and if hypothesis h is the true one, we have, thanks to the law of large numbers:

$$\lim_{t\to\infty} \frac{1}{t}[\log l_t(h) - \log l_t(h')] = \sum_a \mu(a) \sum_{\mathbf{x}} p(\mathbf{x}|a, h) \log \frac{p(\mathbf{x}_t|a_t, h)}{p(\mathbf{x}_t|a_t, h')} \qquad (3.8)$$

Hypothesis h will be declared true if all values of (3.8) are positive.

The convergence of a test exploiting the sign of (3.8) can be refined thanks to tools of large deviations (Dembo and Zeitouni, 1998). Indeed, it is possible to compute the exact logarithmic convergence rate to zero of deciding hypothesis h' while hypothesis h is true:

$$\lim_{t\to\infty} -\frac{1}{t} \log \Pr[\log l_t(h) \leq \log l_t(h')|h] = \rho(h, h') > 0 \qquad (3.9)$$

These number can be used either as a loss to be optimized (Herbin, 2003), or as a way to calibrate the number of samples needed to reach a given performance level (Herbin, 2004).

Eq 3.8 assumes that the conditional probabilities $p(\mathbf{x}|a, h)$ are exact. It is possible to modify the equations by replacing directly the log-ratio $\log(p(\mathbf{x}_t|a_t, h)/p(\mathbf{x}_t|a_t, h'))$ by a more robust and discriminative quantity (Herbin, 2004) or by exploiting multiple votes to secure the difference of distribution supports between hypotheses (Herbin, 2003).

This framework has been applied to the rotational invariant recognition of noisy images, where objects are represented by statistics of pairs of point configurations (Fig. 3.4).

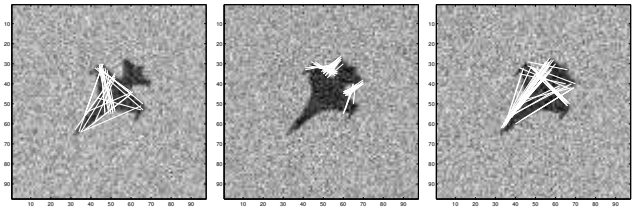


Fig. 3.4: Three examples of bipoint configurations used to define a rotation invariant representation of a shape and that can be exploited from a random sampling strategy.

Markov chain case

The i.i.d. active sampling case relies on a model that can only encode or assume independence of observed input data, but is able to handle multiple modalities. The Markov chain formulation adds to the previous case a model of multiple observation inter-dependence, and can encode more structural dimensions through their underlying transition graph.

With a Markov hypothesis, Eq. 3.7 can be rewritten as:

$$l_t(h) = \mu(a_t) \cdot \Pr[\mathbf{x}_t | a_t, \mathbf{x}_{t-1}, h] \cdot l_{t-1}(h) \tag{3.10}$$

where the probability transition $\Pr[\mathbf{x}_t | a_t, \mathbf{x}_{t-1}, h]$ summarizes all is known about the world.

Similarly to the i.i.d. case, finding the most accurate hypothesis can be done by accumulating evidence through the conditional likelihood and infer the best hypothesis according to their values. The asymptotics of a maximum likelihood test also follows a Large Deviation principle:

$$\lim_{t \to \infty} -\frac{1}{t} \log \Pr[\log l_t(h) \leq \log l_t(h') | h] = \rho(h, h') + \tau(h, h') > 0 \tag{3.11}$$

where $\rho(h, h')$ is a function of the probability ratios between hypotheses h and h', and $\tau(h, h')$ depends on their transition graph differences, the graph of edges whose transition probabilities are strictly positive for both hypotheses. The convergence rate (3.11) gives also a similarity measure (not a distance, however) between two Markov chains exploiting there relative probabilities and structures in a single measure.

The active sampling scheme under Markov chain modeling has been applied to propose a well funded definition of probabilistic aspect graphs of 3D

objects(Herbin, 1996; Herbin, 1997b; Herbin, 1998), and to define a similarity measure between co-occurrence matrices of textures seen as controlled Markov chains (Herbin, 2002).

Active perception use cases

The formal active perception framework described above can be used to solve several use cases. Before presenting them, let's begin by describing a typical situation where an active approach may be useful to produce perceptual outputs.

A generic scenario: dynamic scene description for navigation

In this scenario, the role of the perceptual system is to provide a complete description of the dynamic environment content in which an autonomous mobile platform evolves to plan its trajectory (Fig. 3.5).

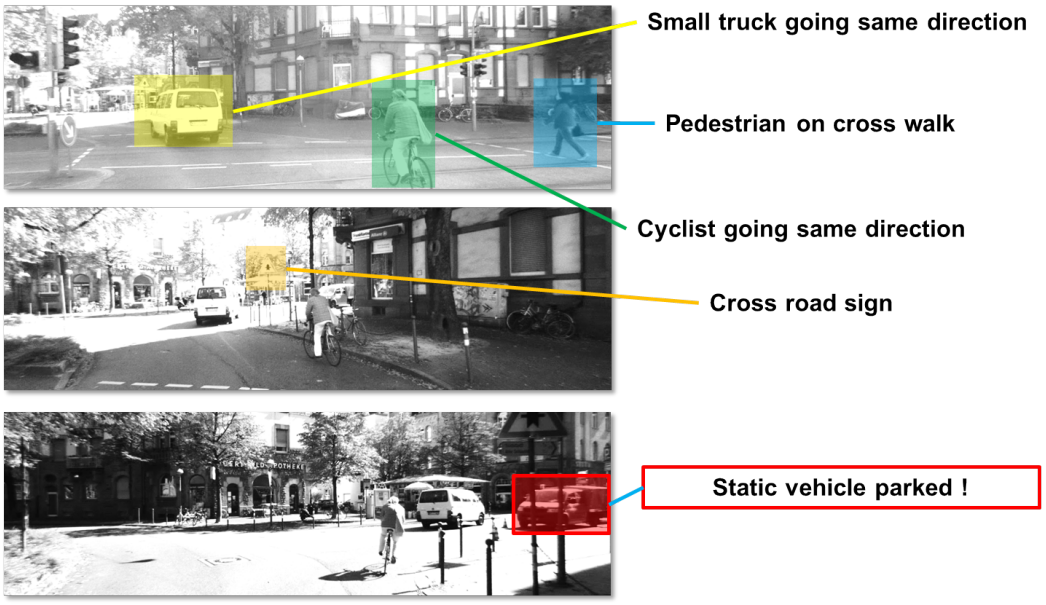


Fig. 3.5: Examples of outputs that are expected from the active perceptual system: object detection and classification, moving direction estimation, and potential obstacle nature.

The description has two roles: to anticipate the presence of obstacles for navigation and to improve the ego-localization of the platform relatively to environmental objects. It relies on capabilities of scene reconstruction, pose estimation, visual detection, tracking, classification, re-identification and prediction.

Multiple agile sensors such as pan tilt zoom (PTZ) cameras or lidars can be used to control the flow of information by focusing on several areas of the field of view or modifying the applied sensor to improve the level of resolution or adapt the physical principle needed to characterize the objects.

Actions on this scenario can be of two types:

- Recruit, initialize and apply an algorithm to update the current state of description;
- Change line of sight, focal length or physical principle of sensors.

All parts of the visual field are not simultaneously observable or not with the same level of resolution. A key feature of this use case is the exclusivity of several algorithms: typically, identification of objects (e.g. road signs) requires high resolution and narrows the field of view, preventing the detection or updating of other moving objects in the whole scene that may have an impact on platform navigation (obstacles). Conversely, when the field of view is large, many potential objects may be detected but not characterized with enough precision or confidence.

The content of the whole dynamic scene cannot be described with a uniform level of detail and confidence: when a new observation is available, each object state and attribute can be either updated if it can be associated to a new measurement or another algorithm output, or simply estimated from the past using a predictive model. What is maintained by the system is therefore a set of predictors about object location, pose, speed, nature, behavioral state, etc. that can be selectively updated from controlled observations and algorithm recruitment.

Predictions are not independent: for instance, a reliable object characterization (whether the object is a car, a pedestrian or a cycle, for instance) may condition its kinematics and its behavior prediction.

Each prediction may be associated with any feature able to represent the current description quality that is meaningful to the user: a confidence score, an uncertainty level, a utility value, etc. The overall objective of the perceptual predictive system is to ensure a good – minimal, average or final – value of this quality.

This rather general use case exemplifies several functions that have been addressed, often independently, in the active perception literature. It also extends the repertoire of actions customarily proposed in robotics (Fig. 3.2) to algorithmic or processing alternatives. The rest of this section describes how some of the functions that contribute to the above scenario can be solved using an active perception perspective.

Multi-view object recognition

Three dimensional object recognition has been one the first functions where an active perception approach has been applied (Arbel and Ferrie, 1996; Herbin, 1996; Dickinson et al., 1997; Paletta and Pinz, 2000; Borotschnig et al., 2000). It is expected to allow the observation of objects from different points of view in order to increase decision reliability with respect to a set of plausible hypotheses. The role of actions is twofold: exploring the viewing space in order to discover more

discriminating or interpretable views, and sample new data that may confirm/infirm current beliefs about object nature.

Most approaches in this field exploit mechanisms of entropy reduction or mutual information with respect to a set of known conditional laws (Denzler and Brown, 2002; Laporte and Arbel, 2006; Deinzer et al., 2006; Deinzer et al., 2009; Defretin, 2011; Atanasov et al., 2013) to choose the next best view (Roy et al., 2004). The final quality of the recognition depends on the discriminating capacity of the image characteristics extracted, the number of points of view exploited and the mechanisms of combination of the perceptive information. Most approaches use heuristics to enhance the independence of different viewpoints.

A large body of work has also been interested in view planning for object reconstruction (Scott et al., 2003; Chen et al., 2008; Remazeilles and Chaumette, 2007; Wenhardt et al., 2006; Wenhardt et al., 2007; Atanasov et al., 2014; Devrim Kaba et al., 2017), a correlated task but where the objective is more directed towards geometric precision and coverage than discrimination.

Object search

Another well studied function exploiting an active information gathering strategy is to control a mobile sensor to efficiently search for objects in a scene. The main idea in this problem is to incrementally build a global representation of the scene that may be able to condition the probability of objects being present in certain locations – typically by modeling object or landmarks co-occurrence – in order to infer the next best view (Andreopoulos and Tsotsos, 2008; Sommerlade and Reid, 2008b; Aydemir et al., 2013; Aydemir and Jensfelt, 2012; Velez et al., 2011; Velez et al., 2012).

Object search, due to its optimal formulation, is one of the few active perception use cases where theoretical results have been produced. Tsotsos and its coworkers in a series of works: (Ye and Tsotsos, 2001) justify the optimality of active perception for object search; (Andreopoulos and Tsotsos, 2009; Andreopoulos, 2009; Andreopoulos and Tsotsos, 2012; Andreopoulos and Tsotsos, 2013) investigate the impact of input noise, occlusion, and the VC-dimensions of the related representation classes on localizing all objects present in the search region, under finite computational resources and a search cost constraint. (Karasev et al., 2012) exploits a simple model that includes uncertainty due to occlusion, scaling, and other types of nuisance processes and introduces the idea of control authority levels to describe various active perception strategies and to derive theoretical performance bounds.

Dynamic scene content description

Because active perception is a temporally unfolded decision process, it makes possible the interpretation of dynamic environment with moving entities such as tracking, behavior prediction, dynamic pose estimation, etc.

Most of the studies have concentrated on using controllable sensors such as Pan Tilt Zoom cameras (PTZ) that are able to dynamically adapt to evolving situations.

A first series of studies are more control oriented, where perception is mainly used to compute an error between an actual observation and a prediction from an internal model to compute the control, i.e. the new sensor features. This problem can be considered as a particular case of visual servoing (Chaumette and Hutchinson, 2006; Chaumette and Hutchinson, 2007). The main active feature to control in those studies was the focal length for problem of object tracking (Tordoff and Murray, 2004; Tordoff and Murray, 2007; Sommerlade and Reid, 2008a).

Another series of studies have concentrated on using a single PTZ camera, or master-slave configuration, to optimally track multiple moving objects using fixed scheduling policies (Costello et al., 2004), internal valued graph representations (Bimbo and Pernici, 2006; Bagdanov et al., 2005; Bagdanov et al., 2006; Melman et al., 2018) or information theoretic formalism (Sommerlade and Reid, 2008b; Sommerlade and Reid, 2008c; Salvagnini et al., 2015).

Very few studies, however, have addressed the dynamic evolution of perceived features for scene interpretation, for instance by zooming on specific parts to improve accuracy. One of those is (Salvagnini et al., 2013) which describes a person re-identification from focused body parts.

Multiple Sensor surveillance

Timely allocating the right sensor, with the right viewing features, is another well studied problem that can fit under the idea of active perception. Compared to the setting of the previous paragraph – a single but controllable sensor – the complexity to handle is also carried by the current system state: the number of sensors and their potential combined contributions to accomplish a global interpretation task.

Multiple Sensor surveillance is multi-objective by nature and requires trade-offs between spatial coverage and quality of detection/recognition/tracking of mobile entities, between accuracy and exhaustiveness.

The idea underlying an actively configurable sensor network is to compensate for limitations due to the limited number of available resources and to constrained geometric configurations. The collection of sensors cannot observe a whole scene in every details due to occlusion, to blind spots and to spatial area coverage boundaries at a given resolution.

The typical problem of a sensor network is multiple target tracking, i.e. the capacity to predict at every moment the presence and location of all targets of interest in the monitored area. Each sensor has to determine whether it should contribute to the tracking of a single target with high resolution, of a group with lower resolution or search for a new target.

The literature in this domain is huge, and has been boosted by the availability of cheaper camera networks and a political interest for security reasons in various countries. Many general surveys (Song et al., 2011)(Wang, 2013)(Natarajan et al., 2015)(Piciarelli et al., 2015)(Liu et al., 2016a) or targeting more focused questions such as coverage (Mavrinac and Chen, 2013) or occlusion (Mittal and Davis, 2008) are available. Multi robot can be considered as an extension to sensor networks where each sensor may have the full 6 Degrees Of Freedom (Bakhtari et al., 2009; Khan et al., 2018; Best, 2019) for instance when hosted by a drone (Xiao et al., 2017).

Re-identification, i.e. finding in a scene an object or a person already observed is a specific perceptual function that is mandatory in multi sensor systems to ensure interpretation continuity and coherence. The causes of loss of visibility or rupture of the quality of the interpretation are numerous (occlusions, sudden change of illumination, exit of the field of view, loss of resolution, blur ...) making re-identification a difficult problem (Quo et al., 2007; Hamdoun et al., 2008; Guinet, 2008; Guo et al., 2008; Arth et al., 2007; Leotta and Mundy, 2009; Tsin et al., 2009; Zheng et al., 2016).

Choosing the right sensors at the right moment can be formalized as a *sensor management* problem (Hero et al., 2007), i.e. to "seek a policy for determining the optimal sensor configuration at each time, within constraints, as a function of information available from prior measurements and possibly other sources" (Hero and Cochran, 2011). Algorithms of this domain rely on formal tools similar to those used in planning (MDP, POMDP, Dynamic Programming). In a recent series of work, (Satsangi, 2019) discusses application of these techniques to multiple sensor person tracking, including a PAC analysis (Valiant, 1984) able to integrate model uncertainty in the decision policies. Most of the models are applied to low dimensional measure or state spaces and to state estimation rather than to environment behavior or content interpretation.

Attention: soft or hard

Attention, i.e. the time dependent selection or modulation of sensory data, is a paradigmatic feature of active perception. The importance of attention in natural vision has already been discussed in a previous section (pg. 41). The separation between engineering and natural models[3], however, is not strict: the artificial

[3] http://www.scholarpedia.org/article/Computational_models_of_attention

intelligence, robotics or computer vision literature proposes models relying on foveated sensors, i.e. that unevenly sample the field of view like the retina (Wang and Bovik, 2001; Weber and Triesch, 2009) or interested in predicting gaze when observing a given scene, i.e. the sequence of informative looking locations (see Fig. 2.15) (Das et al., 2017a; Wloka et al., 2018). The existence of attention is also often justified as a necessary condition for managing the complexity of input sensory data with limited computing resources, both for engineering purposes of for natural perception modeling (Tsotsos, 2011; Borji and Itti, 2013). Here we focus on how attention has been considered in formal models: as a functional pattern to mimic, as an interpretable phenomenon or as an efficient algorithmic principle

A first pregnant usage of attention is inspired by its "overt" features: focus, vergence (Krotkov and Bajcsy, 1993), gaze control, etc. The idea of attention is exploited as a functional pattern that guides perceptual system design and implementation. This trend is well developed in robotics with a specific care about the control of agile sensors such as monocular or stereo PTZ cameras (see previous sections and (Frintrop et al., 2010; Chen et al., 2011)).

Overt attention, which is mainly reduced to the detectable phenomenon of focusing on several part of the field of view, reveals something about how sensory data are exploited during the perceptual process. We have seen in the previous chapter that the pattern of fixation point locations may be conditioned by the current task to be completed (DeAngelus and Pelz, 2009) (Tatler et al., 2010) (Borji and Itti, 2014), although the informational content of such pattern is not clear (Greene et al., 2012). We will see in the next chapter (pg. 127) how attentional features can be used as supplementary explanations or justification of perceptual outputs, i.e. a kind of confidence measure about the quality of perceptual production.

More recently, especially since the advent of the deep learning era and the revival of recurrent networks (Wang and Tax, 2016), attention has been considered as an efficient principle to solve various tasks, ranging from natural language processing (NLP) (Bahdanau et al., 2014; Ma et al., 2018c; Young et al., 2018) to visual object recognition and detection (Ren et al., 2015; Sermanet et al., 2015; Yoo et al., 2015; Ren and Zemel, 2017; Fu et al., 2017; Zheng et al., 2017), visual tracking (Choi et al., 2017; Wang et al., 2018a; Yang and Chan, 2018; Pu et al., 2018; Yun et al., 2017; Yun et al., 2018; Luo et al., 2018; Luo et al., 2019; Zhang et al., 2019b), action recognition (Das et al., 2019), visual question answering (Xu and Saenko, 2016; Lu et al., 2016; Anderson et al., 2018a), captioning (Xu et al., 2015; You et al., 2016; Hossain et al., 2019), person re-identification (Lan et al., 2017; Lin et al., 2019) or visual grounding (Deng et al., 2018).

Formally, attention has been implemented using two different ways to manage resources: soft or hard. The soft way consists in filtering, weighing or modulating the useful parts of the input sensory data, or of inner latent structures; the hard way

selects the useful parts, i.e. discards the resources that have no useful informational role to complete the task.

Most of the proposed algorithms in recent literature belong to the first category, and introduce attention by weighing the input signal through saliency maps (Jaderberg et al., 2015), internal feature channels (Dumoulin et al., 2018) or both (Wang et al., 2017; Woo et al., 2018). Dense multiplicative weighing or gating, a common operation in many deep network architecture, can be interpreted as an abstract soft attention step and is used in several varieties of networks (LSTM, Memory Networks (Sukhbaatar et al., 2015), Neural Turing Machines (Olah and Carter, 2016), FILM (Dumoulin et al., 2018), etc.). Many approaches that exploit soft attention as an internal computing principle fall under the encoder/decoder architectural paradigm and encode input data either sequentially using recurrent networks, especially for NLP applications (Chaudhari et al., 2019; Galassi et al., 2019), or using a more global strategy able to encode efficiently long term dependencies (Vaswani et al., 2017).

Hard attention, i.e. the selection of resources – data, computation, features, etc.– among a given repertoire, has been less studied. This can be explained by the fact that selection often leads to nondifferentiable expressions that prevent gradient based optimization. The main application domain where selective attention has been exploited is object or entity detection in high dimensional data such as images or video where potential locations are ranked and *proposed* to further processing steps. The idea of region proposal is an old strategy used to control the complexity of detection by reducing the number of locations to evaluate (Uijlings et al., 2013; Cheng et al., 2014; Zitnick and Dollár, 2014). Modern deep learning approaches have proposed architectures able to integrate in their pipe-line such selection as an intermediate step (Ren et al., 2015) or by mimicking multiple resolution foveation (Mnih et al., 2014; Gao et al., 2018; Li et al., 2019).

Interactive perception as a solution

Since the seminal papers at the end of the 80's (Bajcsy et al., 2018), active perception has been addressed in a rather large amount of works, especially in robotics (Chen et al., 2011; Patten, 2016; Best, 2019) where time dependent process and control are fundamentals, but also for "pure" perceptual tasks as has been described in the previous section.

One of the main motivations and inspirations for introducing active perception features in artificial intelligence models has been to replicate the performance of natural perception, which is fundamentally active. This is also why this approach has been continuously addressed, and questioned, in artificial intelligence, although with a rather moderate volume of research activity.

We discuss here the conditions under which an active perception approach seems to be useful, even necessary, for artificial perception systems, what are the avenues for improvement and why this subject still deserves to be studied and is a central concern of autonomous perception.

Why is active perception useful to perceptual systems?

To make the system adaptive. Perceiving the world is experiencing contingency. A perceptual system, due to the current limitations of its resources, cannot forecast or predict all aspects and dimensions of world features in every context: it may not be readily adapted to the situation. The world itself may have a configuration where the interesting features are hidden or not accessible with enough reliability, requiring a modification of the perceptual system state (pose, viewpoint, sensing modality, memory update, computing architecture, algorithm, etc.) [4]. The fundamental principle of active perception is to *purposely* modify its available and controllable resources to optimally complete the current task. Considered this way, an adaptive perceptual system is necessarily active.

Because resources are limited. All systems, natural or artificial, have finite resources (time, energy, field of view, resolution, sensitivity, etc.). Active perception, as already mentioned, can be seen as a way to get around these limitations by temporally unfolding the usage of these constrained resources given a way to combine them.

Because perceptual tasks are complex. Active perception, especially vision, was shown to be interesting because a number of problems that are ill-posed for a passive observer (shading and depth computation, shape from contour, shape from texture, and structure from motion) are simplified when addressed by an active observer (Aloimonos et al., 1988). On more high-level tasks, (Tsotsos, 1992; Ye and Tsotsos, 2001) showed that active vision has good optimal properties for object search. No such result exists however for simple semantic tasks such as single object or global scene classification – which may not require active features. For more complex tasks such as question answering or captioning where multiple semantic levels are involved, active or attentional vision is clearly a relevant option.

To make a perceptual system versatile. A complex scenario like the one illustrated in Fig. 3.5 requires the allocation and prioritizing of several sub-tasks and objectives that may vary according to user priorities and world events. A relevant perceptual system for this kind of situation must be versatile in that it must *actively* recruit the current goals and subgoals it pursues.

[4] "The non-invertibility of nuisances such as occlusion and quantization induces an "information gap" that can only be bridged by controlling the data acquisition process." (Soatto, 2013)

What are the remaining problems of active perception?

Active perception has been a subject of interest for a long time in AI and, as the previous section has argued, is still worth addressing. It remains however difficult to develop in practice and can be only assessed in limited operating domain. Perceptual systems able to provide reliable dynamic scene description as depicted Fig. 3.5 are not available yet. This sections discusses several key remaining problems of active perception. More focused research actions will be presented at the end of this chapter.

Self-assessment and uncertainty.

One of the main operating principles of active perception is information gain through sequential sensing and computation actions, often expressed as uncertainty reduction (see the short presentation of utility-based approaches pg. 66). But how express or estimate uncertainty?

Several authors speak of aleatoric and epistemic uncertainties to distinguish what comes from the environment and sensors, and what is the consequence of modeling errors (Der Kiureghian and Ditlevsen, 2009; Kendall and Gal, 2017) and propose a posteriori estimation methods by random perturbation (MC-drop out). The question of calibrating predictors, i.e. making the predictive output score close to the true likelihood, has been addressed recently by several studies and applied to deep networks (Lakshminarayanan et al., 2017; Guo et al., 2017; Malinin and Gales, 2018).

One current way to circumvent the lack of knowledge about uncertainty origin is to avoid its estimation as an intermediate product, and directly estimate the action law by rewarding good trials, usually by reinforcement learning (see pg. 70). However, learning is not the solution to all problems: it is computationally costly and hard to control, it requires real world instantiation or surrogate simulation that brings other type of noise, and in practice can only be applied to small scale active perception problems.

(Gallos and Ferrie, 2019) suggest that a lot can be done already using better uncertainty estimation (Gal and Ghahramani, 2016) and classical utility based next best view prediction. Uncertainty estimation however remains difficult with deep networks and depends on learning database biases (Ovadia et al., 2019).

One of the reasons why an active approach has not been considered mandatory in perception models is perhaps due to the lack of expressiveness of uncertainty formal representation and self-assessment: the recurrent equation 3.5 that is expected to summarize current world experience using conditional probabilities is too elementary to deal with heterogeneous levels of knowledge and complex objectives.

One interesting avenue of research is therefore the improvement of uncertainty representation, combination and estimation for sequential decision processes. To achieve this objective, it may be necessary to escape the framework of statistical information theory which provides powerful and reliable formal tools, but has difficulty in representing and controlling the specificity of the various sources of error.

Learning and modeling.

A machine learning step, typically using deep networks, is now compulsory in artificial perceptual system design to achieve robust and large usage domain performance, at least for several functional components. When applied to the complex and sequential schema of active perception, questions such as What can/must be learned? What can/must be modeled and optimized using other means or knowledge source? do not have clear answers. From a robotics perspective, for instance, deep learning raises specific challenges due to the impact of perceptual output to agent efficiency and survival (Sünderhauf et al., 2018; Kojima and Deng, 2019). One straightforward option is to use learning to provide instantaneous prediction from raw sensory data and integrate the result in a sequential decision loop: this schema implies that prediction is able to provide an uncertainty estimation of its result, which may not be an easy task (see previous paragraph). A second extreme strategy is to learn everything (sensory prediction module and action law) in an end-to-end way (Malmir et al., 2017; Jayaraman and Grauman, 2018; Yang et al., 2019). This solution is appealing as it transfers the burden of ruling the interaction and collaboration between system components to a global learning phase; however, it requires the availability of good learning databases, and it also makes system behavior less intelligible for monitoring or debugging (the question of intelligibility will be addressed in chapter 4).

Evaluation.

Assessing the relevance of an active perception approach poses several problems. When activity is only considered as a feature of the algorithm implementing a given function – i.e. a “static” algorithm may also be able to implement it – performance evaluation is not specific. This is typically the case for attentional algorithmic approaches of images or videos.

Specific problems however arise when actions modify the physical relation of the system with the world, when the perceptual system is embodied. As (Bajcsy et al., 2018) state when evoking the usage of active cameras, "The design and implementation of robotics systems which embody the basic prerequisites for Active Perception is still hard. [...] There is no commercially available control of a camera system, namely of the focus, aperture and field of view, though there are commercially available pan and tilt controllers", and suggest to address this problem through a cyber-physical system perspective (Sztipanovits et al., 2011; Alur, 2015).

The difficulty of developing and fully evaluating real active perception systems can be mitigated by the exploitation of simulated and controllable visual environments which has been recently proposed (Ammirato et al., 2017; Ammirato et al., 2018; Xia et al., 2018; Anderson et al., 2018b; Yan et al., 2018; Savva et al., 2017; Savva et al., 2019; Brodeur et al., 2018) and applied to problems of embodied question answering (Gordon et al., 2018; Das et al., 2018; Anderson et al., 2018c). The question remains whether those environments can faithfully assess real system performance with limited biases.

Towards autonomous perception

Activity is an essential feature of truly autonomous perceptual systems – this is one of the main defended statement of this section. In the active perception approach, inference is the result of an activity. However, perceptual system activity is itself embedded into a more general environment, implying that the value or the quality of what is inferred cannot be settled by the perceptual system itself, but is measured by the degree of user/client requirement satisfaction, as already discussed in the previous chapter.

We propose here directions to integrate also user requirements in the perceptual system activity and argue for the development of a dialog between perceptual system and user/client as a key issue.

Interacting with the user. In the active perception approach – at least the traditional view put forward in (Bajcsy et al., 2018) – interaction with the world is purposive: "An actively perceiving agent is one which dynamically determines the *why* of its behavior and then controls at least one of the *what*, *how*, *where* and *when* for each behavior". However nothing is said about the origin of this *why*. The perceiving agent has no ability to evaluate or discuss the usefulness or relevance of the goal it pursues, whether it is really achievable or nonsense. The actively perceiving agent is not autonomous, in the sense that it has no possibility to prescribe its own objectives.

In the **X** pattern proposed above (Fig. 3.1), interaction has two faces: with the world, as a source of sensory data, but also with a user/client for which the agent produces meaning and possibly receives reward. One may ask: Why is autonomy necessary for perceptual systems? Why should they be able to decide by themselves what are the optimal criteria to optimize? The main reason is because an objective is in fact always a trade-off between various features: available resources, energy, time, beliefs confidence, measure quality, probability of false alarm vs. detection etc. A user/client does not have the full knowledge of what is achievable by the perceptual system and may decide impossible objectives to fulfill if not agreed by the perceptual system.

Dialogue. A key feature of an autonomous perceptual system is therefore the capacity of interacting with the user/client. A spontaneous form of such an interaction would be to initiate a *dialog*, i.e. a sequence of statements, questions, answers, opinions, etc. But what would be the role of a dialog involving a perceptual system?

The design of conversational agents is a well studied area, with three main types of problems: question answering, chatbots and task-oriented dialogs (Chen et al., 2017; Gao et al., 2019). They are not grounded on any information content that could be provided by sensory data. General problems such as good question formulation (Buck et al., 2018) or successful negotiation (Lewis et al., 2017) have been addressed in this area.

The integration of language and sensory data, mostly visual, is an emerging topic that has motivated recently a large quantity of work (Mogadala et al., 2019). One of the first studied tasks has been Visual Question Answering (Antol et al., 2015), where the main difficulty was to design algorithms able to answer free-form natural language questions about image content. This elementary interaction has been extended to full visual dialogs for image retrieval or object discovery tasks (De Vries et al., 2017; Das et al., 2017c; Lu et al., 2017d; Jain et al., 2018; Zhuang et al., 2018; Niu et al., 2019), image generation (Kim et al., 2017; Cheng et al., 2018), navigation (Vries et al., 2018) or pricing (Parvaneh et al., 2019).

What is interesting with the idea of encapsulating active perception in a dialog is that the involved agents ideally exploit awareness of each other skills, behavior and objectives: the questioner (client), for instance, can adapt to what he/she knows about the answerer (perceptual system) (Lee et al., 2018c), even if the objectives of each agent are not the same.

There are several advantages of dialoguing: a first one is to avoid misunderstanding about what can be achieved (function and performance), another is to improve the confidence on perceptual outputs by sharing or comparing knowledge sources or by using it as a justification (this question will be discussed more thoroughly in chapter 4). In general, dialoguing offers a flexible format able to express finely tuned requests and answers.

However, making use of a dialogue as an interactive interface implies specific capacities for a perceptual system: control of the semantic gap between language and internal state, the ability of authorizing interruptions and generating anytime intermediate results, reliable self-assessment and performance prediction etc., features that are rather new to standard perceptual system design and that will be partly discussed in chapter 5.

3.2 Development

The previous section addressed the question of the overall functional architecture of an APES and how it should operate and interact with the external world and, possibly, with its user/client. In this section we discuss how to produce such a system, how to *develop* it.

Perceptual systems, either adopting static **I** or dynamic **X** and **Y** patterns, are purposive: they are specific to the world features from which they acquire sensory data and to their user/client needs. They must be adapted to their environment.

Machine Learning (ML) is currently the favored approach to introduce specificity and adaptation to many predictive processes, including perception, and is often claimed to be one of the key technologies of the digital world transformation, fostering automation in many application domains. The successes of deep learning (Schmidhuber, 2015; Goodfellow et al., 2016) in computer vision, speech recognition and natural language processing have made it unavoidable in contemporary artificial intelligence. We briefly examine in this section how this approach works, why it is successful but also point out several of its limitations.

Supervised learning: the unavoidable paradigm

Machine learning can be seen as the answer to the situation where the specification of the target process cannot be expressed exhaustively by intension, i.e. by a series of computable and verifiable properties, constraints or predicates, and can be substituted or completed by an extensive description, i.e. as a series of examples sampling potential occurrences – the data.

One the most frequent ML use case is called supervised learning, where the desired predictive process is described by samples of input/output pairs. Roughly speaking, prediction in this framework is a form of statistical data interpolation, where concepts of estimator bias, confidence intervals, tests, Bayesian vs. frequentist statistics (Murphy, 2012), information theory (Cover and Thomas, 2012), etc. make sense, but have been renewed to deal with more complex mathematical objects such as Probably Approximately Correct (PAC) Learning (Valiant, 1984) or generalization bounds (Vapnik, 2013) able to better analyze and control higher dimensional data.

Machine Learning for perceptual process development, typically for computer vision or pattern recognition (Bishop, 2006), is not a new trend, but has been made ubiquitous with the conjunction of large annotated database availability, cheap massively parallel computational resources (GPU) and software frameworks easing optimization, monitoring and design. It also has shown dramatic performance improvement on standard problems such as image classification using complex and highly dimensional parametric predictors.

One of the most noticeable successes has been the rebirth of neural networks with AlexNet (Krizhevsky et al., 2012; Alom et al., 2018), a deep convolutional architecture that may be traced back to rather old previous work ((Fukushima, 1980), (LeCun et al., 1989)) but that outperformed current state of the art on the ImageNet competition by more that 9% when submitted.

The main characteristics of a supervised learning approach for the design of perceptual processes, especially when exploiting deep learning techniques, can be summarized as follows:

- The availability of data is a critical step and a lot of effort has to be devoted to their creation; performances are most of the time unsurpassable when such database is available.
- Deep network architectures allow the unification of low and high level vision (end-to-end approach) and are able to generate multi purpose unsupervised informative data features as a by-product of the global optimization process.
- Design objectives shift to architecture innovation and optimization rather than dedicated sensory data modeling. Many studies start with "generic" convolutional architectures such as VGG (Simonyan and Zisserman, 2014), Inception (Szegedy et al., 2015), DenseNet (Huang et al., 2017a), ResNet (He et al., 2016) etc. for image classification, or Faster RCNN (Ren et al., 2015), SSD (Liu et al., 2016b), FPN (Lin et al., 2017) for detection, and even address now the question of architectural meta-learning with Neural Architecture Search (NAS) (Elsken et al., 2019).
- Requirement assessment has been replaced by benchmarking on testing databases using average metrics, confusing various evaluation concepts in a single statistical framework.

Those features explain why supervised learning using deep learning techniques is a key approach of modern artificial perceptual systems that has reshaped the R& D agenda. There are still remaining issues to make perceptual tasks developed this way really mature, safe and robust (this question will be addressed fully in chapter 5) for real-world and critical applications. Several traditional elementary tasks are not yet performing well – object detection, for instance (Liu et al., 2018), is still far from human performance on standard images.

A first and well known issue is *data dependence*. When dealing with highly dimensional sensors and contingent situations, sampling all possible occurrences to build annotated data is impossible, limiting the validity of a pure data-driven approach. We will examine possible answers to this problem in the next section.

A second issue is related to the integration of a perceptual function in a more global system, for instance in a robot (Sünderhauf et al., 2018) or in an autonomous vehicle, where perception is coupled with other objectives, and may benefit from

or alter other tasks. This integration constraint also implies that other performance measures or trade-offs should be involved, for example between accuracy, memory and computational time in image classification (Howard et al., 2017) or detection (Huang et al., 2017c).

The third issue is the real quality of sensory data features obtained after learning, usually in the lower layers of a deep network, and their transferability to other tasks, finely tuned or not. Several studies show that learned features usually improve generalization (Yosinski et al., 2014) often with a specific local adaptation (Long et al., 2015), but clear explanations and controls of this capacity are still missing.

Alleviating data dependency

Relying on data to develop a perceptual system may appear convenient and efficient, but is precisely limited by the availability of data. Although several problems can have access to large quantities – millions of images for face recognition for instance[5] – current practical situations are more characterized by unevenness and small sample size, at least at the scales that are required by modern deep networks to give reliable results. Indeed, collecting, creating, formatting, storing, retrieving, annotating data however are costly tasks and often not applicable in contexts where data are structurally rare (medical (Ker et al., 2018), military, experimental sensors, etc.) Several strategies have been proposed to address the question of small data samples (Shu et al., 2018): we summarize and discuss them in the following.

Transferring information from a similar task

A first idea to compensate for the lack of data is to exploit other abundant sources that are expected to be close to the target domain and provide useful information to help solve the desired task using *Transfer Learning* techniques. "Given a source domain $\mathcal{D}_S = \{\mathcal{X}_S, \mathrm{P}(X_S)\}$ with a corresponding source task $\mathrm{T}_S = \{\mathcal{Y}_S, f_S(\cdot)\}$ and a target domain $\mathcal{D}_T = \{\mathcal{X}_T, \mathrm{P}(X_T)\}$ with a corresponding task $\mathrm{T}_T = \{\mathcal{Y}_T, f_T(\cdot)\}$, transfer learning is the process of improving the target predictive function $f_T : \mathcal{X}_T \rightarrow \mathcal{Y}_T$ by using the related information from $\mathcal{D}_S$ and T_S, where $\mathcal{D}_S \neq \mathcal{D}_T$ or $\mathrm{T}_S \neq \mathrm{T}_T$" (Pan and Yang, 2009; Weiss et al., 2016; Day and Khoshgoftaar, 2017).

Stated this way, transfer learning encompasses a large number of configurations: if the source domain contains annotated data or not, if they are annotated with the same set of labels, if source and target domains or tasks intersect, etc. Several studied problems in this family have potentially clear practical impact:

Domain adaptation: $\mathcal{D}_S \neq \mathcal{D}_T$. (Csurka, 2017; Venkateswara et al., 2017) Transfer occur when input spaces are different ($\mathcal{X}_S \neq \mathcal{X}_T$) but share common features, or

[5] http://megaface.cs.washington.edu/

when only their prior distributions differ ($\mathrm{P}(X_S) \neq \mathrm{P}(X_T)$). The question is to align the two domains by globally compensating their shift (Courty et al., 2016), or to prevent target and source predictor parameters from being too far from each other in a joint optimization (Rozantsev et al., 2018), or to do both (Saito et al., 2018).

Semi-supervised learning: $|\mathcal{Y}_T| \ll |\mathcal{Y}_S|$. (Albalate and Minker, 2013; Zhu, 2005; Chapelle et al., 2006; Zhu and Goldberg, 2009) In this setting, the target domain has much fewer labeled data than the source domain: somehow, the target domain gives the prior statistics, whereas the source gives its meaning, i.e. typical samples of the relation between input sensory data and predicted label. This problem is also sometimes called transductive learning when what is sought is the point-wise prediction given the learning dataset ($\mathbf{x} \mapsto y$) and not the whole predicting function on the potential target domain $f_T(\cdot)$. The solution to the problem relies on two assumptions: smoothness and low dimension of input data distribution, and clusterable structure, i.e. the fact that decision boundaries lie in low density regions. The proposed algorithms can be separated between generative, discriminative and graph-based approaches, and have Deep Network extensions (Shi et al., 2018; Robert et al., 2018; Iscen et al., 2019).

Weakly-supervised learning: $\mathcal{Y}_T \neq \mathcal{Y}_S$. (Deselaers et al., 2012; Oquab et al., 2015; Papandreou et al., 2015; Bilen and Vedaldi, 2016; Durand et al., 2016; Durand et al., 2017; Zhou, 2017; Zhou et al., 2018b; Hong et al., 2017) Available annotations in the learning dataset may not meet the expressivity level required for the target task: for example, the dataset may only contain global tags whereas the task is to detect and locate the objects, or object locations may be encoded by point, scribble or bounding boxes whereas pixel level labeling is expected as output. Transferring relies on the assumption that the target output is the assembly of components that can be identified by a weak annotation and an aggregating process.

The various strategies used to counterbalance the lack of annotated data by exploiting a similar task rely on hypotheses about data distribution structure that are difficult to both clearly identify and verify. Most of proposed methods are only evaluated empirically by comparison with a completely supervised setting to assess their potential, and often with negative conclusions: supervised learning usually performs better, meaning that, when it is possible, it is often more fruitful to invest in data rather than in algorithms.

Sharing features and representations across tasks

Another possibility to address the problem of data paucity is to consider from start a broader set of objectives and to make the target task one of the elements that shares components with the others. One tactic is to anticipate the exploitation of the available data by extracting good features to easily solve the target task, often using

another related data corpus. Typically, what is sought is a discriminating embedding where decisions can be made simpler and reliable with few data, for example using a nearest-neighbor approach or a simple linear classifier. This general scheme has been addressed under different settings.

Representation learning.

Designing good "representations of the data that make it easier to extract useful information when building classifiers or other predictors" (Bengio et al., 2013b) is a (the?) fundamental activity of data-driven artificial intelligence. In modern machine learning approaches, this activity takes the form of a data preprocessing, sometimes called encoding or feature extractor, and fed to a further decision or predictive step. These representations are expected to have some universal – they must be adapted to the largest scope of input data – properties of smoothness, expressiveness, versatility, intelligibility and efficiency. Good pretrained representations are expected to give high performance with small data and has been proved to benefit large sample problems (Sun et al., 2017).

The question of building good representations have been addressed in several directions. A first one has been to identify the informative factors of variation of input data, a problem that can be traced back to PCA or ICA, and provide a *disentangled* representation (Locatello et al., 2019; Higgins et al., 2018; Mathieu et al., 2019) able to reveal the underlying data structure.

Another more recent trend is to exploit the versatility of deep learning architectures and optimizing schemes to solve multiple task problems (Liu et al., 2015; Yang and Hospedales, 2016; Ruder, 2017) or learn a related task that is expected to provide a useful transferable representation from a *self-supervised* learning step (Pathak et al., 2016; Donahue et al., 2016; Jing and Tian, 2019).

In robotics, the idea of designing good representations has been studied mostly as a consequence of building an environment description organized as affordances (Min et al., 2016) or sensori-motor invariants, and often solved as a problem of *developmental learning* (Sigaud and Droniou, 2015). Reward based environment exploration, either coming from an external teacher (Ivaldi et al., 2013) or from internal regulation (intrisic motivation or curiosity)(Pathak et al., 2017; Oudeyer and Kaplan, 2009; Oudeyer, 2018) is a central learning principle of this family of approaches that put the emphasis on structuring robot skills or goals (Laversanne-Finot et al., 2018; Péré et al., 2018) rather than improving sensory data representation or processing, although more computer vision oriented skills have been proposed to learn visual saliency (Craye, 2017; Craye et al., 2018).

One of the reasons for the attractiveness of deep learning is a capacity to provide easily transferable and versatile features in the first layers of a network. Those quasi-universal representations are expected to have good but average performance

for a large repertoire of tasks. This implies that to be really efficient, they need to be adapted to the specific task: one classical technique is *fine-tuning* – a unsolved question being to decide what should be considered plastic (Yosinski et al., 2014). Another strategy is to complete the generic features by more specific parts (Rebuffi et al., 2018; Bucher et al., 2018).

Meta-learning.

The idea of this family of approaches (Vanschoren, 2018) is to prepare representations in a previous global phase using a large database, so that learning with few samples is made easy. Meta-learning is often presented as the problem of "learning to learn" (Caruana, 1997; Wang and Hebert, 2016) and is one of the favorite paradigms to solve one-shot or few-shot learning (Fei-Fei et al., 2006; Vinyals et al., 2016; Ren et al., 2018; Rusu et al., 2019; Wang and Yao, 2019). A seminal meta-learning model is MAML (Model Agnostic Meta Learning) (Finn et al., 2017; Antoniou et al., 2019) that seeks good initialization parameters for further specific adaptation and exploits two intricate stochastic gradient loops: one inner loop to locally adapt the parameters to a current few-shot virtual task, and an outer meta loop that updates the initial parameters using accumulated gradients computed at the updated parameters of the inner loop.

Incomplete cross-modal learning.

Data can be provided by several sources or modalities, in combination or independently. One possibility to compensate for data paucity is to exploit correlations between modalities and estimate missing data as a latent variable in the prediction step. *Zero-shot learning* (Lampert et al., 2009; Lampert et al., 2014) is a specific case of cross-modal learning for classification where the input modalities are images and semantic class descriptions: "zero-shot" means that several classes are only known from their description (unseen classes) and not by samples (seen classes). Various learning configurations have been proposed: standard, where the problem is to predict only unseen classes, *generalized* where input data may come from either seen or unseen classes(Chao et al., 2016; Xian et al., 2018; Le Cacheux et al., 2019a; Le Cacheux et al., 2019b), *instance transductive* where data from new classes are available but without labels (Fu et al., 2014), and *class tranductive* when the whole set of unseen class descriptions is know (see (Wang et al., 2019a) for a recent comprehensive review).

Metric learning for zero-shot classification (Bucher et al., 2016b; Bucher et al., 2016a)

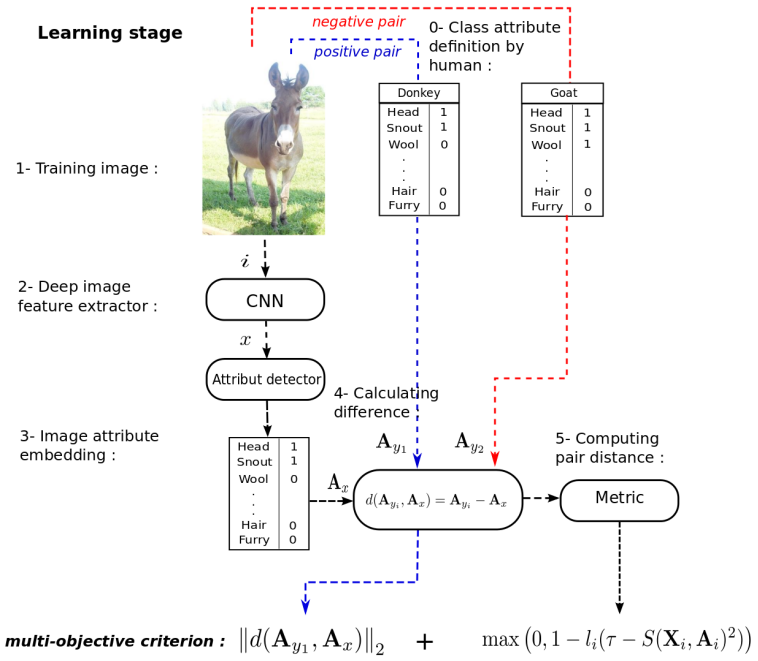


Fig. 3.6: Improvement of attribute prediction consistency using metric-learning for zero-shot classification.

One of the most practiced approach to solve zero-shot learning is to map data and class descriptions – often represented as a series of attributes – in a common embedding space where similarity measures become meaningful. Once the embedding is set, inference can be done using a simple compatibility function and an argmax operation (Akata et al., 2013). In his PhD thesis, Maxime Bucher describes a metric learning formulation to improve attribute prediction consistency (see Fig. 3.6). The idea was to introduce a loss able to simultaneously characterize class prediction and data embedding in the attribute space (Bucher et al., 2016b). An extension to this scheme was proposed to deal with hard negative examples that are selected sequentially according to a trade-off between data uncertainty and intra-class correlation (Bucher et al., 2016a).

Generating new relevant data

Artificial generation is an obvious way to compensate for the lack of data. When dealing with images or videos, a spontaneous idea is to make use of modern procedural rendering engines and models that are able to produce huge quantities of photorealistic outputs in various configurations (Souza et al., 2017; Richter et al., 2016; Müller et al., 2018). Artificial data generation potentially samples a huge variety of situations and is expected to extend to a large usage domain.

However, although often visually satisfying when involving sufficient computing capacity and modeling endeavor, data produced this way are always biased, sometimes in a subtle way, bias that may be amplified when plug into a machine learning process. In other words, algorithm that perform well on artificial data are not guaranteed to have comparable performance on real data.

Another strategy to generate data is to exploit sampling based generative models, where parameters have been learned from a representative database. The most practiced approach exploits an adversarial min-max optimization on a deep neural architecture and is currently referred to Generative Adversarial Network. Since the first paper (Goodfellow et al., 2014a), a large volume of work has proposed several alternative models and optimization schemes able to stabilize learning and increase the dimension of generated data (Creswell et al., 2018; Pan et al., 2019; Wang et al., 2019b; Hong et al., 2019).

The deep network framework allows flexible architectural design and combination, and offers various ways to condition these data-driven generation capacities: as image-to-image translation to achieve style transfer (Isola et al., 2017; Zhu et al., 2017), from semantic maps to image or video (Wang et al., 2018d), from label to image (Mirza and Osindero, 2014), from text to image (Reed et al., 2016; Zhang et al., 2017a), from image and label to image (Wang et al., 2018c) etc. Conditional generative models can be considered as a way to unify data-driven and procedural generation to produce high quality data.

One possible usage of this generation capacity is to augment or transform available data – for instance images obtained by procedural synthesis – to improve predictor accuracy. Generative models have been applied to domain adaptation for semantic segmentation (Murez et al., 2018; Vu et al., 2019), continual learning (Lesort et al., 2018), gaze and pose estimation (Shrivastava et al., 2017), super-resolution for object detection (Bai et al., 2018), 3D object recognition (Wu et al., 2016) etc.

Conditional generative models for zero-shot recognition (Bucher et al., 2017)

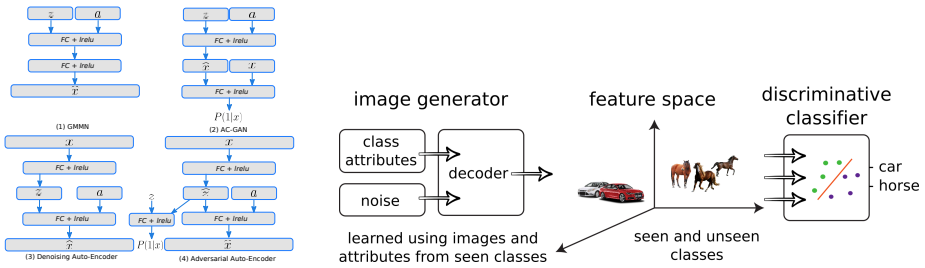


Fig. 3.7: Zero-shot learning using generated features. Unseen class data are conditionally generated using one of the networks on the left, and combined with seen data to learn a discriminating classifier.

When addressing the problem of zero-shot learning, a simple idea is to generate data as surrogate samples of the unseen classes and then apply a standard discriminating classifier. The problem is therefore to build a data generator that is able to produce data from a semantic class description, either from a vector of attributes or from a word vector representation. It was found more efficient to build directly discriminating features (penultimate layer of a VGG or GoogLeNet) instead of images.

Several architectures have been tested, inspired by existing generators: a generative moment matching network (Li et al., 2015) that exploits a Mean Maximum Discrepancy criterion (Gretton et al., 2012), and three networks using an adversarial optimization scheme: Wasserstein generative adversarial network (Arjovsky et al., 2017), adversarial auto-encoder (Makhzani et al., 2016) and denoising auto-encoder (Bengio et al., 2013a).

This simple scheme gave state of the art result on 5 databases (CUB, AwA, SUN, Pascal & Yahoo and ImageNet) and demonstrated the power of generative models to solve both standard and generalized zero-shot learning (Bucher et al., 2017).

Generative models have shown great promises as a main or supplementary tool to produce missing data. However, it is often difficult to assess the intrinsic quality of generated data. Data generators are not defect-free.

A first difficulty is the instability of training. In most of the schemes, there is little guarantee that the finally converged state will allow complete and faithful sampling of the original data. Phenomena of mode collapsing (Lala et al., 2018), i.e. the fact that complete areas in the data space become inaccessible to generation, and overfitting (Webster et al., 2019) are likely to happen. Several metrics besides simple visual inspection have been proposed to evaluate generated data quality, but lack clear validation (Borji, 2019a).

Hybridizing with models

Supervised learning is one of the most powerful paradigm to develop an artificial perceptual capacity but relies on two fundamental hypotheses: the first one is that it can be expressed as a function with meaningful and tractable means to compute predictive error, the second one is that a large corpus of data is available to reach reliable and controllable performance. When those two hypotheses are met – and verifying that they are is another important question – any other type of approach has difficulty to compete. This is however not often the case in practice.

We have presented in the previous section various learning strategies to alleviate data dependency in case of data paucity. Here we discuss how what is usually called *model-based* approaches can be hybridized with machine learning to improve performance.

Model-based approach for perception

Hybridizing data-driven and model-based concepts is a current concern of AI (see for instance the ANITI project[6]) but not a new one. The idea of developing computational architectures integrating the symbolic and neural/subsymbolic levels (an old expression that refers to data-driven machine learning), has been a constant question since the rebirth of neural networks in the 80's (McClelland and Rumelhart, 1987) and has led to the development of hybrid neural systems, for instance (Bookman and Sun, 1994; Wermter and Sun, 2000).

The expression "Model-based" potentially refers to many things: any formal activity that requires a model, i.e. a set of extensive and intensive variables and their relations that abstractly characterize an entity behavior or several of its features. In the field of artificial intelligence (Russell and Norvig, 2016), it also usually refers to the symbolic and logical approaches (Genesereth and Nilsson, 2012), sometimes embedded in a probabilistic framework (Pearl, 1988) with key problems such as knowledge representation, planning or reasoning.

More formally, when applied to perception, i.e. a process that produces signs or measures related to the world, a model-based approach can be reduced to any process that exploits the knowledge of two functions:

- A generative or direct model $G : \theta \mapsto \mathbf{x}$ that maps latent variables θ to observation $\mathbf{x}$;
- An inference, prediction or decision $D : \theta \mapsto \mathbf{y}$ that maps the latent variable θ to the final output $\mathbf{y}$, meaningful to the user/client.

The variables $\mathbf{x}, \mathbf{y}, \theta$ may be multi-dimensional, may have complex structure, and are usually considered random.

The generative model is where knowledge stands: it postulates the existence of a latent cause or condition θ that generates data using a known process G. A classical model is a linear map $\mathbf{x} = \mathbf{A}^T \cdot \theta + \eta$ where η is a random noise, but more complex generative models can be defined, using geometric and radiometric features for instance (image synthesis), sample based generative networks, logical engines, etc. In computer vision, geometric object recognition from a CAD model as has been practiced since the beginning of computer vision (Mundy, 2006) and exemplified in

[6] "The ambition of the ANITI project is to develop a *new* generation of artificial intelligence called hybrid AI, combining data-driven machine learning techniques with symbolic and formal methods for expressing properties and constraints and carrying out logical reasoning." https://aniti.univ-toulouse.fr/

various textbooks before the deep learning era (Faugeras, 1993; Forsyth and Ponce, 2002; Hartley and Zisserman, 2003; Szeliski, 2010).

The inference model produces the output expected by the user. It can be the whole set of latent variables, a subset of it – in object recognition, the user may be only interested in the object pose, not a representation of its texture – or a decision among a set of hypotheses.

A critical step in a model-based approach, is to have access to this latent variable θ knowing a direct model, to estimate it. If the mapping G is invertible (injective), one solution is to apply an inversion process and use the result as input of the inference. However, in many situations, because of the structure of the generation process (think of object parts that are hidden by occlusion) or because of unknowns in the model (e.g. illumination, noise, texture), the generative model cannot be inverted. The question of estimation under reasonable hypotheses becomes central. When the problem is embedded in a probabilistic setting, inversion and inference can be joined in a single but compound process, exploiting optimal formulations such as $\mathbf{y} = \text{argmax}_{\mathbf{y}'} \int_\theta \Pr(\mathbf{y}'|\theta) \cdot \Pr(\theta|\mathbf{x})$.

Hybridization is still a challenging question, especially with the overwhelming domination of deep learning and the temptation of instantiating complex function with "end-to-end" supervised objectives that produces efficient solutions on many problems. In the following, two different hybridization strategies are presented and discussed: the integration of learned components in a core model-based engine, the use of data-driven constraints or regularities in a model-based solver.

Learned components in models

A first simple idea to hybridize model-based and data-driven approaches is to introduce learned components inside the direct model or inference functions, but keeping a generative formal structure. Many "ancient" successful approaches for visual object recognition (Grauman and Leibe, 2011) fall under this category. In the Implicit Shape Model (Leibe et al., 2004), for instance, the relation between local appearance and geometric pose is learned in a previous phase and used to infer object localization through a voting scheme. Another famous example, the Discriminating Part based Model (Felzenszwalb et al., 2009), exploits a geometrically constrained local patch generative model for object detection, where parts and geometric constraints are learned from an image database.

During their PhD, J. Guinet (Guinet, 2008) and C. Le Barz (Le Barz, 2015) have exploited a hybrid scheme mixing learned components, mostly similarity measures, and generative models (a geometry-based image formation and a temporal Hidden Markov Model) for two different computer vision applications.

Three-dimensional object re-identification (Guinet, 2008)

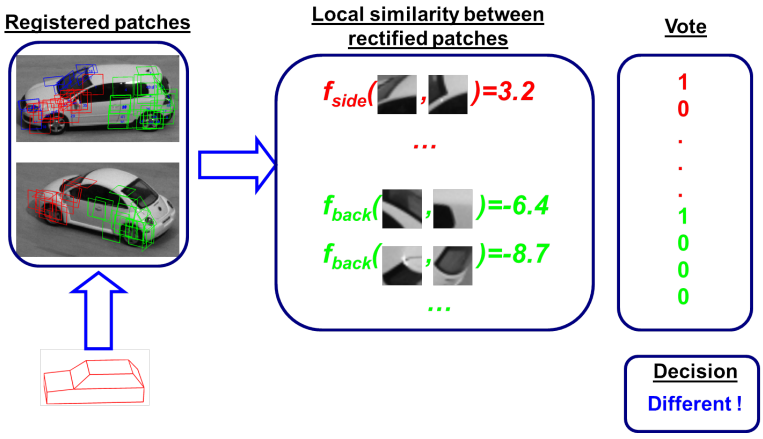


Fig. 3.8: A 3D CAD model is matched to images in order to register and rectify patches which are later compared using a learned similarity measure. A simple vote predicts whether the objects observed in the two images or sequences are different or identical.

In his PhD thesis, Jonathan Guinet exploited a deformable CAD model to predict object appearance observed in a given sequence to other viewing conditions. The geometric model was used to extrapolate their aspect (Guinet et al., 2007) and identify the object parts that could be commonly visible and matched in two different poses for further appearance based comparison. The learned part of the model was a similarity measure between local patches (Fig. 3.8). This chain has been applied to 3D object re-identification in a visual surveillance application where objects could be observed by two different cameras.

Visual localization by image retrieval and temporal integration (Le Barz, 2015)

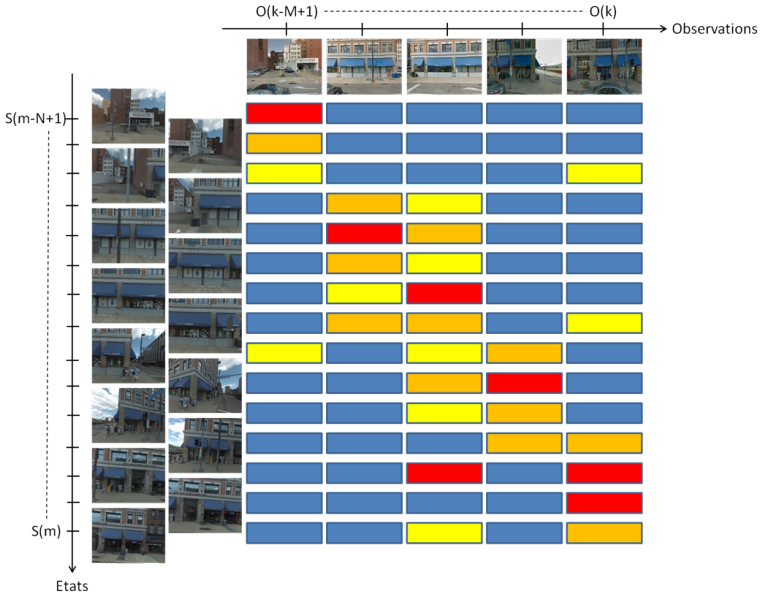


Fig. 3.9: Similarity scores between observed and reference image sequences. The similarity measure is learned off-line. A Hidden Markov Model is used to filter the assignment matrix and globally improve localization.

Ego-localization is a fundamental function in robotics applications and can be achieved by vision. The availability of geo-localized image data like Google Street makes possible absolute localization by comparing current observation and a database of such referenced images. In (Le Barz et al., 2014) it was studied the possibility of temporally filtering local matching through a Hidden Markov Model representing a priori motion – the direct model – to improve global sequence matching between the whole observed video sequence and the database (Fig. 3.9). The similarity measure depended on learned features obtained through standard bag-of-words or adapted to the context through metric-learning (Le Barz et al., 2015b; Le Barz et al., 2015a).

In several application contexts (medical imaging, remote sensing, intelligence, industrial vision, etc.), data are not that diverse due to the common nature of their production (same sensor, same viewing conditions, same type of scene, etc.) The annotation of few data is likely to be sufficient to sample with enough accuracy their variability, given that relevant generative data *models* are available.

Under these hypotheses, we have studied two algorithmic scenarios of how a simple annotation step of one or two images could be exploited for the detection of a not too diverse class of objects in known contexts.

Light annotation for object detection

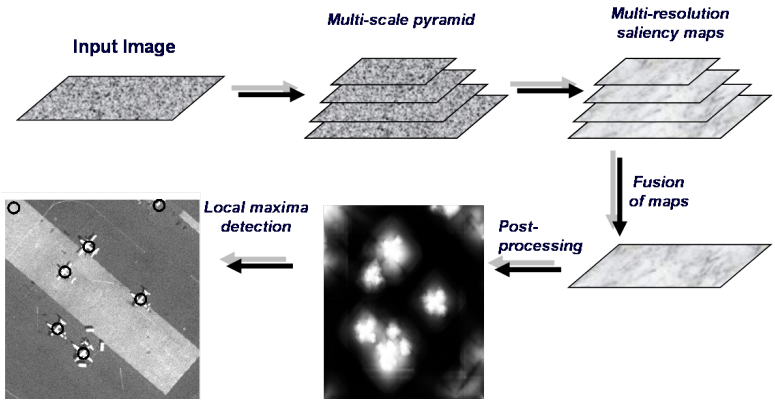


Fig. 3.10: Schema of the saliency detection chain combining multiple scale filters.

The first scenario addressed the problem of detecting objects in remote sensing images using a probabilistic parametric model of object saliency (Borji, 2019b) mixing scale and contrast features. It was studied during the PhD work of Benjamin Francesconi (Chalmond et al., 2006). The model contained few

parameters and can be learned by annotating less than 10 positive (object) and negative (background) patches. The image features were Gabor filters at several scales (Fig. 3.10), the role of the model being to weigh the correlation between filter outputs and object scale.

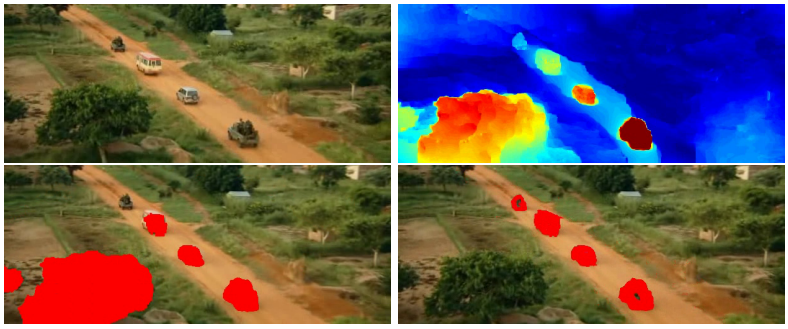


Fig. 3.11: Moving object detection chain combining a direct motion compensation model and semantic classification. Up left: original image. Up right: optical flow magnitude after motion compensation. Down left: moving object detection obtained by thresholding the optical flow. Notice that 3D objects like trees have a motion magnitude similar to vehicles. Down right: Moving object detection obtained by combining motion and semantic classification.

The second scenario addressed the problem of detecting moving vehicles in aerial videos. It was studied during the PhD work of Christophe Guilmart (Guilmart et al., 2011). The underlying generative model exploited two properties: that moving object have high values of residual optical flow obtained after dominant affine motion compensation, and that vehicles drive on road and follow its direction. Those two knowledge statement were instantiated as a semantic segmentation algorithm combining optical flow information and appearance features to produce vehicle/road/background classes. The classification was learned from a single frame roughly annotated in three classes.

Model as a learned constraint

The previous paragraph presented one type of hybridization consisting in introducing learned components that, combined with the direct model, was exploited in the inference step. Another strategy is to modify the direct model itself – or more precisely its inversion – by learned features.

The fundamental idea is to add to the inversion, and possibly to the inference, a constraint that helps disambiguate the set of potential output. It can be considered as an *Inverse Problem* (Idier, 2013), where typical applications are image denoising or reconstruction, tomography, super-resolution, optical flow, stereo-vision, etc. which is "solved" as a global statistical inference from multiple observations that takes

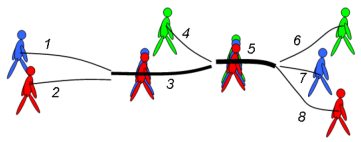
advantage of the hypothetical smoothness of the generative process, but controlled by a *learned* regularization.

In low-level computer vision, numerous studies have exploited geometric constraints as direct model (the projected 2D image motion between video frames can be fully determined by the depth structure and camera motion) to learn to estimate dense features (optical flow, depth maps) (Godard et al., 2017; Yin and Shi, 2018), sometimes also mixing semantic information to add another stabilization constraint (Chen et al., 2019).

Inverse problems can be solved by directly learning the inversion as a supervised learning problem (Lucas et al., 2018) when data is available. Another strategy more respectful of an inverse problem optimization formulation ($\text{argmin}_\theta \|\mathbf{x} - G(\theta)\| + \lambda\phi(\theta)$ where $\phi(\theta)$ is a regularization function) is to add a learned regularized function representing more accurate data prior (Lunz et al., 2018) or to mimic an iterative optimization process as a neural architecture, typically a proximal operator (Meinhardt et al., 2017; Rick Chang et al., 2017) or a truncated Neumann series (Gilton et al., 2019). (Arridge et al., 2019) is a recent review about solving inverse problems using data-driven models.

The expression of a data-driven learned constraint used in the prediction process, either as a regularizer or as an operator in the iterative optimization process, encodes knowledge in a numerical and rather opaque way. Hybridization with more explicit knowledge representation has been proposed through Markov Logic Networks (MLN) (Richardson and Domingos, 2006), a variant of conditional random fields (Sutton and McCallum, 2012) which integrates logical predicate descriptions to specify in a more meaningful way variable interactions.

Markov Logic Networks for tracklet association (Leung and Herbin, 2011)



1. *stCoh(t1,t2) ∧ stCoh(t1,t3) ∧ (t2!=t3) ∧ simApp(t1,t2) ∧ simApp(t1,t3) → isGroup(t1)∧ join(t1,t2) ∧ join(t1,t3)*
2. *join(t1,t2) ∧ join(t1,t3) → isGroup(t1)*
3. *stCoh(t1,t3) ∧ stCoh(t2,t3) ∧ simApp(t1,t3) ∧ simApp(t2,t3) ∧ (t1!=t2) → isGroup(t3) ∧ join(t1,t3) ∧ join(t2,t3)*
4. *join(t1,t3) ∧ join(t2,t3) → isGroup(t3)*
5. *sameApp(t1,t2) ∧ stCoh(t1,t2) ∧ (!∃ t3 stCoh(t3,t2)) ∧ (!∃ t4 stCoh(t1,t4)) → join(t1,t2) ∧ sameObj(t1,t2)*

Fig. 3.12: (Left) One example of a complex occlusion configuration that needs to be handled to improve track continuity. (Right) Example of rules that describe group configuration and tracklet status.

In applications of video-surveillance of crowded areas, multiple static and dynamic occlusions between objects is frequent (Fig. 3.12 (Left)), and lowers the performance of multi-target trackers: they often produce short and unrelated sequences, making the on-line analysis of videos unreliable. A posterior processing is likely to enhance track quality. Tracklet association refers to the joining of reliable track fragments to form longer, coherent tracks.

In (Leung and Herbin, 2011), we have proposed to use Markov Logic Networks as a flexible way to introduce complex object configurations, typically the notion of group, explicitly handling situations where a group is formed or disperses (Fig. 3.12 (Right)). The weights of the MLN are learned using a database of simulated tracklets sampling various configurations. Inference results can be represented by logical formulae and be interpreted more easily.

When is hybridization useful?

There are several reasons to hybridize formal models with data-driven approaches.

A first reason is to compensate for the lack of data, the fuel of machine learning. Models are used as knowledge representations able to provide good approximate predictions that can be adjusted by empirical means. However, when large data corpora are available, model-based approaches give in general no performance gain over "pure" deep learning approaches, typically when accurate generative models depend on many unknown parameters as it is the case for vision.

A second reason has to do with intelligibility. Because models instantiate a form of declarative and shared knowledge, results can be interpreted, analyzed, evaluated and controlled in the light of this knowledge, potentially resulting in a better understanding of the perceptual process for instance by displaying intermediate results on geometric patterns that can be visually checked. We will come back to this issue in chapter 4.

A third reason is validity. When models have been legitimized by other means, for instance when they conform to known physical laws or equations, perceptual predictions may, in a way, be considered *true*. The idea of "physics-informed" machine learning, for instance, is a recent trend that aims at "solving supervised learning tasks respecting any given laws of physics described by general nonlinear partial differential equations" (Raissi et al., 2019). Nonetheless, it seems difficult to get rid of any type of empirical evaluation: models themselves are built on hypotheses or initial values that must also be validated or checked.

A last reason to address hybridization is to use machine learning as a way to provide expressive and practical representations that are physically meaningful and can be used for further reasoning, i.e. as *physical engines* (Chang et al., 2016; Battaglia et al., 2016). In these approaches, knowledge is given as a valued graph structure of pairwise interacting variables (Sanchez-Gonzalez et al., 2018) where links encode physical dynamics. The role of machine learning is to fill the corresponding graph values.

Collaborative development

Specification is the activity of identifying and describing in an exploitable way what the perceptual system actually should do and under what constraints, i.e. expressing the requirements it has to satisfy. Specification and development go hand in hand: specification must have an idea of what is achievable, and development must conform to the requirements.

In the **X** pattern proposed in Fig 3.1, the perceptual system interacts with the world and potentially with a user/client when operating perception. But the user/client may also have an influence in developing perceptual capacities, notably because it is the main part concerned by their results.

Machine learning being one of the essential techniques used to develop a perceptual capacity, the specification step, or at least one part of it, relies on providing good annotations as examples of what the system should output[7]. We presented above several learning approaches that could compensate for low annotation (see pg. 89). But another possibility is to exploit a collaborative setting where both sides exchange information about their states and wishes.

One can think of two ways to develop a collaborative specification.

The first one, commonly referenced as *active learning*, allows the perceptual system to query annotations about data to the user. The provided annotation is expected to be very impactful on system performance, and the global annotation process should be ruled by a expected performance gain and annotation cost. Of course, the activity of annotation itself can also be accelerated by proposing efficient interactive softwares (Andriluka et al., 2018), but the question of what to annotate comes first.

This problem has been addressed in machine learning literature for a long time (Cohn et al., 1996), and applied to problems of classification (Joshi et al., 2009; Jain and Kapoor, 2009; Tuia et al., 2009), image retrieval (Gosselin and Cord, 2008; Wang and Hua, 2011), object detection (Brust et al., 2018), segmentation (Li et al., 2011; Vezhnevets et al., 2012) and co-segmentation (Batra et al., 2011).

Although there are theoretical results about the number of data requiring annotation to reach a given expected performance (Dasgupta, 2005), they are not really applicable to large dimension problems (Ramirez-Loaiza et al., 2017). Most practicable strategies exploit heuristics to improve confidence of decision (Settles, 2009), or input domain coverage. (Fu et al., 2013) in a rather recent survey segments the various approaches in two families: those that evaluate uncertainty locally under an i.i.d. hypothesis, and those that exploit similarity between samples.

[7] We will discuss in chapter 5 the impact of specifying by examples on system performance assessment.

The evolution of processing chains towards deep networks requires an adaptation of the annotation query process, either by developing new ways to estimate uncertainty (Wang et al., 2017; Gal et al., 2017; Beluch et al., 2018; Yoo and Kweon, 2019) or by modifying the query process to handle subsets (Sener and Savarese, 2018).

The second type of collaborative specification assumes that the perceptual system is able to output information assessing its current state of achievement, and that the user from this information is able to modify the system in a way that satisfies his/her/its needs. This setting puts the burden of development on the user, but is only possible if the system prepares it by 1/ providing accurate and intelligible inner representation and 2/ allowing effective action possibilities or "affordances" (Gibson, 2019; Norman, 2013). The question of intelligibility of the perceptual process and identification of influential parameters will be discussed more thoroughly in chapter 4.

Dynamics of development

It has been discussed in the previous section that an essential feature of autonomous perceptual systems is their temporal dimension in operation. Temporality is also constitutive of development – there is a before and an after learning – and often modeled as an iterative, or at least sequential process. Temporality of development can also be involved at a larger scale. Indeed, all situations and application contexts cannot be foreseen during the initial design phase, and should be *continuously* updated or learned to preserve or improve skills.

Besides the challenging scientific question, controlling and understanding long temporal scale learning has obvious practical impact allowing to:

- Enlarge progressively competence, i.e. repertoire of tasks or skills.
- Improve performance with online integration of new information.
- Delay obsolescence of systems by easing maintenance and versioning.
- Integrate and adapt pre-learned components for which the learning database is not available due to protection issues.

The idea of providing a system with never-ending (Mitchell et al., 2018) capacity is an old objective of artificial intelligence, with the dreamed objective of reaching the general adaptivity of human intelligence. It is technically addressed in artificial intelligence under *lifelong* (Thrun and Mitchell, 1995; Chen and Liu, 2016), continual (Zenke et al., 2017a) or incremental (Gepperth and Hammer, 2016) learning, with no clear meaningful distinction between those expressions.

The idea of continuously enhancing by learning the capacity of a predictive system has been formulated in various ways: by increasing the input domain extension, by incrementally refining the output scope, or by augmenting the quantity of tasks

that can be accomplished. (Hsu et al., 2018; Ven and Tolias, 2018; Ven and Tolias, 2019) discuss several scenarios for continual learning and try to better define their functional differences, typically whether task id is given or not as input.

The ubiquity and efficiency of deep learning has rather recently reactivated the question of lifelong learning and given rise to new research activity. (Parisi et al., 2018) presents a recent survey on this domain and puts in correspondence natural science and computing issues.

The major problem to solve is the fact that new information, data or tasks, may interfere negatively with the previously reached level of competence, a phenomenon usually denominated by *catastropic forgetting* (McCloskey and Cohen, 1989; French, 1999).

Proposals to address this question can be roughly divided in three categories:

Structural dynamics. The idea of this family of approaches is to increase the expressive capacity of the predictor by adding new architectural features, for example new layers or neurons in a deep network (Rusu et al., 2016). Although simple in its principle, this family has to deal with complexity and scalability issues.

Regularization. Another type of strategy against catastrophic forgetting is to protect the old model by regularization, i.e. by preventing too much variation during model update. This control of modification can be done very locally reasoning at the weight level (Zenke et al., 2017a; Kirkpatrick et al., 2017) and can be interpreted as a homeostatic variant of Hebb synaptic updating rule (Zenke et al., 2017b), or resulting from a modified learning cost imposing an extra variational constraint, for example using distillation (Hinton et al., 2015) between old and new task outputs (Li and Hoiem, 2017).

Memory management. The last type of approaches introduces extra memory modules to store a summary of previous experience, inspired by the Complementary Learning System hypothesis (McClelland et al., 1994) [8]. The memory module can be a simple fixed size rehearsal buffer which is updated with the data used for the new problem, or a learned generative model (Shin et al., 2017; Lesort et al., 2018). This type approach gives in general the best performance, especially compared to regularization, but requires an extra component.

[8] "According to the theory, effective learning requires two complementary systems: one, located in the neocortex, serves as the basis for the gradual acquisition of structured knowledge about the environment, while the other, centered on the hippocampus, allows rapid learning of the specifics of individual items and experiences." (Kumaran et al., 2016)

Several proposed models combine those three approaches: (Rebuffi et al., 2017; Castro et al., 2018) use replay of stored data and distillation. (Rajasegaran et al., 2019) mixes evolutionary architectural selection, example replay and distillation.

Although an old problem, the formal expression of continual or lifelong learning problems is not fully settled, and has given rise of several studies promoting various definitions and evaluation measures.

(Gepperth and Hammer, 2016) propose five challenges for incremental learning: Online model parameter adaptation, Concept drift, Stability-plasticity dilemma, Adaptive model complexity and meta-parameters, Efficient memory models. (Kemker et al., 2018) propose a general framework for evaluating continual learning using three tasks: Data Permutation Experiment, Incremental Class Learning and Multi-Modal Learning, and apply it on five representative algorithms (Standard Multi-Layer Perceptron, Elastic Weight Consolidation (Kirkpatrick et al., 2017), PathNet (Fernando et al., 2017), two versions of GeppNet (Gepperth and Karaoguz, 2016), and Fixed Expansion Layer (Coop et al., 2013)). (Lopez-Paz, 2017) defines two metrics: Backward Transfer to measure forgetting of past learned tasks, and Forward Transfer to measure the predictive capacity of old task to solve new ones. (Chaudhry et al., 2018) introduces the idea of "intransigence" (inability to update the knowledge to learn the new task), complementary to forgetting. (Díaz-Rodríguez et al., 2018) adds complexity evaluation measures such as model size, storage and computational efficiency. (Farquhar and Gal, 2018) identifies shortcomings and biases of current evaluation protocols and proposes critical desiderata of good continual learning methods. (Pfülb and Gepperth, 2019) describes a large-scale empirical study of catastrophic forgetting.

Incremental learning of Visual Question Answering

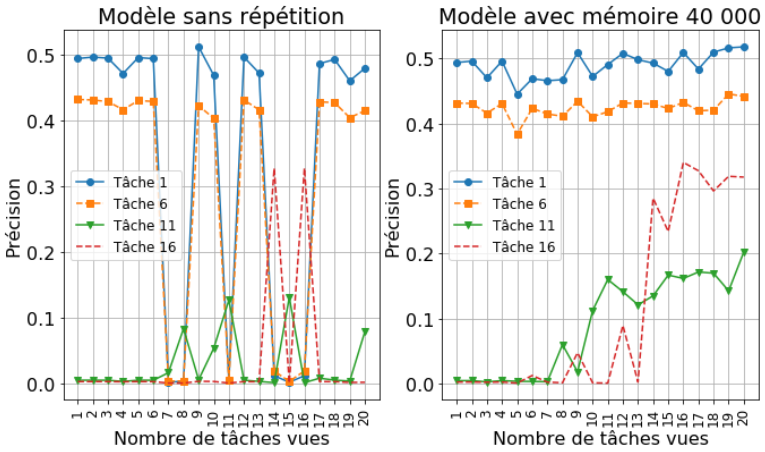


Fig. 3.13: Comparison of incremental learning performance applied to VQA for various tasks, with (right) and without (left) using a rehearsal buffer of size 40000.

In his ongoing PhD thesis Alexis Lechat (2018-2021), addresses the question of continuously improving the capacity of a system able to answer questions in free-form natural language about the visual content of images or scenes (Visual Question Answering = VQA). This problem is complex since it requires to incrementally increase output vocabulary (answer) and input task repertoire (question).

A first study (Lechat et al., 2019) showed the importance of mastering the imbalance between question semantic types (yes/no vs. what/where/who) and the impact of task ordering during learning: certain tasks have very forgetting effect, while others are able to benefit from previous learning. One consequence of this first finding is that controlling and estimating input task complexity, at semantic and visual levels, is a key to apply continual learning on VQA.

3.3 Discussion: machine learning and perceptual systems

The learning paradigm has two facets: from a natural science perspective, it is another world for adaptation or acclimatization to environment; from an engineering or computer science perspective, *machine* learning is a way to specify and design algorithms when efficient, expressive and reliable models are lacking.

Perception, being a very complex faculty due to the high dimension of sensory data, has been using for a rather long time machine learning techniques in association with more analytical models (see the discussion about hybridization above). The spectacular success and ease of use of deep learning techniques on fundamental perceptual functions such as image classification or object recognition has considerably biased the current research agenda: a large part of recent work has been to translate old problems as a supervised learning formalism, usually in an end-to-end learning approach, often if not always leading to a notable performance gain.

Deep supervised learning is indeed currently the paradigmatic design pattern of ML and cannot be completely bypassed nowadays: besides its noticeable good performance, one origin for this primacy is a clear problem expression as functional optimization that makes proposed approaches easily comparable with the definition of train, validation and test datasets. It also has practical and conceptual limitations, however.

We have examined in a previous section its dependency on large corpus of annotated data and presented several of the current directions to mitigate this constraint. In

the same spirit, (Yuille and Liu, 2019) questions the relevance of a design process relying on large data corpora solely: "it seems highly unlikely that methods like Deep Nets, in their current forms, can deal with the combinatorial explosion. The datasets may never be large enough to either train or test them. [..] Although Deep Nets will surely be one part of the solution, we believe that we will also need complementary approaches involving compositional principles and causal models that capture the underlying structures of the data."

(Marcus, 2018) diagnoses other limitations of deep learning besides data hungriness: limited capacity for transfer, no natural way to deal with hierarchical structure, open-ended inference poor performance, not sufficiently transparent, not being well integrated with prior knowledge, cannot inherently distinguish causation from correlation, presumes a largely stable world in ways that may be problematic, often cannot be fully trusted, is difficult to engineer with.

Another big issue is perceptual system versatility and scalability: in supervised learning, there is only one goal-oriented optimized cost, sometimes compound to deal with multi-task problems, but unique. Adding a new task to a learned system is usually detrimental and produces catastrophic forgetting, inhibiting the incremental extension of skill repertoire. This limitation favors the development of task specification other than purely data-based, using maybe more symbolic description (hybridization). Formulating interaction as a supervised problem is also difficult, since the data exchanged between the system and its outside cannot be considered identically and independently distributed.

Deep learning has proven effective on several problems, but its behavior is not well understood, especially its good generalization capacity despite a number of parameters sometimes larger than the learning samples (Poggio et al., 2017)(Kawaguchi et al., 2017)(Zhang et al., 2016). Network architecture, optimization algorithm and data distribution have a combined impact that is difficult to disentangle. A growing number of studies have proposed approaches to extend generalization bounds (Bartlett et al., 2019) (Arora et al., 2018) (Arora et al., 2019), to estimate deep network expressivity (Gribonval et al., 2019) (Zarka et al., 2019), to decipher the role and behavior of stochastic gradient descent optimization (Shwartz-Ziv and Tishby, 2017)(Zou et al., 2018) and the impact of neural architecture to the geometry of the loss landscape (Nguyen, 2019). A global understanding of deep learning behavior leading to its better control is still missing.

Fundamentally, the general question that should be addressed is whether an “everything but data” approach is sufficient to specify, design and evaluate perceptual or cognitive capacities. The limitations of supervised learning to implement desired features of autonomous perceptual systems suggest that this is not the case, and that other development strategies should complement or assist data-driven approaches.

3.4 Research directions

In the previous chapter, it was proposed that an Autonomous PErceptual System should have expressiveness, agency, cognitiveness and trustworthiness properties. In its current state, deep supervised learning alone will have difficulties to provide reliable solutions to implement them all, even if the flexibility of its formalism and the availability of programming environments allow large inventiveness.

The reminder of this section presents several research directions that could contribute to improve the design and development of APES's.

Dynamics of versatile perceptual systems

If the right way to consider perception is as an Autonomous PErceptual System, especially if versatility is the desired feature characterizing autonomy, several issues regarding perceptual dynamics are worth being addressed. Three research directions are proposed:

Structured task spaces

A key question is the capacity for a perceptual system to host and manage multiple tasks with the level of performance required by the user/client. When tasks are dynamic and involve interaction, which is an expected property to reach autonomy, an important issue is to have access to efficient formal structures able to describe and organize in a flexible way their *from what* (input, resources and actions) and *for what* (goal, requirement and output) features.

Designing systems that are able to manage multiple tasks is a rather new concern in artificial perception, and have been mostly aimed at building shared sensory data representations (Ruder, 2017; Kokkinos, 2017; Zamir et al., 2018) or controlling negative interference between tasks during learning using priorization (Guo et al., 2018a) or attentional mechanism (Maninis et al., 2019).

A flexible structured task representation, i.e. with composition operations combining elementary functional components, and handling deep network formalism, is still to be proposed. Several studies have addressed the question of better structuring subgoals by reinforcement learning (Nachum et al., 2018; Nair and Finn, 2019) but to complete a single task. Maybe one direction could be to acclimatize the formal representations developed for image understanding ontologies (Town, 2006; Clouard et al., 2010). However, the difficulty remains to combine these symbolic knowledge representations with machine learning.

Generalized attention

Attention, as a selection or modulation of resources, has become a rather standard algorithmic ingredient in many processing chains (see pg 79). When applied to natural cognition, the idea of attention can be extended to the general management under constraint of resources such as memory, computing, modality, etc. (see the discussion pg. 41).

Dynamic management of resources is a traditional topic of planning in artificial intelligence: several studies involve perception through the next best view selection problem, for visual navigation or scene reconstruction (see the active perception use cases pg. 75). The design of artificial computer game players has also proposed rather global planning problems: (Oh et al., 2016) applies reinforcement learning to "Control of memory, active perception, and action in Minecraft". However, all those studies concentrate on a single optimal objective, and have not addressed the question of a flexible and adaptive strategies that may be able to reconfigure resource management policies conditionally to a given task. Studying the conditions of a generalized attention could be one path towards increasing the versatility of perceptual systems, i.e. their capacity of efficiently completing tasks from a large repertoire.

Contextual and functional dynamics

The idea of *context* has been introduced in the field of artificial or natural perception as a means of representing priors and constraints likely to limit research spaces, and therefore to increase performance or simply feasibility when the environment is complex, contains a large amount of interacting entities, some of which potentially ambiguous when taken alone. Relying on the right priors is therefore critical.

Active perception can be interpreted as implementing prior dynamics: the Bayesian sequential decision of Eq. (3.5) is a simple way to update priors when new data is available, the belief at time $t-1$ acting as a prior to the belief at time t; sequential hypothesis rejection (pg. 69) is a more drastic way to concentrate priors.

Priors depend on the task to complete, on the nature of the environment and on the history of interaction with the world. But dynamics can also operate at a functional level. When interpreting dynamic scenes, the occurrence of certain events may trigger different perceptual subgoals (switch from categorical to fine-grained classification, from tracking to action recognition, etc.) or modify the value of scene content (objects may become obstacle, people may turn dangerous, etc.) and therefore task performance requirements or objectives.

Autonomous PErceptual Systems have therefore to handle two intertwined dynamics: contextual and functional, where priors and subgoals jointly evolve to complete a master task. Formalizing such a coupled dynamics for perception, and developing algorithms able to implement it is a long time research objective. One possible direction could be to address the question by proposing an online version of Neural Architecture Search (Elsken et al., 2019).

A remaining problem is that of evaluation of such perceptual dynamic systems, due to the huge dimensions involved (input, inner states and trajectories in those spaces). The very large parameter space of deep networks brings a new type of complexity to master. This question will be discussed more precisely in chapter 5.

Learning of perceptual dynamical systems

The fact that APES's are temporal entities has an impact for their development. The question of specifying, designing and controlling dynamical systems is a standard issue in engineering and physics, where control goes hand in hand with modeling. When perception is the target function, this collaboration is more difficult, first because the dimensions of inputs, variables or state spaces involved are often huge, but also because perception has to deal with two external contingencies: the environment and the user/client. Two specific issues regarding learning are discussed in the following.

Attention for learning vs. Learning attention

It has been argued previously that attention is a key operating feature of perceptual systems. Deep network formalism has proposed several end-to-end models that can be learned using stochastic gradient optimization schemes, most of them implementing a soft version of attention (see a more detailed presentation pg. 79).

Conversely, one can think of exploiting attention to improve learning efficiency: exploiting a process that selects the *good* experience to integrate in a system should be beneficial, the question being of course to define what it is to be good for learning – selective attention can also act negatively (Schwartzstein, 2014).

Two research problems can be reinterpreted as "attention for learning": the first one is the well studied exploration/exploitation dilemma in reinforcement learning (Ishii et al., 2002), where efficient exploration can be seen as a process that selects the good actions to improve policy reliability. The second, also well studied, is active learning (already discussed pg. 103), with its non interactive hard example mining version (Bucher et al., 2016a; Shrivastava

et al., 2016; Yuan et al., 2017; Smirnov et al., 2018), where learning is improved by selecting the examples that may have the highest impact.

In natural cognitive systems, at least what is understood about their functional structure, learning cannot be reduced to simply optimizing a deep network structure, although complex, through a single optimizing cost. Learning involves many cognitive components, typically various types of memories (long term, working, episodic, verbal/visual etc.) and and raises the question of their dynamic management. The massive corpus of findings and models that have been proposed in the cognitive science literature is still waiting to be exploited as a source of inspiration in artificial intelligence.

Learning to act vs. learning to interact

An important dimension of an APES is its dual interactive relation with an external world and a user/client. The relation to the external world will be assumed rather neutral: the perceptual system does not modify the latent, possibly dynamic, structure of the world, only its ego-relation to it. In other words, the external world is not causally modified by perception. This hypothesis excludes object manipulation for instance that should be considered as involving a bigger system.

However, the relation to the user/client is potentially more complex: both parts may act, the perceptual system by providing answers, justifications or requests, the user/client by expressing requirement, acknowledgement and revision. We have already discussed the usefulness of developing an interactive perception (pg. 81) with the key issue of establishing a dialogue between the two parts. The question is to develop such a capacity by learning (Das et al., 2017b).

One possibility to address this problem, besides defining the nature and meaning of dialogue, is to consider the perceptual system and user/client as two different communicating agents (Foerster et al., 2016) that optimize a common utility, but with different action repertoires and vocabularies. The literature on multi-agent learning is large, and mostly relies on reinforcement learning approaches (Shoham et al., 2007; Bu et al., 2008) and game theory (Lanctot et al., 2017). However, most of the studies have focused on how to define local policies of cooperative multiple but simple agents where the number of agents is large but the vocabulary of actions small.

In the case of a interactive perceptual system, the two agents involved (the APES itself and the user/client) are potentially complex and do not play the same role: they can be simultaneously or sequentially cooperative or

competitive, leading to complex and unstable situations (Lowe et al., 2017). User and perceptual system may not work with the same time scale, and require fine modeling features such as asynchronous event and interruption handling.

Joint dynamics of operation and development

The traditional engineering workflow makes a clear separation between development and operation, the former being usually divided in several steps, for example by applying a *V-Model*. However, in natural perceptual systems, the development phase cannot be definitely ended: although some critical and final learning occurs in the early days of life (see the vision deprivation experiments of (Wiesel and Hubel, 1963)), brain plasticity allows lifelong adaptation and new skill acquisition.

Another argument that favors intrication of development and operation is the fact that operation is a complete experience that can be valued and exploited by an (unsupervised) adaptation rule, for instance through Hebb's associative law (Sejnowski and Tesauro, 1989): actual operation may have an influence to future operations.

A first attempt to combine operation and development is to address the question of lifelong learning, as already briefly discussed pg. 104. The proposed approaches follow a "learn then run" scheme, although repetitively, and show how difficult it is to counter catastrophic forgetting effects in current algorithms.

A question thus is to unify operation and development rather than separating them as two distinct phases, and go beyond the traditional train/test design pattern of supervised learning. One can think of several research directions to address this question:

Online interactive learning

One simple possibility is to sequentially integrate the flow of interaction with the system outside to progressively improve. This requires that the outside shares with the perceptual system some utility value of the produced outputs as one of the interactive content, for instance a desired output or any other type of reward.

In principle, online learning has access to a sample i.i.d. generator, and could be better than batch learning that shuffles the same dataset. Efficient gradient based optimization algorithms have been proposed for low to medium state spaces (Duchi and Singer, 2009; Mairal, 2015) but the question of their

extension to deep networks, where batch learning is usually found to be more efficient in practice, remains.

Another difficulty is to guarantee that interaction yields faithful sampling. Indeed, inputs from system outside may be intentionally poisoned (Steinhardt et al., 2017; Wang and Chaudhuri, 2018), leading to undesirable states (see what happened for instance to the chatbot Tay [a] that was driven to output offensive statements).

Curriculum and developmental learning

A second idea is to embed perceptual skill development in a more global and autonomous behavior learning objective, inspired by the study of children learning where a "self-generated learning curriculum allows infants to avoid spending too much time on goals that are either too easy or too difficult, focusing on goals of the right level of complexity at the right time" (Forestier et al., 2017).

Progressive skill development can be addressed in two directions. The first one, more formal, aims at organizing adaptation steps by controlling data complexity (Bengio et al., 2009; Weinshall and Amir, 2018; Guo et al., 2018c) and can be interpreted as a regularization strategy in a supervised learning setting. The second one, more directly inspired by natural system development (Oudeyer, 2018), organizes experience as a series of subgoals that can be explored and learned by the system in a unsupervised way using a curiosity principle modeled as intrinsic motivation (Péré et al., 2018; Laversanne-Finot et al., 2018).

One limitation of curiosity-driven approaches when applied to APES design is that they fundamentally follow an **O** pattern: agents are only concerned by themselves and with their relation to an external, potentially complex, world. However, as we have argued, an APES is better understood as instantiating an **X** pattern, with an essential role of the user/client that issues queries and receives outputs from the system. The question of faithfully taking into account such an actor that can take several roles (prescriber, teacher, helper, etc.) in a developmental learning framework is open.

Online heterogeneous learning

Supervised learning is often considered as being the most efficient type of learning protocol. However, in practice or in real life, the conditions required to its use are rarely met, and other types of "suboptimal" settings have been proposed: transfer, zero/one shot, multi-task, semi/self/weakly supervised, etc. All those variants have given birth to very specialized subfields, each

with their own protocols and corresponding benchmarks. The question of exploiting them jointly and dealing efficiently with the possible heterogeneity of available data, in format, quantity, annotation level, completeness, has not been clearly addressed.

The concepts of never-ending (Mitchell et al., 2018) or lifelong learning (Chen and Liu, 2016) emphasize the non regressive acquisition of new skills in a system (see short presentation pg. 104), with the underlying hypothesis that coupling and sharing knowledge and representations between a large number of tasks will give a collective benefit to all of them: scaling could be the solution, not the problem. These problem, however, do not formally address the simplest question of progressively and opportunistically improving the performance of a given set of tasks using any type of available data, knowledge or experience.

Finding formalism and algorithms to solve *online heterogeneous learning* has two potential impacts: the first obvious one is practical – having a system or an algorithm able to profitably absorb any piece of information, whatever its type, could be useful from an engineering point of view; the second is scientific, with the underlying hypothesis that addressing and solving heterogeneity in learning will give insights on what generic structures or processes, if any, are involved, and perhaps understand why machine learning is still not able to fruitfully exploit and integrate any new experience as humans do easily most of the time.

[a] https://futureoflife.org/2016/03/27/tay-the-racist-chatbot-who-is-responsible-when-a-machine-learns-to-be-evil

4 INTELLIGIBILITY — Communicating with APES

The advent of machine learning methods in computer vision has dramatically increased empirical performance. Deep learning is the last instance of this trend, and there is no sign that this vogue is likely to disappear at short notice.

Deep networks are notoriously difficult to interpret: without any tools, the mechanical steps generating the predictions are too complex or numerous to be understood by humans.

To allow interaction with perceptual systems, especially if they ambition some kind of autonomy, predictions must be made in a understandable way. This chapter examines how intelligibility can be obtained from APES, and why “Opening the black-boxes of AI” [1] is a major concern of current research.

4.1 Problem formulation

Why intelligibility?

There are several reasons to improve intelligibility of perceptual systems.

1. A first reason is epistemological. Opacity is an obstacle to reliable science. A normal scientist is frustrated when not being able to understand what is behind the phenomena he/she is studying. Doshi-Velez and Kim argue for instance that "the need for interpretability stems from an incompleteness in the problem formalization, creating a fundamental barrier to optimization and evaluation" (Doshi-Velez and Kim, 2017). In other terms, low intelligibility is synonymous of epistemological weakness.

2. A second reason is efficient engineering. An improved intelligibility makes the development of artificial perceptual systems easier to debug, control and tune.

3. A third reason is trustworthiness. By giving some insights of what they are doing, and how, APES’s are likely to be adopted more easily and be trusted.

4. A fourth reason is validation. Explanations can be used as element of proof demonstrating that APES’s behavior conforms to what is expected or is anomalous.

[1] This is an explicit subtitle of the Villani report (Villani, 2018).

In those four reasons we see that humans, which are of course the key recipient of interpretable representations, can play different roles: scientist, engineer, end-user or authority, determining the type of intelligibility sought out. The nature of explanation recipient, however, is seldom identified as a key feature in the literature, with the exception of (Tomsett et al., 2018; Preece et al., 2018) or (Bhatt et al., 2019) that question the current state of the art on explainability regarding their usability for stakeholders — organizations and end-users [2].

We will see in the next section that the search for intelligibility can also be of interest to automated processes.

What is meant by intelligibility?

The idea of intelligibility has given rise to a huge amount of work recently, mostly motivated by the need to better analyze deep network behaviors, to open their *black box* (Shwartz-Ziv and Tishby, 2017), and extract from their distributed complexity meaningful events and structures.

In the artificial intelligence domain, the vocabulary and concepts connected to intelligibility issues is somehow imprecise. The literature speaks of explanation, justification, transparency, interpretability, comprehension, introspection, and associates requirements of faithfulness, reliability, accountability, fairness, completeness... Several recent papers have tried to clarify those expressions (Lipton, 2016; Doshi-Velez and Kim, 2017; Doran et al., 2017; Biran and Cotton, 2017; Hohman et al., 2018; Gilpin et al., 2018; Guidotti et al., 2018; Ras et al., 2018; Arrieta et al., 2019).

The first concept closely related to intelligibility is *interpretability*. Most authors define interpretability as *the ability to explain or to present in understandable terms to a human* (Doshi-Velez and Kim, 2017).

(Gilpin et al., 2018) make a clear distinction between interpretability and explainability, and consider interpretability, along with completeness, as two concurrent ways to evaluate the quality of explanations. "For a system to be interpretable, it must produce descriptions that are simple enough for a person to understand using a vocabulary that is meaningful to the user", whereas "an explanation is more complete when it allows the behavior of the system to be anticipated in more situations." They conclude that "the challenge facing explainable AI is in creating explanations that are both complete and interpretable: it is difficult to achieve interpretability and completeness simultaneously. The most accurate explanations are not easily interpretable to people; and conversely the most interpretable descriptions often do not provide predictive power." One may prefer to speak of predictive capacity

[2] "We found that while ML engineers are increasingly using explainability techniques as sanity checks during the development process, there are still significant limitations to current techniques that prevent their use to directly inform end-users."

rather than completeness here. (Dhurandhar et al., 2017) proposes to give a formal definition of interpretability as a quantity of information.

Associating understandability and interpretability to explainability does not make things much clearer. What is however certain is that it necessarily brings human in the picture, and involves therefore human centered issues addressed for instance in cognitive science, human computer interaction (HCI) or visual analytics. (Abdul et al., 2018) (also extensively cited in (Gilpin et al., 2018)) describes through a comprehensive literature analysis directions for HCI studies towards usable intelligibility (see Fig.4.1). Their conclusion, from an HCI perspective, is that "While researchers in the ML and AI communities are working on making their algorithms explainable, their focus is not on usable, practical and effective transparency that works for and benefits people."

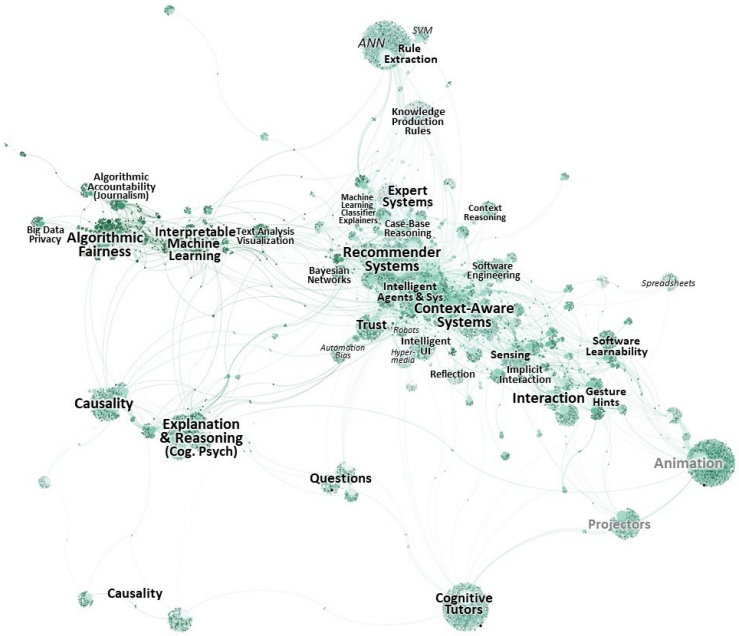


Fig. 4.1: Citation network of 12,412 papers citing 289 core papers on explanations, and identification of topics by clustering. From (Abdul et al., 2018)

As we see from this short discussion, final and shared definitions of the concepts connected to intelligibility still require work and good argumentation. To unify the various ideas connected to explainability that has been discussed in the literature, I – temporarily – propose the following definitions:

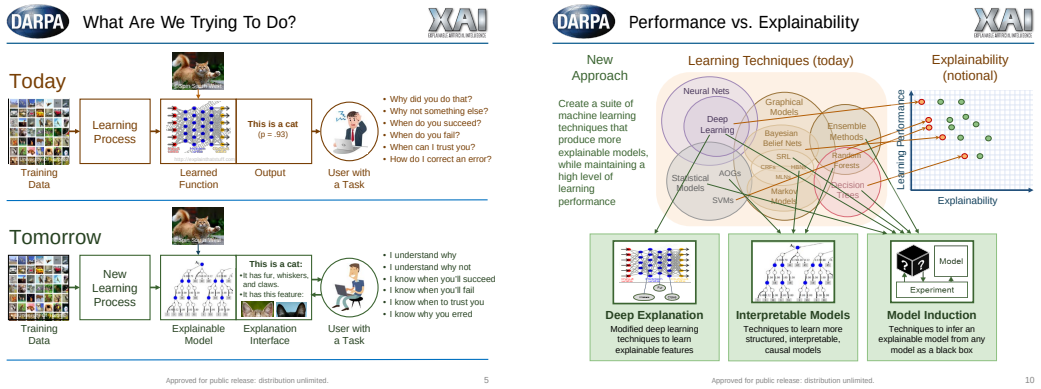
Intelligibility: the capacity of a system to produce explanations.

Explanation: a formal representation that causally depends on the system behavior features (processing and internal states), is interpretable by humans, and contains predictive information.

4.2 State of the art

The idea of providing prediction processes with better intelligibility is not new, and is central to the symbolic approach of AI, sometimes named GOFAI (Good Old Fashioned AI) (Boden, 2014), which promotes explicit step by step understanding and reasoning in its models. The involvement of machine learning techniques in modern methods and the opacity of the resulting prediction processes have encouraged the development of mixed approaches that could benefit jointly from both worlds.

The existence of several recent workshops and tutorials in major AI conferences show that this rather new field of research is creating a research community. A prominent initiative is the XAI program from the DARPA (Gunning, 2017), initiated in 2016, with the final objective of bringing to the user a series of elements that would make him trust and efficiently exploit the predictions made by the automated system (Fig.4.2a). The declared objective of this project is to move the trade off between process interpretability and performance (Fig.4.2b).



(a) Main objective of the XAI project: make user decide with trust.

(b) Mixing machine learning and symbolic approaches to improve the trade off between interpretability and performance.

Fig. 4.2: Two slides from the presentation of the XAI project from the DARPA. From (Gunning, 2017).

In the following, I propose to categorize the work on explainability according to the type of object explained: the whole function prediction or the prediction itself. We describe in the following the main approaches that have been developed in these two categories. Note that other recent surveys have proposed different segmentations:

Explanations of the prediction function

The objective of the first category is to interpret the whole *prediction process* as it should work when applied to a variety of potential inputs (the *how*), either by giving an interpretable account of its functional structure, or by assigning a role to

its functional building blocks. We focus here on the deep network approach which concentrates most of the activity on this problem.

Understanding prediction structure

Explainability in this family of works is to provide hints of what deep networks actually do, what is the role and impact of the functional architecture. One way to answer that question is to exhibit meaningful local or global objectives, typically measures or optimization criteria, that would be able to reproduce the same behavior if applied.

Understanding the prediction function behavior can be traced back to the first studies on multi layer neural networks that interpreted the role of a two layer network as principal component extractor (Baldi and Hornik, 1989) or discriminant projections (Gallinari et al., 1991; Bishop, 1995).

A modern version of this theoretical statistical interpretation trend is the work on what has been called the Information Bottleneck method that gives an interpretation of deep network architectures in the plane of the Mutual Information values, i.e. a two dimension plot of the mutual information between the layer and the input and output variables, as a compressor/predictor structure (Tishby et al., 2000; Tishby and Zaslavsky, 2015). (Shwartz-Ziv and Tishby, 2017) also describes the dynamics of learning as a two phase sequence: fast empirical error minimization followed by slow representation compression. These results were presented as a general understanding of deep network behavior but were found not so general or hard to verify empirically by several authors (Saxe et al., 2018; Amjad and Geiger, 2019). Much work is still to be done to fully understand the behavior of the deep learning approach: the links between architecture, optimization and generalization capacity.

Visualizing feature extraction

One of the big advantages of deep network models is to delegate the design of the feature extraction step – which had been one major research activity in data science, pattern recognition or signal and image processing before the DL Era – to the learning phase.

When dealing with computer vision applications, a natural objective is to find ways to visualize the features encoded in the network, and ideally assign a semantic role to its various components. Probably one of the first studies trying to visually explain features produced by a learned network is the work of Linsker in natural visual system modeling who showed that local image filters emerged in a network with lateral connections learned with an Infomax criterion (Linsker, 1986a; Linsker, 1986b).

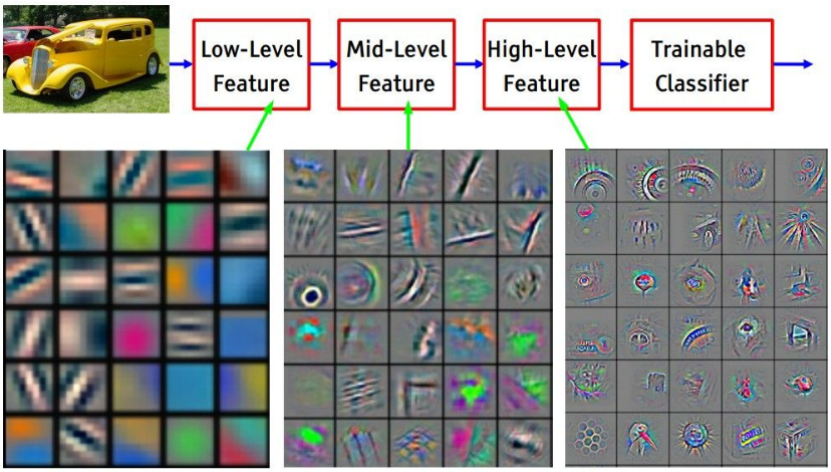


Fig. 4.3: Iconic visualization of the features exploited in a deep network to process a given image (Zeiler and Fergus, 2014).

The visual features in deep network architectures can be more complex. Fig. 4.3 shows an example of what one would expect from a representation explaining the role of its various layers, starting from local filters to more sophisticated pattern detectors.

In this kind of explainability, the idea of feature visualization is not to show a representation of the detector parameters themselves – they have no meaningful structure – but an iconic and interpretable pattern $\mathbf{x}_F$ representing the feature F, i.e. an image or a patch, that will be correlated with its presence in the original input data. Those visualizations can be obtained by optimization, using for instance a criterion with a classical form:

$$\mathbf{x}_F = \underset{\mathbf{x} \in \mathcal{X}}{\operatorname{argmax}}\, S_F(\theta, \mathbf{x}) + \lambda R(\mathbf{x}) \tag{4.1}$$

where S_F is a cost function depending on the network parameters θ, i.e. its architecture and weights, and R is a regularizer used to generate naturally looking images, often a L_2 norm. A typical cost function is the activation of a neuron interpreted as detector for feature F, a whole channel layer activation field, or a classification output score. The set of possible image based feature representations $\mathcal{X}$ may be a given database or an abstract image space. In the first case, the criterion 4.1 is used as a data selector and does not require regularization, in the second case, the image feature $\mathbf{x}_F$ is virtual, giving imaginary non realistic representations (Simonyan et al., 2013; Nguyen et al., 2016).

Explanations of the prediction

The goal of the second category is to provide *justifications* of the prediction made (the *why*), i.e. hints of how the system is working when applied to a given input: they can

be textual (caption), graphical (activity patterns), perceptual (image samples). They can be local (a hidden layer) or global (an activity graph). They can be exact, i.e. the prediction is causally related to their value, the justification being an understandable representation, or proximal, i.e. the prediction can be approximately produced by the justification.

Visual feature contributions

The criterion of Eq. 4.1 is expected to be worth for a whole network. Identifying the features responding to a given input image $\mathbf{x}_0$ is also an interesting explanation. A first possibility is to extend the previous criterion as:

$$\mathbf{x}_F(\mathbf{x}_0) = \underset{\mathbf{x}\in\mathcal{X}}{\operatorname{argmax}}\, S_F(\theta, \mathbf{x}; \mathbf{x}_0) + \lambda R(\mathbf{x}) \tag{4.2}$$

where the cost function S_F now depends on an input image. A typical cost is a Euclidean distance in some representation space $\|\Phi(\mathbf{x}) - \Phi(\mathbf{x}_0)\|^2$, where Φ maps the data to some feature space (Mahendran and Vedaldi, 2015). Another possibility is to invert the features to generate data according to the network, by deconvolution (Zeiler and Fergus, 2014) or by learning (Dosovitskiy and Brox, 2016). The criterion of Eq. 4.2 can also be used to identify the examples from a given data distribution (the $\mathbf{x}_0$) that respond best, and can be structured according to network topology (Zeiler and Fergus, 2014) or classification objectives (Wei et al., 2015). Fig.4.4 shows examples of generated data and associated samples from an image database that respond best/worst to a given feature identified as a network layer. In a recent paper, (Bau et al., 2019) interpret Generative Adversarial Network by identifying the role of several units in the network in the generative process.

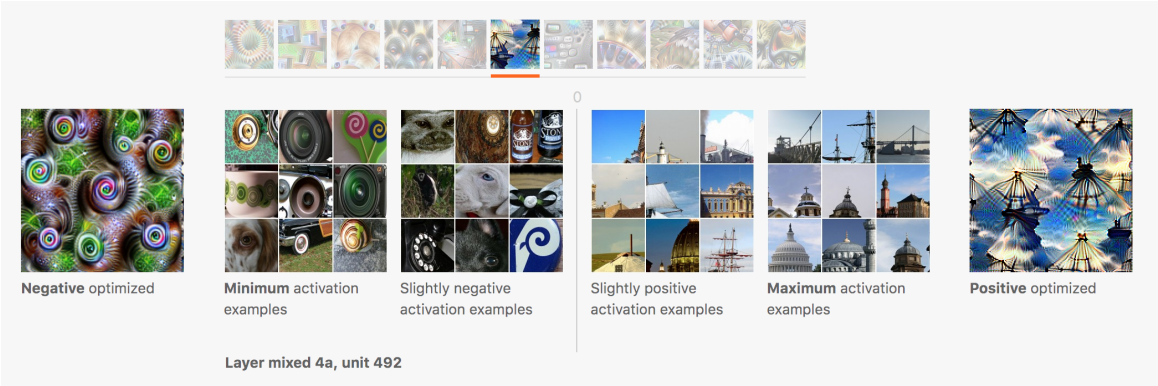


Fig. 4.4: Examples of a generated data and responding input data for a given neuron of a deep network. Worst (left) and best(right) responding data (Olah et al., 2017).

General reviews of visualization issues of deep network has been proposed recently. (Olah et al., 2017) discusses optimization issues and the role of regularization and diversity generation to represent features by image data. (Seifert et al., 2017) is another recent survey on deep network visualization. (Hohman et al., 2018)

analyzes literature analysis on deep network visual explanations from a user oriented perspective.

Approximating the prediction process

To reveal the behavior of prediction process implemented by complex architectures such as deep networks or random forests, a solution proposed by several authors is to approximate it by a simpler model, such as a linear regression, an additive process, a decision tree or a falling rule list (Wang and Rudin, 2015), which is expected to be directly understandable, at least by AI specialists, or if a person can step meaningfully through the algorithm in reasonable time (Lipton, 2016). This approximation is expected to be local, i.e. example dependent.

A typical example of this family is the approach proposed in (Ribeiro et al., 2016) which describes a "Local interpretable model-agnostic explanations" (LIME) for each prediction of a data $\mathbf{x}$ in the form of a simple model $g^*(\mathbf{x})$ that locally fits the prediction function f using a complexity penalty $R(g)$ and optimizes an approximation loss L:

$$g^*(\mathbf{x}) = \underset{g \in \mathcal{G}}{\operatorname{argmin}}\, L(f, g, \pi_x) + \lambda R(g). \tag{4.3}$$

Locality of explanation is achieved by generating data in the vicinity of input data $\mathbf{x}$ through a sampling function π_x, typically a Gaussian distribution centered at $\mathbf{x}$. We see that criterion 4.3 instantiates the interpretability/accuracy tradeoff through this local approximation paradigm.

Most of the other approaches in this family of explanation generators follow the same paradigm. (Zhang et al., 2018d) describe an interpretable proxy (a decision tree) able to explain the logic of each prediction of a pretrained convolutional neural networks. (Lundberg and Lee, 2017) presents a unified framework for interpreting predictions that assigns to each feature an importance value for a particular prediction expressed as an additive model. (Chen et al., 2018) presents feature selector trained to maximize the mutual information between selected features and the output. (Adler et al., 2018) describe an approach that probe black-box models and study the extent to which existing models take advantage of particular features in the data set. (Lakkaraju et al., 2017b) learns a small number of compact decision sets each of which approximate the behavior of the black box model in identified regions of the feature space. (Zhang et al., 2018e) learns how to make each filter represent a specific object part in the input data given a pre-trained convolutional network.

Influential input dimensions

When the prediction approximation relies on sparse inputs, the resulting explanation can be used to indirectly select the most influential features. A rather large body of

research has been proposed recently to specifically identify those features, mostly in the visual domain, without necessarily relying on a low-complexity surrogate approximating prediction. This instance based feature selection process, which can also be considered as a saliency detection, is often called *attribution* in literature.

Most of the approaches define analytical tools that are able to reveal in a given class of predictors those impacting features given the architecture. Fig. 4.5 shows an example of such techniques used to decompose the prediction in terms of contributions of individual input variables, i.e. image pixels, by back-propagating operations. The selected variables may be gathered in the form of a saliency or heat map. Fig. 4.6 shows another flow diagram that generate heatmaps by backpropagating gradients from the main chain.

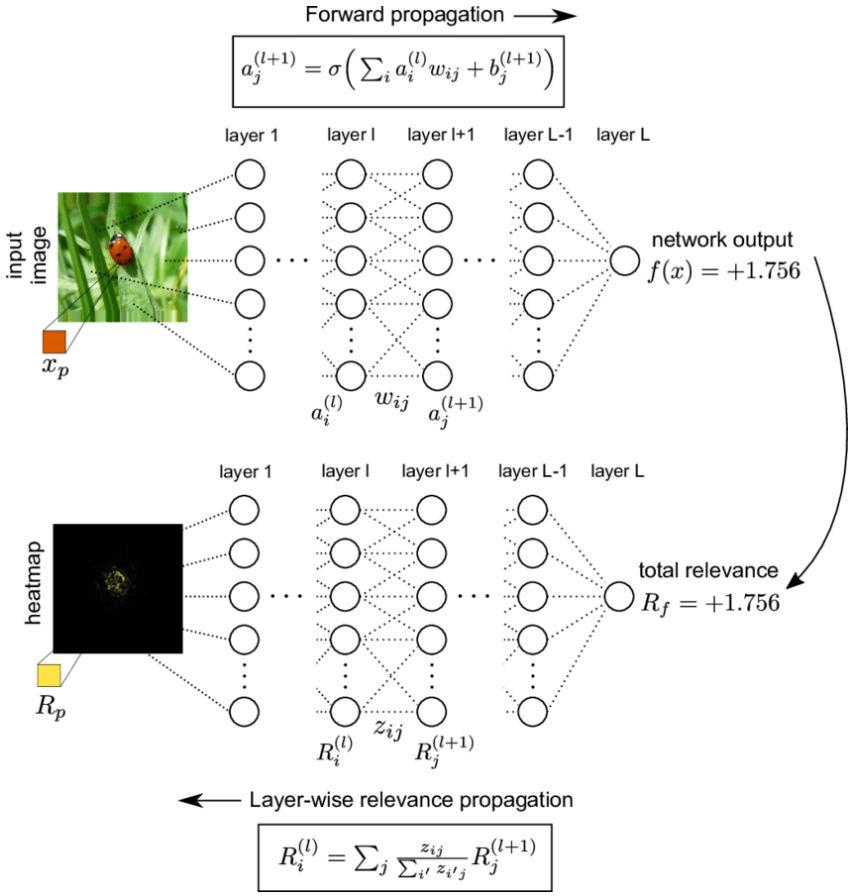


Fig. 4.5: Layer-Wise Relevance Propagation approach used to quantify the contribution of feed-forward neural network components to the prediction (Montavon et al., 2017). The heatmapping site provides several on-line demonstrations.

As recently discussed and unified in (Ancona et al., 2018), attribution methods (Goyal et al., 2016; Shrikumar et al., 2017; Simonyan et al., 2013; Selvaraju et al., 2017; Rajani and Mooney, 2017; Sundararajan et al., 2017; Montavon et al., 2018), rely on a kind of sensitivity analysis of how input data propagates, and exploits combinations of gradients relative to input features to quantify their contribution to the prediction. This type of visual explanation has been applied to various types of

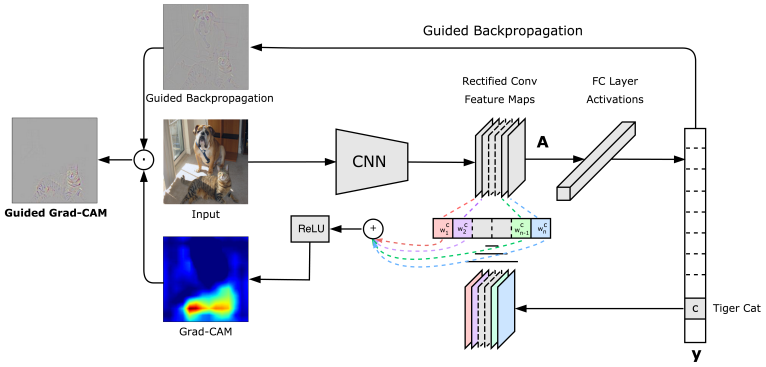


Fig. 4.6: Grad-cam approach that generates heatmaps correlated with prediction (Selvaraju et al., 2017). From `http://gradcam.cloudcv.org/`

image based tasks: classification, captioning or visual question answering, and also in natural language processing (Arras et al., 2017).

One problem with feature based explanations such as saliency maps is that it is not clear whether they can be really trusted. The spectacular demonstration of similarly visual examples fed to a deep network and producing completely different outputs – adversarial examples – shows that the non linear processes involved in deep networks may generate very complex behaviors that are not visually explainable. Adversarial attacks can target feature based justifications as shown in (Xu et al., 2018).

Textual justifications

An alternative to analyzing the inner behavior of the prediction chain is to generate complementary textual justifications that may give hints about the reason why the system produces the current prediction.

(Hendricks et al., 2016) and (Guo et al., 2018b) generate captions that are expected to be more discriminant that general captions for fine-grained classification problems (CUB database). (Li et al., 2018) propose a specific database to learn explanations for VQA problems, and a baseline algorithm. (Park et al., 2016; Park et al., 2018) associate textual and visual saliency explanations on VQA tasks.

Textual justifications are potentially very flexible representations, and may carry a lot of information: they however heavily rely on user linguistics ability and general knowledge to interpret it, which may be the source of ambiguity and misunderstanding. This also makes difficult the quantitative evaluation of justifications: the standard metrics used in computational linguistics to compare texts (Spice (Anderson et al., 2016), Bleu (Papineni et al., 2002), Rouge (Lin, 2004), Meteor (Banerjee and Lavie, 2005), Cider (Vedantam et al., 2015)) are not designed to compare explanatory capacity.

A second limitation of these approaches, which all claim that they can be helpful to debug the operational prediction process, is that they rely on a learned correlation with the actual output. Used as a diagnostic tool, these explanations may not fulfill their objective because they may also introduce measurement noise: when using it as an explanation tool for diagnosis, for instance, it becomes difficult to say who is wrong – the prediction process or the explanation. This problem can be amplified when the explanation itself is learned: the bias of the main operational chain may be propagated to the explanation generation process and lead to good justification of wrong prediction with good faith.

Interpretability by design

Rather than providing supplementary explanations that are expected to give information about the main *opaque* process behavior but with no warranty that it faithfully catches critical aspects of its behavior, an alternative is to design the predictor to be *causally* dependent on some interpretable representations, partly or fully. We describe in the following three different directions to do so.

Attentional mechanisms

Several approaches extract saliency maps as a byproduct of their main attentional based processing flow that may reveal regions of the input space that contribute preferably to the prediction, especially on captioning or VQA tasks (Zhu et al., 2016; Xu and Saenko, 2016; Ben-younes et al., 2017). (Trott et al., 2018) describe a process able to answer to "how-many" type questions that grounds discrete counts in the image. However, the main objective of these maps is usually restricted to an illustration of attention and not of explainability which is not evaluated as such.

A few studies however address more explicitly interpretability issues. (Liu et al., 2017; Das et al., 2017a) try to improve attentional maps in a captioning process with the objective of making model behavior more human-like, and therefore more interpretable. (Xu et al., 2018) show that attention, bounding box localization, and compositional internal structures are vulnerable to adversarial attacks for captioning and VQA tasks, limiting the reliability of explanations generated as attentional maps.

Reasoning

Attention emphasizes what features are important for prediction, but not how to actually make this prediction. Reasoning is one possible functional strategy which relies on a sequential or branching process where each functional step, belonging to a given repertoire, is expected to be interpretable.

Several studies have recently addressed the problem of visual reasoning for scene understanding. (Johnson et al., 2017b) learn a program generator that can be applied to the image to provide the answer. (Hu et al., 2017) generate a concrete network architecture, and then execute the assembled neural module network to output an answer for visual question answering. (Ilievski and Feng, 2017) build reasoning models that combine modules specialized to elementary visual and linguistics tasks. (Mascharka et al., 2018) define a program generator where each visual module is associated with a visual justification (saliency maps), combining reasoning and attention.

All those studies make use of the CLEVR dataset (Johnson et al., 2017a) which has been specifically designed to evaluate reasoning through visual question answering tasks, contains 700K of generated data (image, question, answer) and associated functional programs able to answer questions by an explicit visual reasoning (Fig. 4.7). Several studies have proposed solutions to question answering on this dataset (Perez et al., 2017) without generating an interpretable program, losing the intelligibility capacity of explicit reasoning approaches.

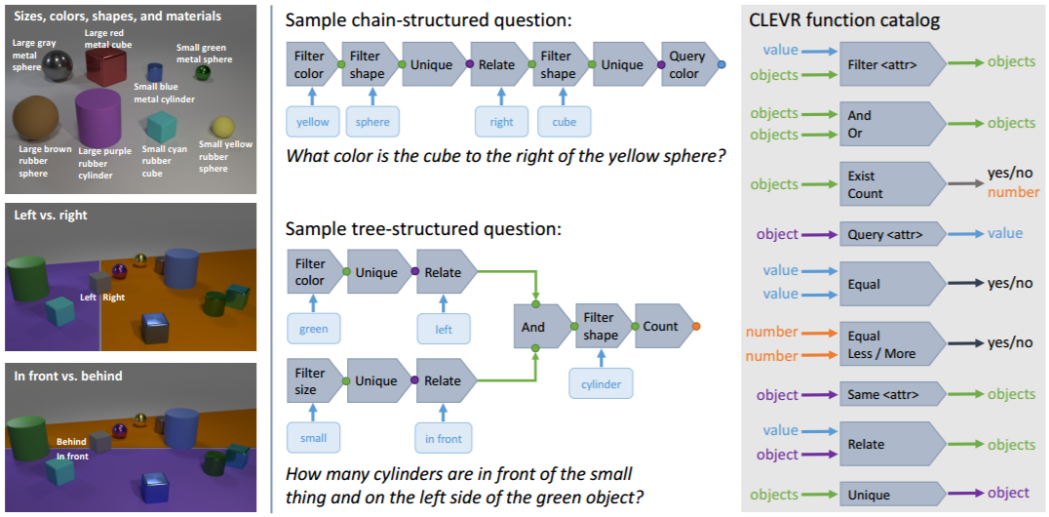


Fig. 4.7: Left: Shapes, attributes, and spatial relationships. Center: Examples of questions and their associated functional programs. Right: Repertoire of basic functions used to build questions (Johnson et al., 2017a).

Causal interpretable inner states

The last strategy to make prediction interpretable by design is to force the processing pipeline to host an intelligible intermediate surrogate state that causes the prediction. Interpretability may take various forms, but the important point is to make the prediction dependent on this intermediate representation, preventing the explanation from being uncorrelated with prediction accuracy.

Semantic bottleneck (Bucher et al., 2018)

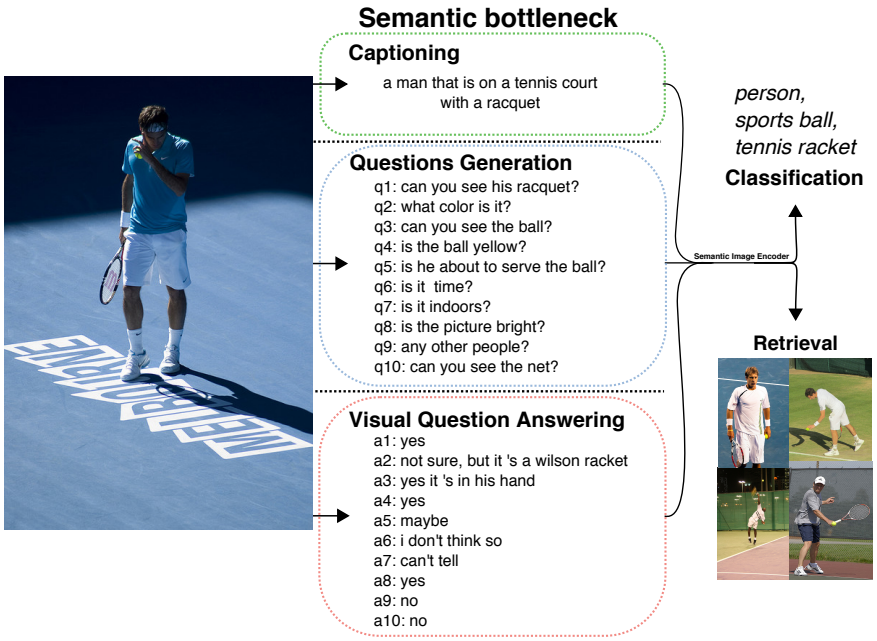


Fig. 4.8: Illustration of the semantic bottleneck approach which enforces a processing chain to generate a textual causal inner representation.

In (Bucher et al., 2018) we proposed a *semantic bottleneck* approach that provides a directly interpretable representation that forces the prediction process itself to be interpretable in some way, since it *causally* relies on this intermediate semantic representation. The textual representation is generated using a combination of image-to-text algorithms (captioning, dialog generation), adapted to the target task (multi-label classification or content based retrieval), and is evaluated as a support for failure prognostic.

Uncertainty expression

The prediction process may be uncertain due to incomplete knowledge or ambiguous input. Providing hints of the information quality is an important aspect of reliable interaction with APES. How to represent, to manipulate and to exploit uncertainty is an old question of artificial intelligence, which has given rise to an extensive literature, with various competing frameworks (probability, Bayesian modeling, possibility theory, fuzzy logic, Dempster-Shafer evidence theory, etc.) (Zio and Pedroni, 2013). Intelligibility of uncertainty may be required at two levels: during inference – how combine inner uncertain representations to produce accurate prediction in an interpretable way – and in the prediction representation – how describe and quantify output hypothesis distribution in an understandable form.

Hierarchical multi-label annotations (Tousch et al., 2008; Tousch, 2010)

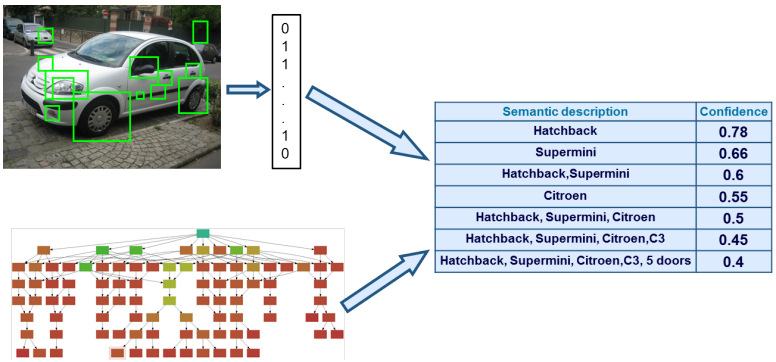


Fig. 4.9: Illustration of the semantic annotation approach expressing prediction as a list of multi-labels scored by a confidence value.

Anne-Marie Tousch in her thesis Tousch, 2010 proposed an algorithm able to produce multi-faceted predictions, i.e. assembling multiple independent semantic view points on the same data, and described a method for evaluating such an algorithm. The starting point was a semantic lattice defining all possible coherent object descriptions through inheritance and exclusion relations. This domain knowledge was used in a learning process which outputs a set of coherent explanations of the image valued by their confidence value (Fig. 4.9). The first contribution was to design this method for multiple complexity level image description. A secondary focus was to develop rigorous evaluation standards for this computer vision task, i.e. able to measure the trade-off between semantic expressiveness and accuracy.

Other approaches not for perceptual tasks

Several other lines of research have investigated the general design of interpretable models, but applied to small dimension data: rule sets (Lakkaraju et al., 2017b; Wang et al., 2017), rule lists (Wang and Rudin, 2015; Yang et al., 2017; Angelino et al., 2017), scoring systems (Zeng et al., 2017), case based reasoning (Bichindaritz and Marling, 2006; Richter and Weber, 2016), hybrid models (Wang, 2018). How they can be applied on high dimensional perceptual data is not obvious.

An unsolved issue: the evaluation of interpretability

The question of clearly evaluating the quality or usability of explanations remains an active and ill posed problem given the variety of so-called explanations, the vagueness of their objective – what they represent, who is going to use them, for what purpose. Several recent reviews argue for the necessity of defining clearer metrics and objectives (Arrieta et al., 2019; Mohseni et al., 2018; Carvalho et al., 2019).

Indeed, the state-of-the-art about the evaluation of explainability, as found in published studies, is mostly from qualitative arguments on selected examples. (Samek et al., 2017) presented a comparative method based on randomly perturbing the input image and measuring the impact on a heat map explanation. (Kindermans et al., 2016) studied the influence of noise in the explanation process and uses it to visually compare several heatmap based explanations on MNIST classification.

Several studies provide theoretical elements. (Kindermans et al., 2017) question the stability of saliency based visual explanations by showing that a simple constant shift may lead to uninterpretable representations. (Montavon et al., 2018) propose to quantify explanation quality by measuring two desirable features: continuity and selectivity of the input dimensions involved in the explanation representation. (Gilpin et al., 2018) discuss four different types of evaluation based on various trade-off points between completeness and interpretability of explanation and explanation objective.

A few studies define quantitative metrics that can be automatically computed given an augmented ground truth. (Zhang et al., 2017b) and (Bau et al., 2017) describe geometric metrics to assess the quality of the visual explanation with respect to landmarks or objects in the image. (Ancona et al., 2018) propose a quantitative but comparative evaluation metric based on computing the correlation between attribution score and random input perturbation.

Very few studies however explicitly address user-centered evaluation(Hoffman et al., 2018) for perceptual functions. (Selvaraju et al., 2017) experimentally evaluated the grad-CAM visual explanation approach on two problems (class discrimination information and trust on the explanation) using Amazon Mechanical Turk (AMT) and compare it to three other baselines (Fig. 4.10).

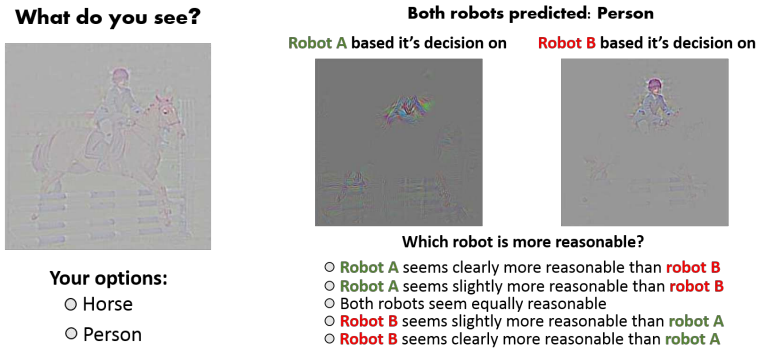


Fig. 4.10: Example of an experimental evaluation of a visual explanation. User was asked to predict the class given the explanation (left) or compare the reliability of two prediction processes from their visual attribution (right). (Selvaraju et al., 2017).

An evaluation protocol should clarify what quality is expected to be improved by explanations and for whom. Evaluation of explainability should therefore depend on the four potential types of potential explanation users defined previously:

- A scientist will expect that explanations will make her/him understand better the laws or rules underlying perceptual systems. The benefit of exploiting explanations for scientific discovery is person history dependent and makes the evaluation of their generic capacity difficult and even not meaningful.
- An engineer will expect that they will help her/him increase the accuracy and efficiency of a given system. A simple evaluation would be to compare the capacity of several types of explanations to tune or debug better given the same input data. (Nushi et al., 2017) propose a human-in-the-loop methodology to predict which fixes in a machine learning based system are most likely to improve its behavior.
- An authority will mostly exploit explanations as arguments demonstrating that the system behavior is *extensively* good: explanations should therefore aim at identifying operating domain boundaries, and be evaluated according to their capacity to express domain coverage or sensible situations.
- An end-user will expect that explanations will give her/him enough reliable information to trust predictions made by the system. A good way to measure this trust is to quantify the capacity to predict bad behavior from explanations, with or without human-in-the-loop. (Zhang et al., 2014) describe a general framework to learn how to detect failure of various vision tasks (segmentation, zero-shot learning, vanishing point estimation). (Bansal et al., 2014) learn an attribute description of failure cases given a trained visual classification system and uses it to anticipate prediction failure. (Nushi et al., 2018) describe a set of hybrid human-in-the loop and machine learning methods that facilitate the process of describing and explaining failures in machine learning systems through performance views explaining under which circumstances the system is most likely to err.

Note that (Mohseni et al., 2018) proposes another segmentation of potential users of explainability (machine learning experts, data experts, and AI novices) that can be crossed or completed with the above categories.

One example of user-centered evaluation of explainability has been proposed in the PhD Thesis of Maxime Bucher.

Failure prediction through semantic bottleneck analysis (Bucher et al., 2018)

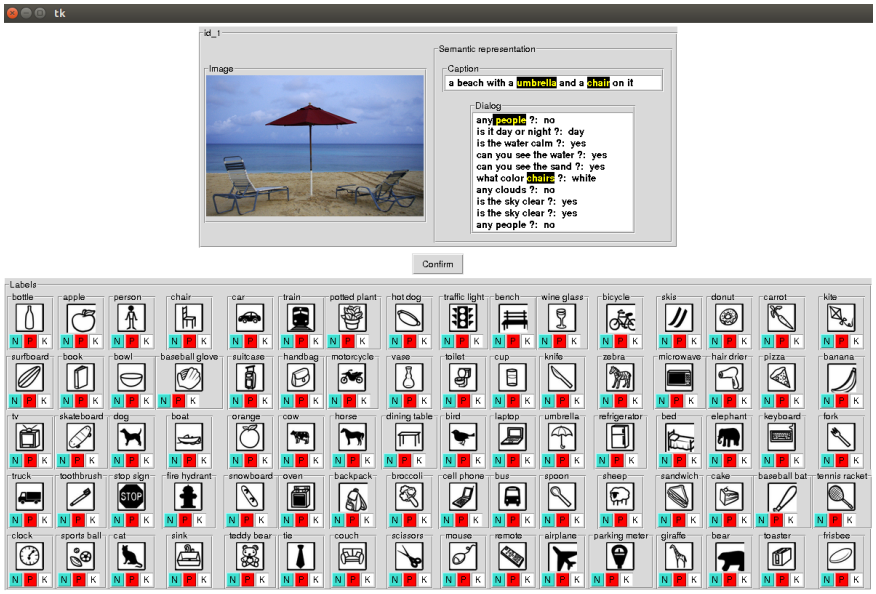


Fig. 4.11: Interface used to evaluate the capacity of the semantic bottleneck to predict prediction failure for a multi label classification task. Users were asked to anticipate labels that may be missed (False Negative) or falsely detected (False Positive) by the prediction chain based on the image and the semantic bottleneck.

In our semantic bottleneck approach, we evaluated through a dedicated interface (Fig. 4.11) the capacity of the textual representation to anticipate bad behavior, i.e. to detect potential wrong predictions. The interpretability capacity of the provided explanation is measured by the performance increment when rejected either the whole data, or only the identified label when computing accuracy. User performance is also compared with a learned automatic failure predictor exploiting the same input data. On a experiment involving 1000 images, we showed that human subjects were able to identify half of the failures with a precision of approximately 60% using the semantic bottleneck representation, whereas the automatic algorithm detected around 40% of failures with a precision of 50%.

4.3 Research directions

The question of making prediction more understandable is in several respects an old and a new problem. Old because artificial intelligence, at least its symbolic tradition, makes explicit modeling of intelligent processes a key objective. New because recent successes relying on deep neural networks which are opaque in practice due to their complexity require new explanatory and analytical tools to really understand their behavior, in both testing and training phases.

Several questions need to be clarified:

What is actually explained? Recent literature, as shown, investigates multiple paths but with no clear common objective of what is the object of explanations. Do explanations have to identify the effective informative features or do they have to describe how prediction is built from data? Should they justify, i.e. be coherent with, normal behavior, or should they be used to detect bad behavior?

For whom? Explanations will serve different purposes when targeted to a scientist, an engineer, an end-user or an authority. For each of them, the usefulness of the explanation will be different, and should be evaluated accordingly.

How? What is the best way to provide an explanation? As a symbol, a text, an image? What causal relation between the explanation and what it is expected to explain?

For what purpose? The goal of explanations is to provide meaningful understanding of system behavior, but may also be used as an information element for decision or action: control, tuning, approval, etc.

I propose to address those issues with the following research actions:

Clarification of the intelligibility objectives

The problem of explanation is a longstanding research issue in humanities and social science. (Miller, 2017) argues that the field of explainable artificial intelligence should build on this existing body of works to remove the AI specialist bias, and proposes a preliminary categorization. This first study should be completed, and extended.

One possible research direction, for instance, is to consider an explanation as a *sign*, and analyze the various explanatory schemes and objectives using tools of *semiotics* (Chandler, 2007). Signs, as discussed in chapter 2, are one of the main perceptual output types able to refer to the external world in a triadic relation involving a user; explanations as signs refer instead to the internal state of the perceptual system with some respect.

This type of research is typically a multi-disciplinary action, and should involve researchers from several fields: artificial intelligence, philosophy, psychology, and cognitive science.

User centered design

Interpretability design and evaluation

The final recipient of an explanation is a human, which as we have seen may play various roles (scientist, engineer, end-user, authority). This implies

that interpretability should be designed and evaluated from a user center perspective.

A first trend of research, as exemplified by (Olah et al., 2018) which present rather spectacular interfaces oriented to deep network interpretation, is to build efficient and *usable* interfaces able to reveal prediction process behavior for various types of users. This means that explainability solutions should include issues and findings from Human Computer Interaction (Abdul et al., 2018) domain.

A correlated action to make progress on these questions is to build shared and accepted benchmarks and protocols. This is a difficult question since user centered evaluation is not a well established routine in artificial intelligence research, except in specialized sub-field such as crowdsourcing.

Good and validated interpretability will come from multi disciplinary studies.

"Monitor, Fix and Evaluate": fast collaborative design of perceptual systems

Specifying and designing a perceptual system for a given application context, and assessing its performance, is a tedious work, event with the availability of unifying programming framework such as TensorFlow or Pytorch. Meaningful explanations that reveal the reasons for success or failure of a given system state could therefore be helpful.

Various types of explanations could be used as answers to design oriented questions like: What is the current system state? Where to act? What is the impact of action? and initiate a collaboration between user/client and system to improve performance interactively.

Monitoring the system state has been the main purpose of proposed explanability studies until now, however not with the purpose of actually exploiting them for action, to provide affordances (Norman, 2013). Several works have addressed the question of identifying critical input units of a deep network and use them for data generation (Bau et al., 2017; Bau et al., 2018). Other works act on several features of a system to evaluate their generalization (Morcos et al., 2018) or selectivity (Zhou et al., 2018a) properties. (Dalvi et al., 2019) describes a toolkit for analyzing individual neurons impact for natural language processing functions.

However, what is still lacking is an integrated user oriented strategy exploiting explanations or various visualizations of inner states and action capacities to improve design cycle.

Text as pivot computational representation

Designing good representations is often the key issue of artificial intelligence, although several researchers, as we have seen, have proposed to eliminate them. What is expected from such formal objects is expressiveness – the capacity to stand for a diversity of objects or references – and computability – the ability to support various types of formal operations, transformations and mappings.

A naive statement could assert that natural language, and its textual encoding, could be such a good representation. Fig. 4.12 illustrates various usages: human interaction, image or world description, as knowledge representation, digital encoding, etc. and shows how text can serve as an intermediate between those usages.

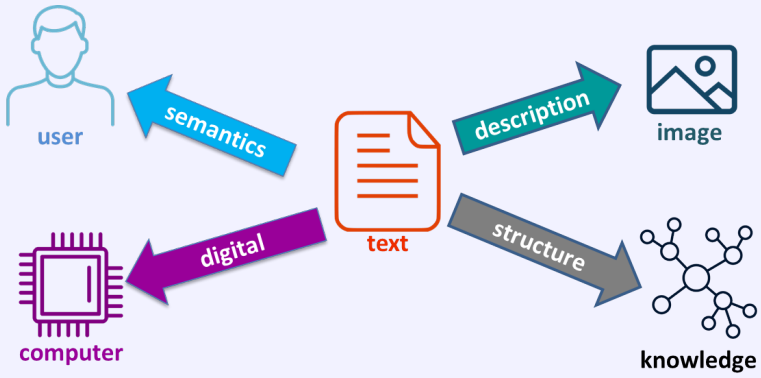


Fig. 4.12: Various usages of text as representation.

Research associating sensory data, especially image or video, and text have flourished with the rise of deep learning, mostly as multi-modal integration, with functionalities such as captioning, question answering, alignment or grounding, data generation, etc.[a] (Mogadala et al., 2019; Guo et al., 2019; Baltrušaitis et al., 2018). Many approaches have proposed efficient algorithms to achieve mappings or fusion between modalities: the "semantic gap" has never been so narrow.

A different usage of language in the perspective of giving some autonomy to a perceptual system is to rule its interaction content with the user/client (see the discussion about interactive perception pg. 81): a text can express several dimensions of the perceptual output, typically the outer world content, but also a self-assessment of its production, a justification, a level of uncertainty

("The animal on the sofa is a cat because it has fur, whiskers and is sleeping, but it can also be a small dog"). It can also be used in a dialogue to settle and negotiate the perceptual task assigned to the system.

Our first work (Bucher et al., 2018) on constraining the prediction process to host a semantic bottleneck was a preliminary proof-of-concept on two visual tasks (content-based image retrieval and multi-label classification) which requires further investigations. Several semantic problems have not been fully resolved, both in terms of reference to the input data – **what** kind of information it represents – and to the downstream prediction process – **how** is the information used.

Using an intermediate textual representation is very flexible, but the nature of what it expresses requires a better specification. One question is to study how the semantic bottleneck vocabulary can be adapted to account for potentially multiple and miscellaneous visual prediction tasks such as retrieval, classification, captioning, event detection, ego-localization, navigation, and how such representations can be used generically to anticipate behavioral problems.

[a] https://github.com/pliang279/awesome-multimodal-ml

SAFETY – Trusting APES 5

Artificial intelligence is predicted to invade our day-life, at home or at work, and there is a legitimate concern about knowing what it is actually doing, controlling how it is working, verifying that it is doing well, in order to *safely* use it.

The idea of a safe AI, and the way to achieve it, has become a issue by itself either from a technical or from a social perspective (future jobs, biases in decision making (Osoba and Welser IV, 2017)), most of the discussions being driven by three topics: ethical issues of general AI, security of personal data and autonomous driving. Safety of AI is however a beginning research area, but a real concern of the scientific community[1], probably because until recently the performance of autonomous decision making systems prevented them from being actually used, except in a very restricted situations. The integration of machine learning in the design process adds another dimension to control.

This section focuses on how safety is to be considered for APES – i.e. for autonomous or adaptive systems able to produce some information about the environment from sensors – what has been achieved, and what are the future directions of research with the underlying goal of providing a clear framework for their *certification* or normalization,[2] the definition of protocols for design and testing to deal both confidently and efficiently with those complex objects, i.e. towards a "grown-up" perception.

5.1 Problem formulation

Asserting that an artificial system is safe can be done along two lines of thought: either by *proving* it, typically through formal verification, or by *convincing* users, authorities, patients, media, etc. through evaluation, design protocol and analytic tools that it can be relied on. In practice, a formal proof of all the use cases exploiting complex structures such as perceptual systems is not achievable, and is de facto transformed into an ancillary objective of a more global picture.

To clarify what is expected to be safe, a first task is to be able to state what the purpose of the system is, and under what conditions it is supposed to be used. We therefore start by examining a specific operating domain, where safety issues have some content, in order to derive more generic problems.

[1] For instance, the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems that seeks to "advance a public discussion about how we can establish ethical and social implementations for intelligent and autonomous systems and technologies, aligning them to defined values and ethical principles that prioritize human well-being in a given cultural context." Also, the Villani report (Villani, 2018) clearly discusses ethical issues and societal impact of artificial intelligence.

[2] https://marketing.afnor.org/livre-blanc/intelligence-artificielle

Avionic safety as a reference domain

Aircraft design and operation is a domain where safety and certification has a long history and a formatted practice, and where sensors – often simple detectors or probes – are generally involved with software components. Other domains such as chemical processes, food, medical devices have also developed safety approaches, but non concerned with perceptual issues. Two other application domains, automotive industry (Janai et al., 2017; Salay and Czarnecki, 2018) and medical diagnosis may also be interested in having safe perceptual components, but have not yet developed a comparable framework.

This paragraph briefly describes the main tools and protocols proposed for avionics software safety.

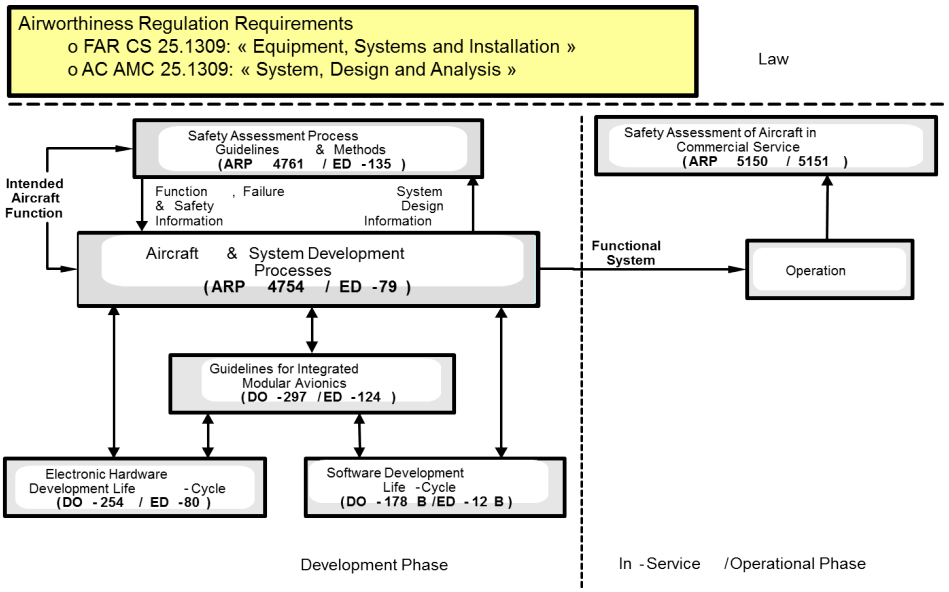


Fig. 5.1: Set of documents and logical relations between them as used for software certification in avionics.

Avionics software safety is mostly grounded in the definition of good practices for software design, development and integration agreed among administrations and industry. Fig. 5.1 shows the various documents and standards implied and their conceptual relations.

One central reference document is the *DO-178C, Software Considerations in Airborne Systems and Equipment Certification* (RTCA, 2011) which defines the "acceptable means, but not the only means, for showing compliance with the applicable airworthiness regulations for the *software* aspects of airborne systems and equipment certification" (US Federal Aviation Administration).

It distinguishes *Reliability* – i.e. the system does what it is supposed to do (no failures) – from *Safety*[3] – the system does not do what it is not supposed to do (no hazards). From a software production point of view, reliability is obtained by tracing each requirement to its implementing code and verification, with no missing functionality, whereas for safety, the objective is to trace back each piece of code to a requirement, guaranteeing no additional functionality, or "dead code".

Requirements are organized according to a hierarchy of Design Assurance Levels (DAL) for each software component. Each DAL targets a specific failure effect level (Catastrophic, Hazardous, Major, Minor, No Effect) and specifies a corresponding series of objectives to be met (between 0 to 71, see Tab. 5.1) according to the failure effect level, establishing the rigor necessary to demonstrate their compliance with safety goals.

Tab. 5.1: Failure effects and acceptable occurrence rates.

Level	Failure effect	Objectives	Failure Rate	Examples
A	Catastrophic	71	10^{-9}/hour	Flight surface controls, engine controls, etc.
B	Hazardous	69	10^{-7}/hour	Primary Flight Displays, Cabin Pressurization, etc.
C	Major	62	10^{-5}/hour	Flight Management Systems, COMM, NAV, DATALINK,etc.
D	Minor	26	10^{-3}/hour	Transponders, cabin lighting, etc.
E	No Effect	0	n/a	In-flight entertainment, satellite phone, etc.

The DO-178C document describes a series of recommendations to assert safety of avionic software, and is completed by more specific guidelines. The ARP4754A Guidelines for Development of Civil Aircraft and Systems (SAE, 2010) (Fig 5.2) targets the software development process.

The *ARP4761 Guidelines and Methods for Conducting the Safety Assessment Process on Civil Airborne Systems and Equipment* (SAE, 1996) defines a series of actions and processes for using common modeling techniques to assess the safety of a system. Functional Hazard Assessment/Analysis (FHA) is a key process whose objective is to identify all the risks that a system may encounter, their causes, and consequences. The steps to fulfill this process are the following:

1. Identification of the operational context where the system evolves.

[3] The term safety is often used without making a clear distinction between the two. We will introduce later more precise issues aiming at making system safer and more reliable, and use the term *safety* in the broad sense.

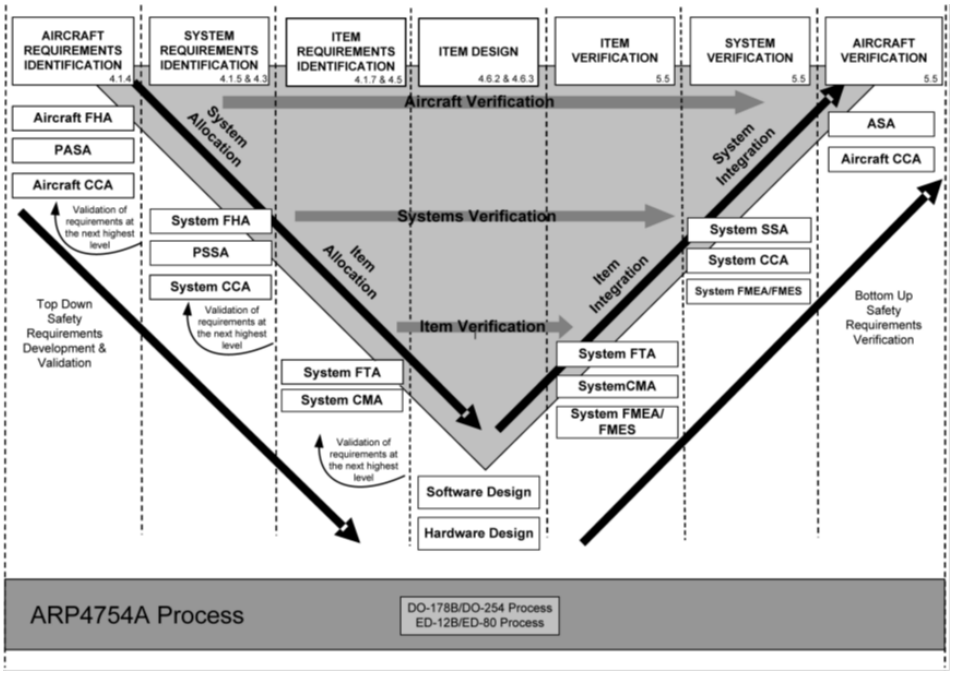


Fig. 5.2: Processes used to implement the safety guidelines for software certification.

2. Identification of potential hazards and the severity of their consequences.
3. Definition of corrective actions.
4. Verification of completeness of the resulted list of failure conditions.
5. Preliminary indications for the system architecture in order to provide mitigation means.

and results in a summarizing spreadsheet (Fig. 5.3).

These documents and processes to guide and assess safety are essentially code development oriented: their motivation is that every line of code be directly traceable to a requirement and a test routine, and that no extraneous code outside of this process be included in the build.

The next paragraph examines whether this formal approach is applicable to artificial perceptual systems, and what it may actually guarantee.

Specificity of perceptual systems

Addressing safety for perceptual systems is a rather new concern. There are of course related software issues, since until now artificial perception is essentially implemented on a computer linked to an electronic sensor. However, the main scientific question to address, and hence the major potential cause of failure if not answered satisfactorily, is to describe the way perception does what it does, *how it works*, not how it is implemented on a computer.

1. Function	2. Failure Conditions, Hazards	3. Phase, State, Mode	4. Effects on Aircraft	5. Classification	6 support	7. verification
Decelerate aircraft on the ground	**Loss of deceleration capabilities**	Landing/ RTO/ Taxi	See below			
	a) unannunciated	Landing/ RTO	Crew unable to decelerate, resulting in a high speed overrun	Catastrophic		FT
	b) annunciated	Landing	Crew selects a more suitable airport, prepares occupants for overrun	Hazardous	Emergency landing procedures	FT
	c) unannunciated	Taxi	Crew unable to stop the aircraft, resulting low speed contact with obstacles	Major		
	d) annunciated	Taxi	Crew steers the aircraft clear for anu obstacles and calls for a tug	No safety effect		
	Asymmetric deceleration	Landing/ RTO	See below			
	a) unannunciated	Landing/ RTO	Offside excursion from the runway	Major		
	b) annunciated	Landing	Crew is prepared and counters with rudder & nose wheel steering input	Minor		
	c) Asymmetric deceleration	Taxi	Slightly diverts from intended course	No safety effects		

Fig. 5.3: Functional Hazard Assessment example for a Deceleration on the Ground function.

A way to point out where the safety issue critically lies for artificial perception compared to software production, is to use the three *independent* levels of analysis of information processing system as proposed by Marr (Marr, 1982): computational, algorithmic and hardware (Tab. 5.2).

Tab. 5.2: The three levels at which any machine carrying out an information processing task must be understood (Marr, 1982).

Computational theory	**Representation and algorithm**	**Hardware implementation**
What is the goal of the computation – *the problem it solves, the* ***function*** *it implements* – why is it appropriate, and what is the logic of the strategy by which it can be carried out?	How can this computational theory be implemented? In particular, what is the representation for the input and output, and what is the algorithm for the transformation?	How can the representation and algorithm be realized physically? *On what machine or substratum?*

These three levels were primarily thought to be applied on formal models of natural vision, although Marr's work was clearly influenced by computing aspects, and had a huge impact on artificial vision models (the International Conference on Computer Vision awards a *Marr Prize Paper* every two years). (Peebles and Cooper, 2015) discusses the legacy of Marr's levels in cognitive science.

When dealing with computer vision, it would be useful to add also an intermediate *software* level between algorithmic and hardware to take into account the way

algorithms are coded and formally described in a program, and in a way that is not bound to a specific physical implementation.

As an illustrative example, let's take the simple boundary detection function in an image. The computational level can be simply defined as *detecting the relevant contours of an image*, i.e. provide a map encoding the probability that there is a contour at a given pixel. The algorithm is the way pixel values are exploited to provide this map: given the complexity of the problem, there is no single way to compute it, and a huge number of algorithms have been proposed in the literature. Fig. 5.4 shows several examples of boundary detection outputs on the same image. One can observe that for the same function, various algorithms give various results, and behave differently on the same input: the first algorithm usually gives more details than the second one, and is scored positively for the first image, but negatively for the second, when compared to the target "ideal" map.

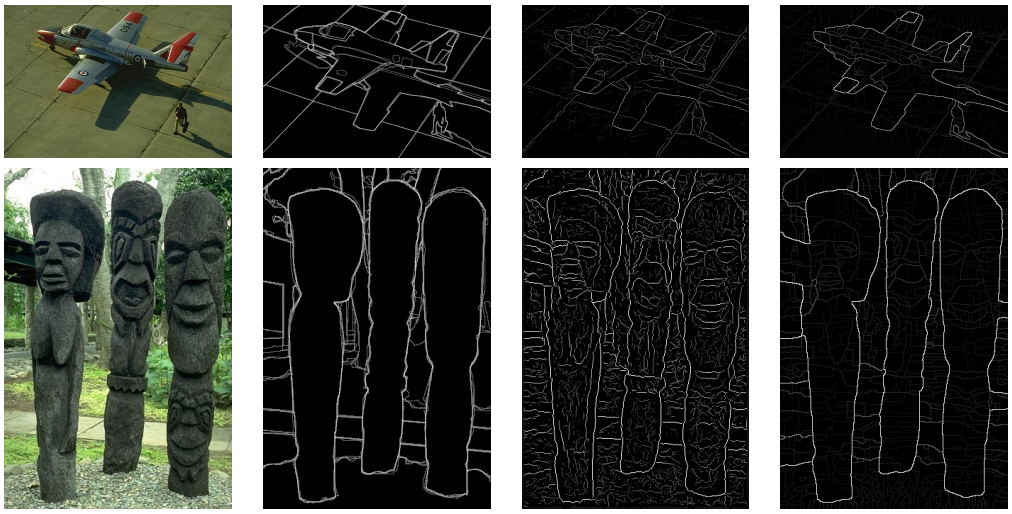


Fig. 5.4: Examples of boundary detection results for two algorithms. First column is the original image, second column is the ideal boundary map defined by several users, third column is a result from the best algorithm in average, fourth column is the second best. From the Berkeley Segmentation Dataset and Benchmark

Expressed in a safety-like vocabulary, one can state that the first algorithm of Fig. 5.4 *fails* on the second image (too many noisy contours), and the second algorithm *fails* on the first image (not enough details) because they are not consistent with the desired output defining the computational level.

By generalizing this idea, a way to assert that an information processing system is safe is to ensure that the three representation levels (or four when adding software) are *consistent*, in a sense that must be agreed upon and made specific to each couple of levels; another way to put it is to state that a system will fail when there exists some inconsistency between the levels.

When it comes to artificial perception, the main safety problem is not to assess that the software is developed according to safe guidelines and standards, but rather to state if the underlying perceptual function can actually be implemented through calculation. One can interpret the goal of most of the studies produced in the field of artificial perception as trying to make the first two levels consistent: designing algorithms able to reliably implement the function defined at the computation level (detect relevant boundaries, classify and localize objects, track objects, answer questions about visual content...) The way the algorithm is defined can exploit any available means: a learning dataset, a knowledge base, an optimal cost functions, noise or uncertainty models. . . , but is expected to produce results close to what the computational level specifies.

The software production protocols and standards for avionics do not address any algorithmic issue, i.e. they do not question the capacity of an automatic calculation to realize the function. The three last levels (algorithm, software, hardware) may be consistent, but may result in a faulty system: one may rigorously code an algorithm, compile and operate it on a given architecture, but if the algorithm doesn't solve reliably the computational problem (e.g. detect an obstacle), it may generate critical hazards.

The consistency of the three or four levels should be addressed globally: software and hardware levels, for instance, may also impose several constraints on the algorithm structure or family due to their own requirements (no randomness, single precision calculation, computation load, bounded stopping time, budgeted memory etc.) Algorithm state of the art may also restrain the functional requirements and the expected operating domain: why target a given function if there is no algorithm available able to implement it?

Ensuring the global consistency of the three (or four) description levels for perceptual systems, autonomous or not, is a difficult task, and cannot be solved using the protocols and formalism practiced for software certification. The next section examines how this issue can be addressed using a data driven perspective.

Data driven safety assessment

Datasets are now crucial components of artificial intelligence and almost mandatory to computer vision. They are of course a key ingredient of machine learning techniques, but are also used to empirically evaluate the performance of algorithms, i.e. to statistically estimate their errors. Those two usages – learning and error estimation – are common practice in machine learning where it is customary to separate learning and test sets, sometimes in a virtual way for cross validation.

The idea that failure results from an inconsistency between the computational and the algorithmic levels leads to three different problems when expressing require-

ments through datasets: function output description, operating domain coverage and evaluation of the operating discrepancy between function and algorithm.

Dataset to describe function output

In perceptual systems, except for very simple and restricted input spaces, pure analytical expressions of the computational level are in general not accessible. In most cases, the function is only accessible through samples of the joint distribution of inputs and outputs $\{(\mathbf{X}_i, \mathbf{Y}_i)\}_{i=1}^{N}$, i.e. as a *test set* if we use machine learning vocabulary. The question now translates into finding the right ways to obtain these samples, the input and output pairs.

In most of the datasets containing real data, not simulations, the expected outputs $\mathbf{Y}_i$ of the function given the input $\mathbf{X}_i$ are created by manual annotations, and rely on the hypothesis than humans host the golden standard of perceptual systems. Manual annotation is known to be tedious work, especially for low level features, and is also annotator dependent. Relying on human expertise to define the function output however generates uncertainty: Fig. 5.5 depicts examples of various annotations for boundary extraction on the same image and show that user define relevant contours with various levels of detail, yielding to an inherently random *ground-truth*.

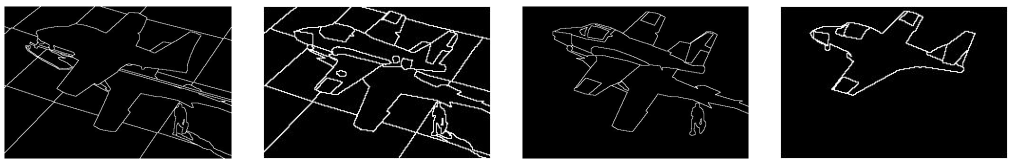


Fig. 5.5: Examples of boundary detection proposed by four different users. From the Berkeley Segmentation Dataset and Benchmark

The current practice for annotating large sets is to use crowdsourcing resources such as Amazon Mechanical Turk[4]. In such settings, the problem is to design a complete protocol ensuring that annotation is of high quality while effort is minimized (Kovashka et al., 2016). Although annotations are primarily used to define the main target output, they may also contain other hidden features useful to improve learning, e.g. the pose of an object as a supplementary information when object categorization is the functional objective.

Dataset to cover operating domain

The dual question of annotation is to select or collect the data that define the input samples $\mathbf{X}_i$ and ensure that they adequately cover the distribution of inputs that the perceptual function is likely to process.

One may think of using active learning techniques (Fu et al., 2013) to select from an available pool or from a controllable generator the data that minimizes the expected

[4] https://www.mturk.com/

error. This kind of approach has been used in many applications where annotation is expensive (crowdsourcing (Kovashka et al., 2016), supervised learning classification (Li et al., 2013)) or when data is unbalanced such as in object detection (Canevet and Fleuret, 2015) or zero-shot learning (Bucher et al., 2016a). However, the way data is collected in these approach depends on the algorithm used and may possibly reinforce its biases: it cannot guarantee that the annotated input data will cover the operating domain. All the other sorts of transfer learning or domain adaptation techniques suffer from the same bias limitation. They can be used however as a way to design the algorithm, but not to generate the data used to test it.

Another way to define the input samples that should be used to define the function is to exploit expertise on the events leading to algorithm failure. Many benchmarks have been designed to sample specific difficulties. For instance, the data used for the Visual Object Tracking challenge[5] are labeled for each frame with a tag (Occlusion, Illumination change, Object motion, Object size change, Camera motion) that is likely to cause drift. Since most of the evaluation metrics rely on average and not worst case analysis, building a dataset from identified difficulties may prevent the global performance to be dominated by easy and frequent situations.

A generalization of this way of designing test sets has been proposed in Zendel et al., 2015 as an application to computer vision of the Hazop (HAZard and OPerability analysis) methodology, used in safety analysis of large systems. The principle is to systematically list the various sources of risk as deviations of a parameter located in some part of the system and predict their impact. The various types of deviations are described using a limited vocabulary named *guide words*. Fig. 5.6 shows the various parameter locations of a computer vision system,

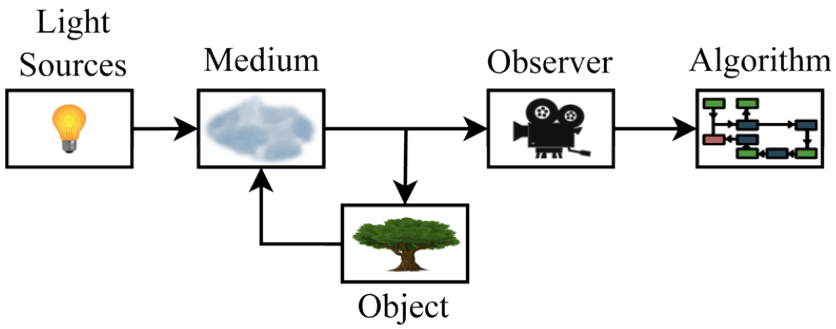


Fig. 5.6: Information flow used to define hazard locations. Light travels from the light source and the objects through the medium to the observer, which generates the image. Finally, the algorithm processes the image and provides the result. (Zendel et al., 2017b)

and Tab. 5.3 shows examples of hazard description with their associated location/-parameter/guide word. The full list containing more than 1400 hazards is available at https://vitro-testing.com/cv-hazop/.

[5] http://www.votchallenge.net/

Location/Parameter	Guide Word	Meaning	Consequences	Hazards
Light source / Intensity	More	Light source shines stronger than expected	Too much light in scene	Overexposure of lit objects
Object / Reflectance	As well as	Obj. has both shiny and dull surface	Diffuse reflection with highlight/glare	Object recognition distorted by glares
Object / Texture	No	Object has no texture	Object appears uniform	No reliable correspondences can be found
Objects / Close	Reflectance	Reflecting Obj. is closer to Observer than expected	Reflections are larger than expected	Mirrored scene taken for real
Objects / Positions	Spatial periodic	Objects are located regularly	Same kind of objects appear in a geometrically regular pattern	Individual objects are confused
Optomechanics / Aperture	Where else	Inter-lens reflections project outline of aperture	Ghosting appears in the image	Aperture projection is misinterpreted as an object
Electronics / Exposure	Less	Shorter exposure time than expected	Less light captured by sensor	Details uncorrelated due to underexposure

Tab. 5.3: Extract of the list of hazards for computer vision systems (Zendel et al., 2017b).

The HAZOP approach can be used to finely analyze the hardness of datasets. (Zendel et al., 2017a) apply it on stereo vision benchmarks (Fig. 5.7), and shows that algorithm performance is strongly correlated with frames marked as hazards. They also identify 32 types of missing hazards in datasets. Their conclusion is that benchmark design should focus on finding data instantiating hazards, rather than on increasing their size.

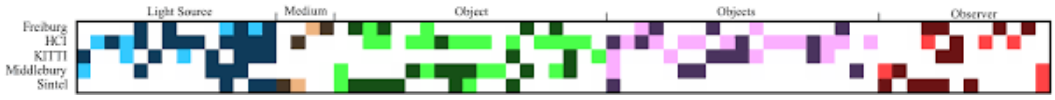


Fig. 5.7: Distribution of hazards per dataset used in stereo vision: Dark cells show identified hazards while light cells represent entries with no GT, too small area or disputed ones; color represents CV-HAZOP category (Zendel et al., 2017a).

Another way to produce data is to artificially generate it, and possibly target specific hazard where data is missing. This is done in application domain where data is scarce or expensive but with low safety requirements. The question is to estimate the bias of these data, and if they can be used to evaluate real hazard occurrence. However, it is important to make the distinction between generation used to artificially generate surrogate data for learning and simulation of corner cases to control the hardness level of benchmarks: as discussed above, statistical relevance improvement is expected from the first case, yielding to possible problem of domain adaptation, whereas the second case targets better coverage of the operating domain.

Evaluation metrics

The third issue to exploit datasets for safety assessment is to measure the discrepancy between what the algorithm is actually producing from the input data $\mathbf{X}_i$ and what it should, the $\mathbf{Y}_i$.

All datasets used as benchmark come with performance scores or metrics used to compare algorithm proposals. Most of them are statistics in the form:

$$\text{Performance score} = \frac{1}{N}\sum_{i}^{N} \mathbf{w}_i \cdot \mathbf{D}(\widehat{F}(\mathbf{X}_i), \mathbf{Y}_i) \tag{5.1}$$

where $\mathbf{D}$ is a discrepancy measure, possibly multi dimensional, comparing the actual output $\widehat{F}(\mathbf{X}_i)$ with groundtruth $\mathbf{Y}_i$, and $\mathbf{w}_i$ is a weight vector encoding failure impact of each dimension, and may also be used for re-normalizing the performance score in various ways, for instance to account for unbalanced number of samples between categories.

Most algorithm have parameters which can be used as control: this is for instance the case for detection or classification algorithms wher final decision is obtained by thresholding a score or likelihood calculated for each possible hypothesis. A usual way to produce performance measure that are independent of such controls is to average their scores:

$$\text{Performance score} = \frac{1}{N}\sum_{\lambda}\sum_{i}^{N} \mathbf{w}_i \cdot \mathbf{D}(\widehat{F}(\mathbf{X}_i, \lambda), \mathbf{Y}_i) \tag{5.2}$$

where λ is the parameter controlling the output $\widehat{F}(\mathbf{X}_i, \lambda)$.

Those scores however do not address the question of evaluating algorithm reliability for real applications in their operating context. It is hard to figure out if the best algorithm from the state of the art, i.e. the *winner* on a given benchmark, is really usable and safe. What is the semantics of the evaluation score value – e.g. is an Average Precision score of 85% for detection really good for applications?

With the ubiquity of machine learning in perceptual system design, there is also somehow a general confusion between data needed for learning, which requires statistical relevance to catch variability, and data used for testing – the benchmark – which is expected to sample both hard and easy cases. The high competition for publishing in high rank journals and conferences mostly concentrates endeavor on known competitive benchmarks without questioning their value.

The way perceptual algorithms are currently evaluated has two main weaknesses if used to assess safety.

A single score averaging miscellaneous behaviors in a single scalar, if useful to produce ranking, cannot take into account all the features of a given application context. This is why most modern benchmarks (e.g. MS COCO[6], VOT (Čehovin et al., 2016) and MOT[7] challenges) analyze algorithms along a pool of metrics, which may

[6] http://cocodataset.org/#detection-eval
[7] https://motchallenge.net/results/MOT17Det/#metrics

have a certain degree of independence. However, those supplementary analytical tools are usually statistics in the form of (5.1) or (5.2) and may numerically hide rare but catastrophic events among frequent but easy situations. Introducing worst case measures, or HAZOP analysis, as presented above, could be investigated to improve benchmarks from a safety point of view.

Another weakness of current evaluation frameworks is the difficulty of handling complex or multidimensional outputs. This is already the case in the classical object detection problem or, more recently, in dense video captioning, where the output associates textual description and numerical spatio-temporal localization: the tasks of object localization and characterization are not obviously independent? They can be correlated negatively – i.e. one cannot have simultaneously good localization and characterization – or positively – to have good characterization, it's better to have good localization, and vice versa?

With the constant improvement of algorithm performance, a current trend is to propose more versatile or multi-task perceptual systems (see pg. 90). Their evaluation requires even more complex tradeoffs or correlations to be identified and mastered to derive meaningful indicators.

Safety issues for APES

The formal framework developed for software safety in avionics presented in the previous paragraph is not fully applicable to autonomous perceptual systems. As we have suggested, the proposed protocols and standards do not guarantee the consistency between the algorithmic and computational levels of artificial perceptual system, which is the main source of failure of such systems.

Several ideas may be retained however from this framework.

The first idea is to distinguish between requirement specification and assessment – how to define a normal situation and demonstrate that the system operates in it – during the design, development and operation phases. The main question is the formal definition of what should be and can be required to define the operating domain; especially when dealing with perceptual systems, how measure discrepancies between the computational and algorithmic levels, and define acceptable bounds of their values. Formal proofs or verification procedures may be useful, but, as we shall see, cannot encompass all the aspects of perceptual systems. They are also quite difficult to develop given the way modern algorithms are designed – i.e. through machine learning – and perhaps more fundamentally given their adaptive model-free nature, and the huge dimension of their inner state and input sensory data (image, video, text, sound).

The second idea is hazard management – how to detect abnormal conditions or dreaded events and what to do in such cases. Given the contingency of the environment and the various autonomy levels allowed to the system, APES may cause, as an adaptive agent, or be the subject of, as an interactive entity, new types of risks. The main difficulty is to make the system able to detect failure conditions, and provide mitigation solutions. Machine learning techniques introduce a new actor in the picture: a learning phase, and introduces new questions such as operating domain coverage, data poisoning, malicious adversarial attacks, harmful exploration, etc. Again, when dealing with huge dimensional data, these questions suffer from the curse of dimensionality problem, and require specific tools or tricks to control it.

The last idea is that safety assessment, and certification, must be a collaborative process involving all the actors (scientists, engineers, authorities and users), and should promote the design of specific tools and protocols to do so. Formal proof of requirement fulfillment is difficult, or even impossible for adaptive agents, and a surrogate approach is to improve their *transparency* during design and operation, i.e. output interpretable representations revealing their expected behavior or possible failure. The search for transparency is to convince certifying authorities by producing good behavior reporting, but also to bring confidence to system users, either novice or expert, that they can reliably interact with it and accept it. Intelligibility, as discussed in chapter 4, is one feature of this transparency objective.

A few studies have addressed the problem of safety, mostly from a general AI perspective. (Amodei et al., 2016) discusses safety issues, namely the problem of accidents in machine learning systems, defined as unintended and harmful behavior that may emerge from poor design of real-world AI systems, with application to reinforcement learning or agents acting in a real environment (e.g. autonomous vehicles). (Yampolskiy and Spellchecker, 2016) presents future AI safety issues from cybersecurity perspective. (Papernot et al., 2016b) discusses vulnerabilities of machine learning based algorithms under an adversarial optimization framework. (Seshia et al., 2016) defines several principles for what is called *Verified AI*. (Huang et al., 2018) is a recent survey about deep network safety and trustworthiness. (Cheng et al., 2019) addresses safety of machine learning based components through an architectural point of view and proposes to reach dependability through "diverse redundancy, information fusion, and runtime monitoring". (Ashmore et al., 2019) presents a global survey about assuring Machine Learning, i.e. "generating evidence that ML is sufficiently safe for its intended use" and organizes activities in data management, model learning, verification and deployment issues. (Rahwan et al., 2019) argues for the development of an interdisciplinary research field concerned with "Machine Behavior" that would address the impact and role of AI systems in society.

The next section discusses the state or the art on these questions with a focus on perceptual systems, and is structured around three research issues:

Requirement fulfillment The goal is to answer the question "How to make sure that the instantiated algorithm actually implements the target function?" and to develop means of validating & verifying that the requirements are satisfied.
Keywords: evaluation benchmarks, adversarial example design, unknown unknowns impact limitation, formal verification.

Run-time safety The question to ask in this case is "How to prevent the algorithm from generating hazardous or unexpected behaviors?", which can be answered by developing specific functions used to detect bad operation and mitigation means.
Keywords: self-diagnosis, anomaly or novelty detection, malicious attack detection.

Certification equipment and tools The question to be answered is "How to demonstrate to users and authorities that the algorithm is correctly doing what it should?" and to propose tools able to either show that the algorithm actually performs well on the current data or that the process has been correctly designed.
Keywords: explainability or justification of predictions, output qualification or self-assessment, transparency.

5.2 State of the art

Requirement fulfillment

This section presents the tools and frameworks developed to state that perceptual systems are actually doing what they are required to do. We focus here on requirements assessing the consistency between computational and algorithmic levels.

Evaluation benchmarks

The question of benchmark quality, such as those used in academic studies, is not new: (Ponce et al., 2006) states that the "hardness of different datasets is not well understood" and identifies several issues to be addressed: annotation quality and content (semantic, geometric, viewing conditions, object pose), impact of context or extra information, rigorous evaluation protocols. (Thacker et al., 2008) gives a thorough account of evaluation practices on several tasks before the deep learning era.

Many datasets are now available thanks to the availability of modern sensors and storing capacities. The CVonline site[8] maintains a rather up-to-date list of current sets used in computer vision, showing the variety of data and annotations that have been gathered. Many datasets have dedicated sites that maintain associated leaderboards to monitor the evolution of performance. The Kaggle[9] platform organizes data science competitions and hosts several multimedia datsets.

Several specific domains have gathered large amount of data, especially to be used as learning databases. This is the case for instance for data targeting autonomous vehicle (Berkeley Deep Drive[10], Cityscape[11], Kitti[12], etc.), or remote sensing [13]. Other domains, e.g. image based medical diagnosis, however, have less furnished or very unbalanced datasets regarding machine learning requirements.

As previously discussed, the CV-HAZOP analysis (Zendel et al., 2017b) has revealed that identifying difficult cases is a key ingredient to build datasets that may evaluate safety issues. Most of the available benchmarks however do not address the explicit definition of hazards and rather favor the diversity of sources. The main reason of this situation is that data acquisition or collecting is usually opportunistic, and is not able to fully control their content.

One possibility to overcome the lack of data instantiating hazards is to simulate data. Computer graphics simulation has been used for a long time in robotics, for instance, using modern game engines (Shah et al., 2018; Mueller et al., 2017): data realism is achievable with such generators, but it essentially depends on the quality of the models fed to the engine. They are in practice very costly to create, and what is often exploited by these simulations is more the controlled diversity of situations than the realism of sense data. [14] Another simulation strategy is to exploit complex or multi-modal data acquisition, e.g. omni-directional sensors or combination of lidar and optical cameras, to generate new data with a variety of viewpoints. (Zajc et al., 2017) applies this approach to produce various visual motion patterns for single object tracking problems.

Another commonly used simulation strategy is to randomly augment existing data, usually images, by geometric or photometric transformations. More recently, techniques of style transfer have been applied to enhance data quality from low resolution models and have shown to improve performance (Shrivastava et al., 2017; Wang et al., 2018c; Atapour-Abarghouei and Breckon, 2018). The goal of both approaches,

[8] http://homepages.inf.ed.ac.uk/rbf/CVonline/Imagedbase.htm
[9] https://www.kaggle.com/
[10] http://bdd-data.berkeley.edu/
[11] https://www.cityscapes-dataset.com/
[12] http://www.cvlibs.net/datasets/kitti/
[13] https://github.com/chrieke/awesome-satellite-imagery-competitions
[14] The site `https://github.com/unrealcv/synthetic-computer-vision` lists resources exploiting simulated data for computer vision.

however, is more to increase the number of learning samples than to design a good test set.

As already stated, all benchmarks come with associated evaluation metrics aiming at measuring the discrepancy between algorithm output and required ground truth. The current trend is to compute a series of measures, possibly correlated, each one being used to address either a certain type of phenomenon or specific input data, and to select a master one for ranking. Tab. 5.4 shows several metrics used in common benchmarks for various functions.

Dataset	# measures	Names
Pascal VOC detection [a]	1	AP
Kitti object detection [b]	4	Moderate, Easy, Hard, Runtime
MS COCO detection [c]	12	AP, AP50, AP75, APS, APM, APL, AR1, AR10, AR100, ARS, ARM, ARL
Kitti road detection [d]	7	MaxF, AP, PRE, REC, FPR, FNR, Runtime
VOT challenge (2017) [e]	3	Robustness, Accuracy, Expected Average Overlap
Kitti tracking [f]	7	MOTA, MOTP, MT, ML, IDS, FRAG, Runtime
MOT challenge [g]	9	AP, MODA, MOTP, FAF, TP, FP, FN, Precision, Recall
MS COCO captioning [h]	8	CIDEr-D, METEOR, Rouge-L, BLEU-1, BLEU-2, BLEU-3, BLEU-4, SPICE

Tab. 5.4: Examples of evaluation measures used in several benchmarks.

[a] http://host.robots.ox.ac.uk/pascal/VOC/
[b] http://www.cvlibs.net/datasets/kitti/eval_object.php?obj_benchmark
[c] http://cocodataset.org/#detection-eval
[d] http://www.cvlibs.net/datasets/kitti/eval_road.php
[e] http://www.votchallenge.net/vot2017/
[f] http://www.cvlibs.net/datasets/kitti/eval_tracking.php
[g] https://motchallenge.net/results/MOT17Det/#metrics
[h] http://cocodataset.org/#captions-eval

This list points out three issues about evaluation metrics as currently practiced:

- They are multidimensional: algorithms may fail in various ways and for various types of input data, which motivates the proposition of several corresponding measures.

- The same function can be evaluated by various sets of metrics, although several benchmarks are more used than others and become de facto standard.

- Most of them are empirical means over some quantity, and do not therefore identify difficult or extreme situations that have low occurrence probability.

If the goal of evaluation benchmarks is to rank algorithms according to a clear rule of game – although questionable – available datasets and associated metrics fulfill their role. Specific metrics adapted to restricted domains can also be proposed: for instance, (Fritsch et al., 2013) focus on road detection evaluation and apply them to the Kitti benchmark.

Because it is sometimes difficult to clearly formulate and quantize the operating interpretation of a discrepancy between algorithmic and functional levels, one solution is to define it from the dataset itself. This is for instance proposed in (Cui et al., 2018) which learns a metric able to distinguish between human and machine-generated captions as a captioning evaluation measure.

Evaluation of visual object recognition tasks

Issues related to the design, modeling and evaluation of perceptual functions are commonly addressed in my activity at ONERA (Herbin et al., 2012) as a national research agency performing technical analysis for the government. Most of this work however is not public.

One public output has been the co-organization of the ROBIN competition[a] whose goal was to evaluate the performance of object recognition algorithm in various operating contexts (video surveillance, remote sensing, aerial imaging) and in cooperation with several industrial data providers (Duclos et al., 2008a; Duclos et al., 2008b). An evaluation protocol [b] has been proposed and applied on a series of 6 datasets. One of the main features of ROBIN was to orient the evaluation towards applicability, with an emphasis on taking into consideration control points and rejection capacity.

[a] http://robin.inrialpes.fr/
[b] http://robin.inrialpes.fr/robin_evaluation/downloads/ROBIN_metrics_v6.pdf

However, the real operating value of metrics is not explicitly addressed by current benchmarks: in particular, they do not allow clear assessment of algorithm safety, especially in the case of high dimension inputs.

Testing robustness – generation of adversarial examples

An important issue of safety is to state whether a perceptual system is robust to hazards that may have been known, or not, during the design or learning phase. The HAZOP methodology is an expert analysis aiming at describing generic hazardous situations, anticipating possible failures on state of the art algorithms. A complementary approach is to start from a given instantiated function and discover its possible failure cases through specific stress tests or *attacks*.

A particular and notorious approach to build hard examples for deep networks, the current state of the art approach for perceptual functions, is the existence of the so-called *adversarial* examples, i.e. inputs that are tailored to fool a system (Fig. 5.8). Those examples reveal the fact that that current deep networks, at least the way they have been learned, may be unstable and that small perturbations may have a dramatic impact on their behavior: this is a worrisome issue for safety concerns.

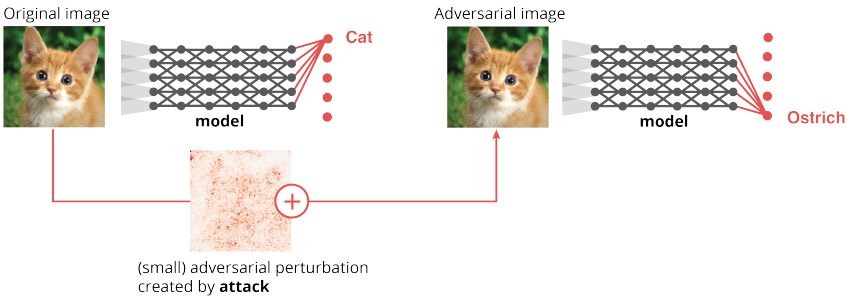


Fig. 5.8: Illustration of adversarial example paradigm. A small and visually undetectable perturbation added to the original image may drive the deep network classifier to erroneous prediction (output 'ostrich' when looking at 'cat'). From NIPS 2018 Adversarial Vision Challenge.

Since the seminal articles of Szegedy et al. (Szegedy et al., 2014) and Goodfellow et al. (Goodfellow et al., 2014b) that have identified the phenomenon, adversarial examples, both from the attacking and defending sides, have generated a huge literature in very short time. (Akhtar and Mian, 2018) is a recent survey in the computer vision domain, and contains more than 180 references. It distinguishes white box (Carlini and Wagner, 2017a) to black box (Papernot et al., 2017) strategies, universal (Moosavi-Dezfooli et al., 2017) to image specific (Moosavi-Dezfooli et al., 2016) attacks, and whether the fooled output is controlled, i.e. its output predicted class is a parameter, or not. (Xu et al., 2019) is another recent survey that addresses other application domains and other settings (learning data poisoning).(Goswami et al., 2019) gives a general panorama of adversarial defenses and mitigation means applied to face recognition. (Zhang et al., 2019c) addresses adversarial attack and defense for natural langugage processing, including multi-modal signal to text functions (captioning, VQA, OCR).

Adversarial attacks and robustness evaluation

The main and first attacked function by adversarial examples is classification, although continuous functions such as visual flow estimation (Ranjan et al., 2019), semantic segmentation (Fischer et al., 2017; Hendrik Metzen et al., 2017; Arnab et al., 2018) and detection (Xie et al., 2017) have been addressed recently.

A typical strong attack based on a source data $\mathbf{x}$ that one wants to erroneously classify as t is obtained as an optimization (Carlini and Wagner, 2017b):

$$\text{attack}(\mathbf{x}, t) = \underset{\mathbf{x}'}{\text{argmin}} \|\mathbf{x}' - \mathbf{x}\|^2 + c \cdot l_\kappa(\mathbf{x}', t) \tag{5.3}$$

where the loss function l is defined as $l_\kappa(\mathbf{x}', t) = \max(\max\{Z_i(\mathbf{x}') : i \neq t\} - Z_t(\mathbf{x}'), -\kappa)$, and $Z_i(\mathbf{x}')$ is the logit function used to classify data as class i, and κ is a margin used to force the wrong class output logit to be arbitrarily high. (Carlini and Wagner, 2017a) experimentally studied the resilience of several attack detectors and showed that this way of forging adversarial attacks was able to bypass ten of them. This approach is considered as being among the strongest so called "white box" attacks, i.e. attacks that have access to the entire prediction algorithm, its weights and architecture.

In this strategy, adversarial examples are specifically tailored to fool a *known* system. Adversarial examples may also threat – be transfered to – another classifier (Papernot et al., 2016c; Liu et al., 2016c), which allows the estimation of simpler surrogate classifiers of a black-box system (e.g. an SVM instead of a deep network) that can be used as a model to find adversarial examples.

The basic explanation of the existence of adversarial networks is that small perturbation distributed over input signal may be cumulated and emphasized by deep networks that have a large Lipschitz constant (Cisse et al., 2017). This also means that the fundamental signal that is exploited by most of the attacks is the gradient w.r.t the input as a measure of data sensitivity. Many other types of attacks have been proposed (see the cited surveys), but all rely on an optimizing locally or globally a data based criterion to search the most impactful input example, and exploit gradient computation, or its approximation.

The evaluation of attack strength can be done according to two dual measures. The first one estimates the average smallest perturbation $\rho(f)$ able to fool the classification function $f(\mathbf{x})$ and can be expressed as:

$$\rho(f) = \mathbb{E}_{\mathbf{x},y} \left[\min \|\delta\|_p \text{ s.t. } f(\mathbf{x} + \delta) \neq y\right]$$

where the norm of the perturbation $\|.\|_p$ is defined for $p \in \{1, 2, \infty\}$

The second measure estimates the impact of perturbing input data by less than ϵ on the prediction loss $l(y, y')$:

$$\mathcal{R}_{\text{adv}}(f, \epsilon) = \mathbb{E}_{\mathbf{x},y} \left[\sup_{\|\delta\|_p \leq \epsilon} l(f(\mathbf{x} + \delta), y) \right]$$

Those two adversarial criterion do not have the same usage. The robustness ρ is a measure of classification sensitivity to perturbations, and qualifies the whole

function. The second measure is more an optimization criterion, that can lead to regularized criteria such as those of Eq. 5.3, and to distribution free (Diochnos et al., 2019; Attias et al., 2019) or distribution dependent (Yin et al., 2019) bounds to control the gap between empirical and real adversarial risks.

Note that perturbations may not stay on data manifold (the $\mathbf{x}'$ is not constrained to be sampled from original data, and δ is isotropic): they may be completely artificial and not observed in nature.

A second important point is that those measures are local and are not, by construction, related to the global risk that measures accuracy:

$$\mathcal{R}(f) = \mathbb{E}_{\mathbf{x},y}\left[l(f(\mathbf{x}), y)\right]$$

The relation between the two has been discussed recently where several studies argue that accuracy and robustness are related antagonistic phenomena (Tsipras et al., 2019) (Su et al., 2018) that can lead to mixed optimization criteria (Zhang et al., 2019a), whereas others separate the two by distinguishing between on and out data manifold adversarial examples (Stutz et al., 2019).

Adversarial defenses

The discovery of adversarial examples has motivated the development of defense techniques to improve robustness. Two principles have been proposed: data augmentation and regularization.

Adversarial examples can be interpreted as unknown learning biases. One way therefore to counteract their impact is to augment the learning database by artificially generating adversarial examples (Kurakin et al., 2017). This *adversarial learning* strategy generates one of the best defenses against adversarial attacks, but can be costly: there are potentially many possible perturbations per example that would fool the predictive system. (Sinha et al., 2018) augments model parameter updates with worst-case perturbations of training data in a Wasserstein ball. (Tramèr et al., 2018) studies a technique that augments training data with perturbations transferred from other models. Another data augmentation strategy is to interpolate between available data and their labels as a mixing regularizer (Zhang et al., 2018a). (Tramèr and Boneh, 2019) discusses the difficulty of handling multiple adversarial perturbation types in training and proposes new training criteria fusing several adversarial risks. (Shafahi et al., 2019a; Wong et al., 2020) propose faster ways to achieve adversarial training.

A second type of strategy is to improve the learning step itself, typically by adding more layers/sub-networks, by changing the loss/activation functions, etc. (Cisse et al., 2017) controls the Lipschitz constant of each layer through regularization. (Madry et al., 2018) studies the adversarial robustness of neural networks through a

robust optimization perspective. (Papernot et al., 2016a; Papernot and McDaniel, 2017) exploit the notion of *distillation*, i.e. the extraction of class probability vectors produced by a first model to train a second one of reduced dimensionality without loss of accuracy, to generate more regularized deep networks.

However, as (Goodfellow et al., 2018) states it, "few strong countermeasures exist for the many attacks that have been demonstrated". Detecting attacks to start mitigation means or to prevent error propagation instead of trying to counter them is an alternative strategy that will be described in a following section.

Real world attacks?

Whether adversarial examples are a real threat for real-world or embedded applications is however still a debated question. (Evtimov et al., 2017) describe real world attacks and shows that simple stickers put on road signs may fool the classifier for various viewing conditions (Fig. 5.9). (Sitawarin et al., 2018) is another recent study that shows examples on black-box attacks on road signs. Other attacks fool object detectors by inserting in the scene an adversarial patch close to the object (Brown et al., 2017). (Jan et al., 2019) generates adversarial examples against deep neural networks by explicitly modeling the digital-to-physical transformation.

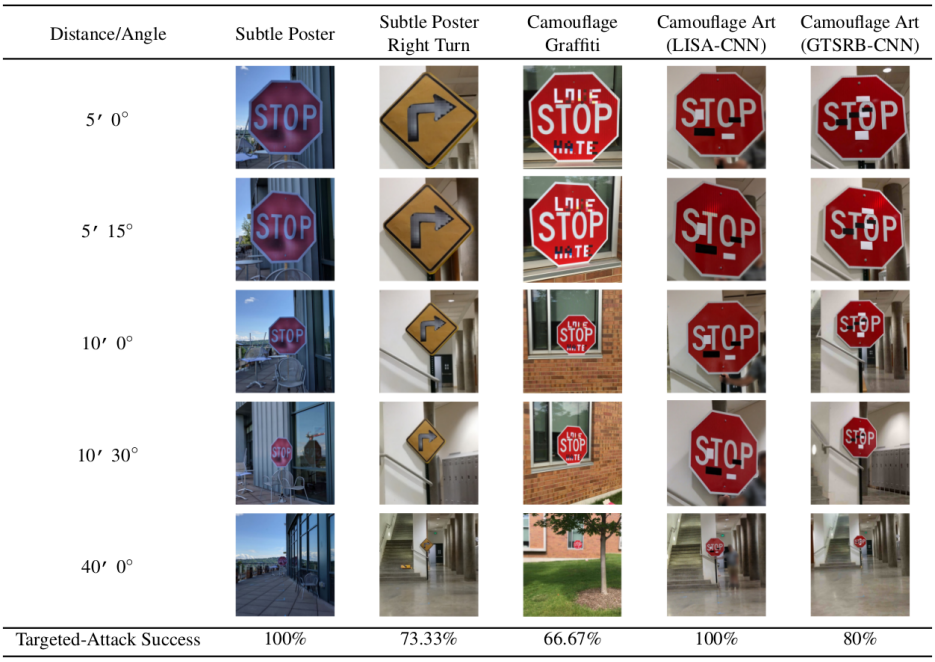


Fig. 5.9: Example of modified real world objects fooling a known deep network used to classify road signs (Evtimov et al., 2017).

However, some advocate that the theoretical existence of such phenomenon is not critical for embedded applications such as autonomous driving (Lu et al., 2017a), especially for object detection (Lu et al., 2017c) where the technique proposed in (Evtimov et al., 2017) is hard to reproduce. More global and simple perturbations,

however, may induce wrong decisions: (Afifi and Brown, 2019) studies the impact of incorrect color constancy on classification and semantic segmentation based on deep networks.

Given the maturity of this research domain, it is hard to say if adversarial examples are a real concern for safety issues or if their occurrence in real situations is negligible compared to other hazards (Gilmer et al., 2018).

Besides the theoretical issues raised by the existence of adversarial instabilities, one possible use of this already large body of techniques developed may be used to tailor benchmarks of various difficulty levels or simply to improve robustness of algorithms.

The high interest of the research community has promoted several challenges on designing defense methods against adversarial attacks: for instance, NIPS 2017: Defense Against Adversarial Challenge Attack[15] and NIPS 2018 Adversarial Vision Challenge[16]. Benchmarks in these competitions are usually of medium size (number of samples and data dimension): cifar-10, MNIST, Tiny ImageNet, Traffic sign [17]. Those challenges often come with adversarial example generation toolboxes such as the adversarial robustness toolbox[18](Nicolae et al., 2018) or CleverHans library [19] as baselines. A thorough benchmarking action is proposed in (Su et al., 2018) with the objective of examining the existence of empirical trade-offs between robustness and accuracy using multiple robustness metrics, including distortion, success rate and transferability of adversarial examples [20]. Their conclusion is that low error networks are highly vulnerable to adversarial attacks and that network architecture has a larger impact on robustness than model size.

As regularly mentioned in papers, attacks such as those calculated using Eq. 5.3 fool most of currently proposed defenses, but are also increasingly detected. As (Goodfellow et al., 2018) asks, "can we expect an arms race with attackers and defenders repeatedly seizing the upper hand in turn?", as is for instance instantiated in the NIPS 2018 Adversarial Vision Challenge.

A critical question regarding the safety of APES would be to know whether attacks and defenses can be universal, and in what sense. The origin of the existence of adversarial example, however, is unclear. Several authors hypothesize that accurate learned classifiers may in fact catch non-robust but informative latent features to make their prediction (Tsipras et al., 2019; Ilyas et al., 2019; Joe et al., 2019), implying that adversarial examples are possible because the resulting classifier has

[15]https://www.kaggle.com/c/nips-2017-defense-against-adversarial-attack
[16]https://www.crowdai.org/challenges/adversarial-vision-challenge
[17]http://benchmark.ini.rub.de/index.php?section=gtsrb&subsection=dataset
[18]https://github.com/IBM/adversarial-robustness-toolbox
[19]https://github.com/tensorflow/cleverhans
[20]https://github.com/huanzhang12/Adversarial_Survey

not been able to apprehend the "good" regularities in data. Other authors impute the existence of adversarials to the data distribution itself and its geometry (Khoury and Hadfield-Menell, 2019; Khoury and Hadfield-Menell, 2018), which leads to non intuitive concentration of measure phenomena (Shafahi et al., 2019b; Mahloujifar et al., 2019), and could partially explain how attacks can also be performed on simple non-deep classifiers such as k-nearest neighbors (Sitawarin and Wagner, 2019b; Sitawarin and Wagner, 2019a).

All those recent studies point out the fact that many phenomena encountered in deep learning are not well understood, and that the objective of ensuring safe learning is not yet achieved.

Testing robustness – unknown unknowns

Machine learning based prediction systems, or more generally AI systems, have by essence a limited knowledge of the world in which they will live: they entirely depend on the learning database and on the underlying regularity assumptions that are exploited in the design phase. As Dietterich puts it (Dietterich, 2017) "An AI system must act without having a complete model of the world". (Boult et al., 2019) states that solutions able to handle unknowns are insufficient, especially for deep network predictors.

There are two reasons why limited knowledge may produce erroneous predictions: the model is too uncertain for reliable prediction, or the system predicts from wrong, invalid or non existing grounds. The first case can be dealt with supplementary elements describing the prediction uncertainty level (confidence coefficient or variance). The second case is more problematic since the system doesn't know it may be wrong and outputs false predictions with high confidence. This last case has been called *unknown unknowns* (UU).

Looking for UU's should be one dimension of the safety assessment of predictive systems. Very little work has been done in this direction: the main studies are (Lakkaraju et al., 2017a; Bansal and Weld, 2018) which describe sequential sampling strategies applied to the prediction system considered as a black box.

Outlier, novelty or out-of-distribution detection approaches may also be used to identify UU's: they are mostly use to detect failure on line (see section 5.2) and usually require the availability of the training database.

Testing robustness – domain coverage

The previous paragraphs addressed robustness issues by trying to find specific error generating cases, with various levels of universality or transferability. A dual concern is to ensure that all operating situations have been examined: the question is to design strategies able to *cover* them with a certain confidence level.

The aim of the CV-HAZOP approach (Zendel et al., 2017b), presented previously, was precisely to try to exhaustively anticipate hazardous situations through an expert analysis: this analysis is not algorithm dependent, or in a rather loose way – computer vision experts know in which general circumstances algorithms are likely to fail.

Recent works have tried to define algorithm dependent policies aiming at ensuring their operational value. Several studies have been inspired by coverage criteria as practiced in software testing: (Pei et al., 2017) and (Tian et al., 2018) exploit neuron activity coverage on a known deep network (*white-box* approach) as a way to generate incorrect "corner cases". (Wu et al., 2019) describes a strategy for generating corner cases, i.e. data that defines the boundaries of good functioning domain, by applying a series of image transformations to the learning database. (Ma et al., 2018a) defines bad behavior criteria for deep network based on testing if neuron output values belong to sets of real valued intervals. (Gopinath et al., 2017) proposes a data-guided approach for automatically identifying safe regions of the input space, within which the network is robust against adversarial perturbations, using verification techniques (SMT). (Wicker et al., 2018) proposes a feature-guided black-box algorithm for evaluating the resilience of deep neural networks against adversarial examples. (Odena and Goodfellow, 2018) adapts techniques from software engineering (coverage-guided fuzzing) to find numerical issues in trained neural networks, disagreements between neural networks and their quantized versions, and undesirable behaviors in character level language models.

Many studies have addressed safety issues in the context of autonomous driving. (Dreossi et al., 2017) describes an image generator that produces synthetic pictures by sampling in a lower dimension image modification subspace to test the deep network used to predict driving commands. (Zhang et al., 2018c) presents an unsupervised framework to automatically produce large amounts of driving scenes through Generative Adversarial Networks to test the consistency of driving behavior.

Formal verification and proofs

Deep networks are rather complex objects: their behavior is not fully understood, and there are not definite results stating the impact of optimization, architecture, data sets on performance stability and accuracy. However, several approaches have tried to adapt several formal results or practice of "validation & verification" techniques.

A first series of methods makes use of verification algorithms to evaluate the stability of network, i.e. their output invariance to perturbations at a given operating point. (Huang et al., 2017d) presents work on verifying the absence of adversarial inputs in generic feed-forward multi-layer neural networks using Satisfiability Modulo Theory (SMT), while (Katz et al., 2017) develops Reluplex, a simplex formulation of local

invariance for networks combining linear and ReLU type non linearities. (Tjeng and Tedrake, 2017) formulates verification of piecewise-linear neural networks as a mixed integer program. Those verification processes are exponential in the number of features, and their scaling for large images is an issue. (Wang et al., 2018b) addresses the scaling issue using interval analysis and linear relaxations. Other works that improve property verification scaling are (Singh et al., 2019b)(Salman et al., 2019)(Singh et al., 2019a). (Hains et al., 2018; Liu et al., 2019) present a general recent account of formal methods developed to assess safety of deep networks, and conclude that "there is a trade-off between completeness of a verification algorithm and its scalability. Complete algorithms run slower on larger networks, while incomplete algorithms are more conservative."

A second series of studies examines global network from a functional point of view, and measure stability through an evaluation of their Lipschitz constant (Scaman and Virmaux, 2018; Weng et al., 2018; Fazlyab et al., 2019).

Finally, (Cullina et al., 2018) takes a statistical learning perspective and extend the Probably Approximately Correct (PAC)-learning framework to account for the presence of adversaries. (Varshney, 2016; Varshney and Alemzadeh, 2017) formally define machine learning safety in terms of risk, epistemic uncertainty, and the harm incurred by unwanted outcomes.

Those method are related to the emerging topic of *Verified AI* which proposes to extend the current validation & verification practices to AI (Menzies and Pecheur, 2005). Seshia et al. (Seshia et al., 2016) identified five main challenges from a formal method perspective (environment modeling, formal specification, system modeling, computational engines, and correct-by-construction design) and defined several corresponding design principles:

"

1. *Introspect* on the system (i.e. identify assumptions that the system makes about the environment that are sufficient to guarantee the satisfaction of a given requirement) and actively gather data to model the environment;
2. Formally specify *end-to-end behavior* of the AI-based system, and develop new quantitative formalisms to specify learning components;
3. Develop *abstractions* for and *explanations* from Machine Learning components;
4. Create a new class of *randomized and quantitative formal methods* for data generation, testing, and verification;
5. Develop techniques for *formal inductive synthesis* of AI-based systems, supported by an *integrated design methodology* combining design-time and run-time verification.

"

Those principles target generic AI systems and are general, with a twist towards model-based approaches as a prerequisite of many formal methods. The question whether they are relevant to APES is open since modern perceptual algorithms are mostly data-driven designed.

Run-time safety

The techniques described previously aim at characterizing a given system *before* it is actually operated. They exploit potentially hazardous input data or an analysis of the algorithm architecture. Another strategy to improve safety is to develop methods detecting potential problems *during* operation. Here again, one can make a distinction between methods trying to characterize input data, and those that exploit knowledge of the algorithm to diagnose bad behavior. The new challenge of deep network instability, revealed by the existence of adversarial examples, has motivated the development of specific defense approaches.

Anomaly or novelty detection

A *safe* system should be able to warn its user when there is a risk of catastrophic consequences when exploiting a false prediction and suggest to *reject* it. In a prediction system, there are mainly two causes of rejection: uncertainty – the input data can be meaningfully associated to more than one prediction – or novelty – the input data has not been considered during the design phase or is abnormal with respect to the underlying models governing prediction. We focus on this last case in this section.

Novelty, anomaly or outlier detection are synonyms of the same formal problem: to decide whether a given data belongs to an underlying known distribution, usually described as samples or characteristic prototypes. It does not address the question of designing a system that is robust to anomaly or outlier but aims at equipping a predictor with an explicit rejection capacity or out-of-distribution detector. In machine learning, this problem is also named “one-class classification”. Introducing a supplementary rejection class in a global decision process is sometimes referred to selective classification (Geifman and El-Yaniv, 2017; Geifman and El-Yaniv, 2019), the question being to control the good operating trade-off between decision and error rates. Note that the expression “anomaly detection” sometimes refers to a way of building “saliency” detectors (Borji et al., 2015) – an anomaly being a pattern considered different from most of the others – and not in the sense of building a rejection process.

Novelty detection is not a new problem, and is used in many applications, for instance in data stream analysis to detect intrusion (see (Chandola et al., 2009; Markou and Singh, 2003; Zimek et al., 2012; Pimentel et al., 2014; Akoglu et al., 2015) for various surveys). However, when data is highly dimensional, like images,

applying generic methods is not powerful enough, and depends on a projection on a much lower dimension feature space, using for instance Principal Component Analysis (PCA), auto-encoders or non-linear kernels, to make statistically relevant inferences. (Zimek et al., 2012) discusses the issue of high dimension and its relation to the curse of dimensionality phenomenon.

A first strategy to detect anomaly or novelty is to find ways to apprehend the extension and structure of the data manifold. This is what deep learning is expected to do, either for generic tasks (classification) or to specifically improve anomaly detection. (Chalapathy et al., 2017) describes a robust auto encoder that learns a nonlinear subspace that captures the majority of data points, while allowing for some data to have arbitrary corruption, and evaluates the approach on three image datasets. (Zhai et al., 2016) investigates two decision criteria (energy score and reconstruction error) for performing anomaly detection from an energy based distribution representation computed on a deep network architecture. (Erfani et al., 2016) presents a hybrid model where an unsupervised deep belief network (DBN) is trained to extract generic underlying features, and a one-class SVM is trained from the features learned by the DBN. (Ruff et al., 2018) extends one-class support vector approach to deep network, using the same concept of minimum volume hypersphere boundary.

Another series of works exploit or modify the output scores before decision, and use them to detect out-of-distribution data coming from datasets that contain classes different from those found in the in-distribution. (Hendrycks and Gimpel, 2017) shows the performance of a baseline approach on several datasets relying on the idea that correctly classified examples tend to have greater maximum softmax probabilities than erroneously classified and out-of-distribution examples, allowing for their detection. (Liang et al., 2018) describes a method improving the detectability of out-of-distribution from the output scores by adding small perturbation to the input and output temperature scaling. (DeVries and Taylor, 2018) proposes a method that learns a confidence score jointly with the actual prediction by retraining the last layer of a classification network, and uses it on the task of out-of-distribution detection. (Mandelbaum and Weinshall, 2017) also learns a confidence coefficient from inner layers of a classification network and prediction but with another loss measuring pairwise distance between different classes. (Lee et al., 2018a) exploit hierarchical class structure to detect data coming from new classes using confidence-calibrated classifiers, data relabeling, and leave-one-out strategy for modeling novel classes under the hierarchical taxonomy. Note that all these approaches rely on detecting bad score prediction behavior and therefore require that the underlying data distribution is structured in classes.

As a binary decision problem, the evaluation of novelty detection algorithms depends on measures of false positive/false negative tradeoffs (AUC under ROC curve,

Precision at given Recall). Most of evaluation frameworks exploit data acquired from "real" situations, e.g. by labelling several classes as outliers, or importing other datasets of similar origin and label it as novel (Cifar-10 vs. Imagenet). Algorithms are believed to be more fairly compared under such settings. (Campos et al., 2016; Swersky et al., 2016) discuss the suitability of available benchmarks (datasets and metrics) and compare several algorithms using such metrics. Their evaluation however is limited to low dimensional data, and whether their conclusion scales to higher dimensional perceptual data is open. (Snoek et al., 2019) describes the results of a large-scale benchmark on classification problem and investigates the effect of dataset shift on accuracy and calibration.

It is however difficult to tell using such evaluation approaches if the state of the art of novelty detection algorithms is mature enough to assess on-line safety of APES. A fundamental question is to define what is a realistic or useful out-distribution to validate novelty detection: How different should it be from the in-distribution? What are hazardous situations likely to happen? The design of adversarial data (see 5.2) is a complementary way to build evaluation in a way that is more, perhaps too, specialized to instantiated algorithms. The next section will present in more details how to build defenses against adversarial attacks.

Detecting adversarial examples

Countering attacks, i.e. producing mitigating means to correct the impact of malicious perturbations on the decision, is difficult as we have seen. Whether it is possible remains a question. However, one can instead try to detect when such attack happens and delay the final prediction to other actors. This is typically an anomaly detection problem but dedicated to adversarially perturbed input data.

The main difference with anomaly detection is the malicious intention of the attack: data are generated purposively to fool the system. Various settings can be defined according to the attacker knowledge (white/black box, access to learning data or not).

Most of the studies addressing adversarial detection analyze behavioral difference of a known network when activated by normal or perturbed data. (Gong et al., 2017; Metzen et al., 2017) learn a separate binary adversarial detector from a set of generated attacks. (Grosse et al., 2017) builds a two sample statistical test to separate benign and corrupted data. (Lu et al., 2017b) learns a Radial Basis Function SVM to detect out of distribution data from the last stages of a deep network where adversarial examples are expected to have the most different behavior. (Li and Li, 2017) defines a cascade classifier from convolutional filter outputs of various layers in a deep network.

Another type of strategy modifies training or input data to make attacks more salient. (Dathathri et al., 2018) detects adversarial examples by testing the validity of Neural Fingerprints, a set of fixed perturbations that are expected to have a controlled behavior when added to a real data and not when added to an adversarial example. (Xu et al., 2017) uses feature squeezed (pixel encoding depth reduction and spatial smoothing) data to compare predictions from the original and the squeezed images. If a large difference is found, the image is considered to be an adversarial example. (Akhtar et al., 2018) learns a specific network able to rectify a perturbed data, and detects universal attacks by analyzing the difference between the original image and the rectified one.

Other approaches exploit the same intuition that adversarial examples are far from the manifold of clean data and can be identified by out-of-distribution method in a given subspace spanned by inner activation layers of a deep neural network. (Meng and Chen, 2017) detects adversarial examples by projecting the data to the learned manifold of clean images. (Feinman et al., 2017) uses kernel density estimates and Bayesian uncertainty through drop-out to detect out of distribution adversarial data. (Ma et al., 2018b) uses local intrinsic dimension estimation of adversarial regions and apply it to the detection of adversarial examples. (Lee et al., 2018b) proposes a method for detecting any abnormal samples based on computing the Mahalanobis distance between class conditional Gaussian distributions with respect to (low- and upper-level) features of the deep models obtained through Gaussian discriminant analysis.

A complementary question to adversarial detection is to ensure that a given prediction cannot be impaired by any bounded perturbation for given data points. This is the objective of several formal methods that have been developed to compute such data relative bounds as described pg. 162, and that can be used as a test for adversarial detection, the value of the bound indicating the risk of accepting the prediction. (Kolter and Wong, 2017; Wong et al., 2018) exploits a convex outer approximation of the set of activations reachable through a norm-bounded perturbation for piece-wise linear activation function networks. This idea has been extended to non-linear activation functions (Singh et al., 2018)(Zhang et al., 2018b). (Jordan et al., 2019) provides tighter bounds on the average case. (Balunovic et al., 2019) extends verification to geometric transformation (rotation and translation) and solves it using a combination of sampling and optimization to compute asymptotically optimal linear constraints. (Cohen et al., 2019) injects Gaussian noise in the input data to produce a smoothed random classifier that can be guaranteed to be accurate under any perturbation bounded by a computable radius.

Self-diagnosis and recovery

Self-diagnosis is the capacity of a system to detect bad behavior by probing its inner states. Novelty or adversarial attack detection, as described previously, are potential candidates to fulfill this goal for predictive systems.

Self-diagnosis requires some kind of *introspection*, i.e. a way to express and represent in a common formalism what is happening inside the system. The role of intelligibility for perceptual prediction (see chapter 4) will be discussed in more details later as a general way to provide insights about system behavior. We are more interested here in examining how self-diagnosis can be used on-line to improve robustness and make the predictive system safer.

The big question is how to integrate self-diagnosis capacity in a general processing framework so as to improve safety and to help the system recover from its detected failures. In other words, what action to make once failure has been detected?

There are mainly two families of algorithms that may profitably make use of on-line self-diagnosis: vision based dynamical systems implementing functions such as tracking, navigation, SLAM, etc. which, because of their sequential nature, can incrementally correct their behavior; ensemble methods that may weigh or select the contributions of each of the components in a fusion step according to their failure prognostic or confidence.

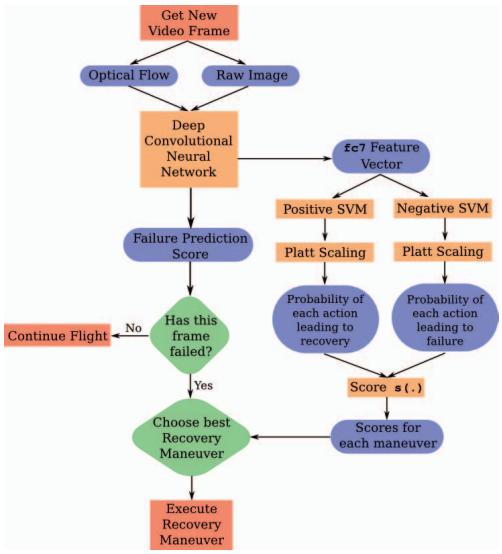


Fig. 5.10: Block diagram of a flight controller exploiting vision failure detection. (Saxena et al., 2017).

Very few studies take into account explicitly vision failure in a global processing pipeline. Most of them target autonomous driving (car or UAV) as application. (Saxena et al., 2017) learns a recovery maneuver predictor (translate right, translate left, rotate right, rotate left) applied when an uncertain input image is detected, i.e. when a classifier is unable to ascertain whether the scene is collision free vs.

collision prone, for example in case of improper illumination (see Fig. 5.10). The role of the maneuver is to drive the vehicle to a location or an altitude where the input data can be safely interpreted. (Khan and Hebert, 2018) ensures UAV flight safety by choosing the best trajectory leading to past non failed states after failure has occurred. (Richter and Roy, 2017) exploits an autoencoder to declare visual scenes familiar or not during a robot navigation task that collects new images to learn a collision predictor. The detection of novelty makes the robot switch to a safer behavior (lower speed). (Hecker et al., 2018) learns a "Scene Drivability" predictor that decides whether the environment is safe or hazardous for an automated car to navigate using an automated image-based maneuver generator (speed and steering), and let the driver take over sufficiently ahead of time.

In Isabelle Leang thesis, we have been interested in studying failure predictor and recovery means for the fusion of single object trackers.

Online drift prediction for fusion of single object trackers (Leang et al., 2015; Leang et al.,

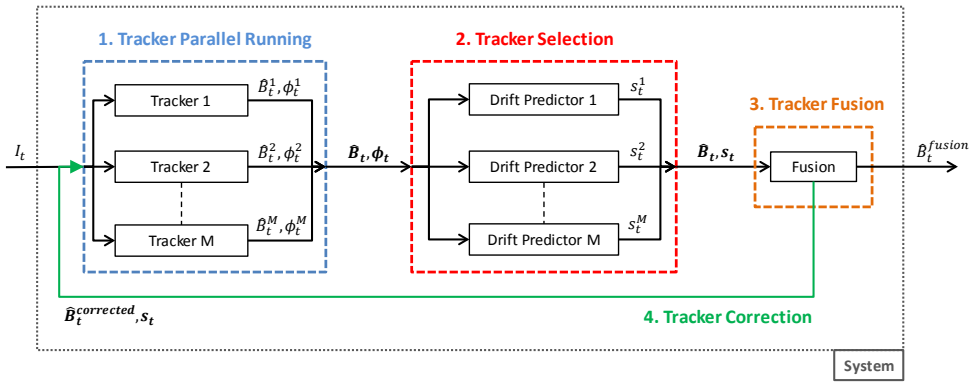


Fig. 5.11: Block diagram for multiple tracker fusion.

The work focused on the design of good strategies for the on-line fusion of trackers. The emphasis was on controlling the overall robustness of tracking measured as a number of drifting events, i.e. the number of times the target is lost when applied on a given database. Trackers deal with critical situations differently (illumination, occlusion, appearance changes, camera motion); the idea was to exploit their complementarity on various fusion strategies.

Fusion can operate at two levels: by selecting the appropriate set of good trackers and/or by correcting either their output or their inner state (Fig. 5.11). Drift prediction based on various features has been proposed and more specifically studied as a key component of the selecting step. The overall fusion strategies resulted in 46 different schemes that have been extensively evaluated on 4 databases (VOT2013+, VOT2015, VOT-TIR2015 and OTB-100) and a repertoire of 9 trackers with available source code (NCC, KLT, CT, STRUCK, DPM, DSST, MS, ASMS, CCOT).

The results of the experiments has been summarized as a series of recommendations (What trackers use? What to fuse and how?) when trying to apply on-line fusion given a target database or application context and a set of trackers with their individual robustness evaluation on the database:

1. Fusion is helpful when fusing trackers with comparable individual performance (robustness) and gives an important gain. By contrast, fusing very heterogeneous trackers can be harmful when noisy outputs contaminate the other trackers and degrade their behavior.

2. A selection step is useful, the simplest methods based on bounding box reasoning – temporal filtering and consensus – leading to comparable results to more specific methods trying to give independently a hint of each individual tracker behavior (score or likelihood maps).

3. The correction step is sensitive to individual tracker behaviors: passive fusion cannot recover from target confusion, and active fusion may be contaminated by bad target localization.

Fusion performance also depends on tracker complementarity besides their individual performance. To quantify the complementarity of a set of trackers, we defined an incompleteness measure based on off-line individual drifts that is predictive (with a certain variance) of the fusion performance of 2 to 4 trackers. This measure can be used to choose the best combination of trackers for a given database.

Self-diagnosis associated with recovery means should be an essential feature of safe systems. Self assessment of prediction quality is usually unstable and hard to set up – it is often pessimistic. In our work on drift detection it has been observed than better prediction was obtained using an out-of-consensus approach, implying than redundancy monitoring, which is a simpler failure prediction scheme, is safer than introspection. This assertion however requires further investigation.

Certification equipment and tools

It seems difficult or even theoretically impossible to prove the safety of APES for all their potential inputs: the sensory data are contingent and their high dimensionality makes the bad behavior of predictors difficult to anticipate. If safety cannot be proven, an alternative strategy is to address trustworthiness directly and to provide APES's with additional tools, outputs or devices that would convince users or certification authorities.

APES trustworthiness can be improved along two different strategies: by making its operations intelligible to detect possible wrong behaviors, and by improving its usability for better acceptance.

Explainability (for authorities)

Explainability is the ability of a system to justify the cause or origin of its prediction by providing a dedicated representation: a text or a visual sign. Chapter 4 has presented what kind of explanations are possible, especially when dealing with deep networks. It has been pointed out that the usefulness of an explanation depends on its recipient: an end-user, an engineer, a scientist or an authority. Explainability may play several roles to improve system safety.

The fact that a system is able to deliver reliable explanations is an element that may be used to improve its trustworthiness. The values of explanations can be checked to verify on specific cases that everything goes right. Explainability allows better and more efficient monitoring.

Another use of explanations for authorities is as a evidence of good or bad activity. They can be recorded for further analysis in case of failure, have usually smaller size than the whole system inner states, and encode directly informative features.

Using explanations to improve safety implies that their production is reliable. As already discussed in Chapter 4 page 130, the assessment of explanation quality is not a very well settled question.

Usability

One way of looking at the question of system safety is to consider that the best judge is the user: a system will be recognized as safe if the user declared to be so. Another way to justify the concern about usability for safety is the importance of human-computer interfaces. This places *usability* as a critical feature of safety assessment when a user is involved as a recipient or as a prescriber of the predictive process and enters into an interactive loop with the system.

The computer vision literature addresses human-in-the-loop processes to generate data interpretation such as segmentation (McGuinness and O'connor, 2010; Zhao and Xie, 2013) or annotations (bounding boxes around objects, categories, attributes), often leveraged by a crowdsourcing approach (Russakovsky et al., 2015; Kovashka et al., 2016). The interactive processes are usually evaluated as a trade-off between accuracy and interaction load (Veit et al., 2015), typically measured as mouse click counts, number of workers involved in a crowdsourcing setting (Branson et al., 2017) or time to accomplish the task.

Human computer Interaction is a now well established inter disciplinary research field, at the intersection of computer science and behavioral sciences (psychology, neuroscience, ergonomics), and should take an active part in analyzing and evaluating usability of interactive perceptual processes.

In his book (Norman, 2009), Don Norman one of the foremost researchers in this field analyzes the way user (should) interact with machines, and suggests six general rules that should govern system design:

1. Provide rich, complex, and natural signals.
2. Be predictable.
3. Provide a good conceptual model.
4. Make the output understandable.
5. Provide continual awareness, without annoyance.
6. Exploit natural mappings to make interaction understandable and effective.

These rules are general and the question remains of how instantiating and evaluating them for APES so as to improve usability and safety. This question has been partly addressed in the following project.

SATIE project (2014-2015)

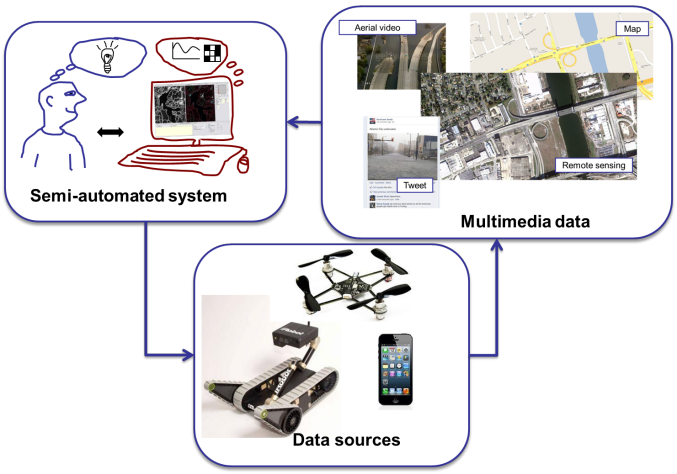


Fig. 5.12: SATIE project.

SATIE ("Semi-AuTomated Information Extraction") was an ONERA research project aiming at studying human interaction design and evaluation of interactive system exploiting multimedia data and sources (Fig. 5.12). It originates in requirements or critics about available systems expressed by experts on image analysis for defense or intelligence applications.

The project identified several topics that appeared relevant and were found not being sufficiently addressed:

Usability	The system must be seen as an aid, not a constraint.
Acceptance	The system is a partner, not a substitute to human skills.
Familiarization	System design must take into account the maturation of human-computer relation.
Personalization	System interactions should allow and exploit user idiosyncrasies (personal history, preferences, etc.).

When addressing more specifically the usability of prediction processes, it had been found at that time that several important dimensions were not sufficiently mature, or even addressed:

Affordance	Processes can be controlled at several levels of the chain (input parameters, decision thresholds, on-line annotations, etc.) but there is no global and validated account of action impact.
Intelligibility	State of the art processes are rather opaque, and the way they build their predictions, especially when they are wrong, is often unclear. We have seen in chapter 4 that this question is now an important research subject.
Predictability	The impact of user actions on the system is often difficult to foresee (parameters have various sensibilities, several combinations are more useful than others). Knowing the potential role of actions is however crucial for efficient feedback.

The sequel of this project would have been to instantiate and evaluate those interactive system features using psycho-physics measuring devices (eye tracking, electroencephalography) in an experimental psychology study. This is work that is still to be done.

5.3 Discussion

The safety of AI algorithms and software is a recent issue and motivated by their economical prospects. Making AI based products safe is becoming a key concern for stakeholders [21], citizens, government or market. This section discusses several of the currently envisioned solutions.

[21] See the section "Risks for safety and the effective functioning of the liability regime", pg. 12 of the recent White Paper on AI from the European Commission.

Use of formal methods

One now well studied research domain is the application of formal verification methods to deep neural networks (see the presentation of several works pg. 155 and pg. 162). Given the ubiquity of such objects in contemporary perceptual models and systems, this seems an interesting objective.

As already questioned, formal methods developed for software validation & verification may not be well suited to the nature of perceptual systems.

A first difficulty is the way safety is expressed as a satisfiability problem ($\mathbf{x} \in \mathcal{X} \Rightarrow f(\mathbf{x}) \in \mathcal{Y}$): how can this kind of formulation be adapted to data-driven function specification as usually done for perception? Most of the proposed approaches ensure stability around known operating points, but this may drastically restrict the system operating domain to known but limited situations, without being able to assess the safety of other ones, unknown but likely to occur. Ensuring joint local constraint satisfaction and generalization capacity is still to be achieved.

Another difficulty for formal verification of modern perceptual systems is that adversarial examples seem to be rather easy to find due to the high dimension of deep network parameters (computer vision networks have several millions of weights and biases). This means that formal methods alone will probably not be enough to prove safety in a large operating domain even though problematic cases such as those generated by adversarial attacks may not happen during the system life-time. If the existence of adversarial examples is a conceptual obstacle to universal formal stability of predictive processes, it may not be that problematic for real systems. The idea of confronting defenses and attacks could be an interesting research direction to improve their accuracy and robustness as a design principle.

Empirical testing

Empirical validation methods, such as those practiced in most of computer vision challenges, is another way to assess safety, and partly substitutes empiricism to formal proof.

The emphasis is then on designing the good testing dataset that will cover the various potential failure causes, and the evaluation protocols. The CV-HAZOP approach (Zendel et al., 2017b), which tries to optimize hazard covering as discussed previously, is one step in that direction.

A complementary action could be to develop testing features taking into account the actual instantiated algorithm, for instance by forging black-box attacks, whether adversarial or not: i.e. to design algorithm dependent “stress tests”.

User centered validation

Safety issues should involve the recipient of APES's production. A first reason, as already mentioned, is that users are the best judge of how useful is an APES to their needs – it will be for sure useless if untrusted or unsafe. Somehow, safety can be considered as a feature of usability. This way of thinking has however several limitations: safety constraints sometimes appears contradictory to ease of use, for instance when cybersecurity requires complex passwords or even more complex protocols to log into a computer system.

Another reason for placing user at the center, whether it is a human or another artificial system, is to resist to over specification that may impose too large security margins at the cost of lower efficiency. It is also a way to bypass the question of formally defining a target operating domain, which is a difficult task, and almost impossible when dealing with high dimensional data. Subsuming safety under usability is a way to circumvent the problem of forecasting its operating domain independently of an actual usage and leads to *virtual* verification, proof or testing.

Systemic safety

A current trend of modern industrial design is to assemble COTS (Commercial off-the-shelf) components, relying on data sheets characterizing their performance, and ideally their operating domain. APES are systems that may be considered as COTS of a bigger system, and may also itself contain COTS (think of deep image features such as currently used in computer vision).

The problem with the classical practice of decomposing a system in several components to assess global system safety as a combination of unitary verification and validation procedures is difficult to apply to APES.

- Ideally, we would like to know how, when and on what applying a given algorithm to provide a satisfactory solution. APES are complex, and the way their performance is described (see benchmark section pg. 152) is usually not sufficient or correctly focused to predict how errors may propagate to the bigger system and cause failure. Fundamentally, there is no clear way to identify or describe an operating domain, especially when machine learning designed components are involved.

- ML based components are likely to be updated regularly in a system, without modifying the hardware or software structure, simply by modifying their parameters (think of a whole deep network as a parameter described by representations such as Open Neural Network eXchange files)[22]. A new version will generally improve the component performances *on average*, but will find

[22] https://onnx.ai/

difficult to guarantee non regression on several data. The impact of this updating on the global system performance and safety becomes difficult to predict.

- Assessing the safety of APES as a system assembling several components also raises questions. The main obstacle is the way modern perceptual systems are created using Deep Learning techniques (end-to-end design, fine-tuning, multi-task, memory networks) and global optimization that make almost impossible to assign a functional role to the various parts of the networks.

5.4 Research directions

It is too soon to state what should be the good ways to assess safety of complex high-dimensional machine learning based processes such as found in APES, in particular, it appears unclear when and where to use formal verification methods, if they can be completed or even replaced by more empirical strategies. Four more focused research directions to mature this problem are proposed in the following:

Joint design of run-time failure prognosis and recovery

Equipping a system with a run-time failure prognosticator is a possible step towards improved safety, but not a final objective. The question remains of what should be done once a failure prognostic has been detected: should the system simply send a warning, or start more complex recovery actions to escape, counter or protect itself from the failure cause?

In either case, a prognosticator should take into account the failure origin: an accidental out-of-distribution data cannot be considered the same way as a malicious adversarial attack, may have a different impact in case of not failure detection and should be countered accordingly. Recovery means can also be harmful or have a too negative impact on usability, potentially leading to hazardous behavior of the global system.

A key research direction is therefore to design a functional architecture able to handle various types of failures with corresponding recovery or defense strategies.

Explainability for user centered validation

Pure formal verification of APES is difficult, especially when they have been designed by machine learning techniques. One proposed alternative of a third party evaluation is to involve explicitly the user/client for contextual validation as the better judge of the satisfaction of his/her requirements.

In chapter 4, it has been suggested to measure explainability quality as failure prediction: it is proposed to extend this idea and make explainability a tool for assessing the safety of the perceptual system by giving an overview of its behavior to the user.

A first objective is to figure out what kind of explanation or justification is useful for this purpose and to examine what category of APES is likely to be validated by usage, i.e. what type of prediction is expected, from what data, with what level of interaction. Also, if explainablity is to be considered as a sign production, it will be interesting to examine on what kind of media they will be conveyed (visual, sound, haptic, etc.), in what way they will refer to the system behavior features (as index, icon or symbol), and how they can be reliably produced.

A second series of problems is to differentiate the role of explainability at three different stages of a system life-cycle: as a tool to help design, at run-time to check system behavior and for a post-hoc analysis, targeting three different types of recipients: engineer, end-user and authority.

Multi-task system safety

Designing multi-task systems that partly share common resources is a way to regularize inner representations and features through machine learning (Caruana, 1997; Evgeniou and Pontil, 2004; Ruder, 2017) and is expected to make the system globally more robust and accurate. When addressing safety issues, multi-task approaches can be regarded as a difficulty or as an opportunity.

As a difficulty, a problem is to define a generic multiple objective evaluation framework for APES. What trade-offs and/or correlations are active and acceptable? Do errors propagate between tasks? What are the components involved, and what is their main impact to performance? A clear view on those questions is necessary to understand and describe a multi-task system behavior, and is therefore critical for safety assessment.

However, the fact that the system is designed to jointly perform several tasks can be beneficial to failure detection, as each task to be completed provides a different view of the situation, and may be used to transfer behavioral information from one task to another, for instance in case or corrupted input data.

Certification of APES

An important dimension of system safety is the design of norms and protocols that would lead to their *certification* by authorities allowing their actual usage.

Modern perceptual systems such as APES make a central use of machine learning, which exploit several data sources and annotations with various controls and heuristics. These techniques (and sometimes tricks) make perceptual prediction performance apparently improve, but not with enough convincing justification.

There are two issues related to certification: qualification of the predictive function in its operating domain, but also of the way it has been settled – the machine learning phase.

One feature of the first issue is the design of testing datasets which should 1/ reliably cover the expected operating domain and 2/ reveal algorithm resilience to known hazards. The design of adversarial attacks as algorithm dependent "stress tests" is probably an activity that should be emphasized given the current instability of deep networks.

Regarding the second issue – the impact of the machine learning phase – one of the well known problem is the reproducibility of results: many works do not show in their experiments a correct account of the bias-variance dilemma, making the claimed performances and the generalization capacity questionable. Evaluation and design protocols should give means to assess faithfulness of the results, of the ways operating points are controlled, of the robustness to annotation noise or data poisoning, for instance.

Synthetic data simulation or generation could be useful to help addressing those two issues. Simulation techniques have now reached a state where photo realistic data is available at low calculation cost, for instance (see the last Turing architecture from Nvidia which integrates hard wired ray tracing), although designing models remains an expensive task. Generative models are also able to produce data with nice variety and look, and have been used to complement data distribution. Their possible contribution to safety assessment is still to be justified.

Conclusion 6

Summary

Features of Autonomous PErceptual Systems

The purpose of this document was to discuss the consequences of applying an idea of autonomy to perception with the underlying hypothesis that developing such a property is the key to its reliability, and therefore to its potential usability for demanding applications. The schedule of this practical objective is justified by the maturity level that can now be envisaged thanks to the latest development of Deep Learning.

A first direct implication of autonomy is perceptual *agency*: when speaking of autonomy, we indeed implicitly assume the idea of an agent, i.e. a dynamic system with components interacting between themselves and with an outside.

The specificity of a perceptual agent is its ability to *express* for a known recipient a measure of or a sign about the world. The value of what is expressed depends on the needs of this recipient: autonomy is granted if the perceptual system takes its responsibilities and agrees in an explicit form with the user on the quality and requirements that it should satisfy, i.e. if it is able to establish and fulfill a *contract* defining the nature of the expected perceptual *service*. A formalized interaction with the user/client is a key feature of a *trustworthy* perception.

The conceptual and functional separation between perception and cognition is unstable: there is no such process as *pure* perception that can operate without memory, learning, reasoning, knowledge, decision, etc. Conversely, perception is also a fundamental faculty of what is called cognition. Acknowledging the *cognitive* dimension of perception is required to provide it with some autonomy.

Instantiation of Autonomous PErceptual Systems

The design of APES's rests on two conceptual pillars: interactivity and learning.

Interactivity means that the system is able to act and to be receptive to some inputs. The specificity of APES's functional pattern is the double source and destination of interaction: the outer environment and a user/client.

The interaction with the environment can be apprehended under a "classical" active perception approach. Two of its essential features: attention, mostly as a soft modulating process, and dynamic sequential information integration have been

incorporated as standard functional patterns in deep learning formalism and architectures.

Formalizing the interaction between the system and the user that receives the products of perception can be one answer to a better control of the system behavior: each part can be involved in a dialog to specify in an explicit and trustful way what is expected and what is achievable, to monitor the state of achievement or to even collaborate to produce the usable perceptual outputs. The introduction of such a dialog at the perceptual level requires the development of a specific formalism able to express several aspects of system behavior and implies reliable predictive self-assessment capacities to avoid misunderstanding.

Machine learning is now a prerequisite to artificial perception design. Its supervised framework shows the best performance, but relies on the availability of large databases and restrains the dependability of predictions to the boundaries defined by the training database. Several strategies or settings have been proposed to respond to this limitation, typically variations of the supervised pattern combining multiple types of data or hybridization with knowledge based models, but often with lower performance.

The concept of autonomy usually refers to on-line adaptive operating capacities: one idea is to extend it to the development of those capacities, not only their usage, and make learning a constitutive part of APES's life, with the underlying hypothesis that the acquisition of reliable perceptual skills depends on endowing a perceptual system with some developmental autonomy and capacities such as continual or never-ending learning able to improve consistently their quality and the extension of their repertoire.

Another argument for more autonomy when learning is the complexity of cognitive skills such as perception that involve multiple interdependent high dimensional functions and components sharing global informational resources in an intertwined way: the precise specification of each component and their relations is an overcomplex task. One answer to master this complexity, and one of the reasons for its success, is end-to-end deep learning with its capacity to transfer specification complexity to a global optimization problem given a parametric architecture and learning samples.

Such a design principle allows great creativity, but has also several limitations: it suffers from a well-known black-box phenomenon making difficult the understanding and control of its behavior. This weakness leads to a new important feature requested for autonomy: intelligibility. Equipping components and systems with this property is currently a major concern of AI, with related objectives such as justification, explanation, explicit reasoning, uncertainty representation, transparency, accountability, responsibility, etc. but is still in its infancy: agreed and practical

definitions, clearly identified problems with metrics and evaluation protocols are still lacking.

With the arrival of AI components and techniques in many application and scientific domains, their safety and trustworthiness are becoming topics of increasing interest. Classical validation and verification techniques based on formal proofs are no longer relevant or applicable for systems that rely on machine learning principles. Artificial perception fall in this category. Explainability, robustness testing and prediction of bad or unpredictable behavior are current research objectives aiming at providing new safety tools. A question that requires further investigation is whether aiming at autonomy and letting systems develop new skills on their own, but in a accountable way, could be an alternative strategy to gain safety.

Perspectives

Several challenges and possible research actions that could contribute to mastering Artificial PErceptual Systems have been identified in the core of this document. This section proposes a more organized research program.

Short term actions

Incrementability of perceptual skill acquisition

The general problem is to study the capacity of a visual system to acquire incrementally new interpretation capacities without forgetting the old ones. A first research action is the PhD Thesis of Alexis Lechat (2018-2021) where the objective is to address the incremental learning of visual question answering (see box 'Incremental learning of Visual Question Answering' pg. 106).

Dialog as collaborative specification and explanation

A future PhD Thesis (2020-2023) is expected on the problem of designing an explainable by design perceptual system. The main idea is to represent the behavior of an algorithm as a natural language-record of the steps that have led to collaboratively realize a perceptual task (detection, recognition, "captioning", monitoring, tracking, etc.) by two agents having distinct roles but able to dialog which each other to exchange various types of information.

Self-diagnosis and certification of perception

The question of mastering the use and integration of machine learning enabled perceptual systems gives rise to new issues: How to ensure a given level of performance? How to avoid failure?

The proposed contributions to answer those questions will be at two levels: development of specific tools to assess operating domain (adversarial defense, out-of-distribution detection), and definition of generic methods or protocols able to validate perceptual functions. The proposed actions will be carried out as part of ONERA research projects and participation to discussion groups about standardization of AI.

Long term objectives

Several long term objectives have been described throughout the document. We recall them briefly in this section and organize them in several clusters.

Systemic and dynamic dimension of perception

As argued in chapter 3, bringing autonomy in perception breaks the static feed-forward functional pipeline classically exploited in artificial perception design (the **I** pattern) and requires a more dynamic and systemic approach (**Y** or **X** patterns).

Systemic complexity management: Addressing perception as a system containing multiple interacting dynamic components is a difficult objective, both from practical and conceptual points of view. Deep learning formalism provides one answer when deployed as an end-to-end approach, but remains opaque and rigid. Other more flexible alternatives do not compete yet in terms of performance.

Versatility of perception: Perception potentially serves multiple goals that depend on the user needs. Designing or modeling a perceptual system able to do so addresses three specific issues: optimality of resource allocation for multiple objectives or tasks, incrementability of tasks and evaluation of such versatile system.

Dynamics of versatile perceptual systems: These two essential features of an Autonomous PErceptual System (systemic and dynamic dimensions) should be addressed jointly. Three research directions are of interest: better models and control tools of task spaces, generalized attention as an online task-based dynamic selection of (computing) resources and control of contextual priors and functional dynamics.

Learning of perceptual dynamical systems: APES's are complex dynamic objects with potentially highly dimensional state spaces and two contingent worlds to deal with: the external environment and the user/client. Learning such systems is therefore difficult, and two research directions are proposed to exploit an autonomous dimension: adapt attention skills to learning itself, and develop specific tools to acquire reliable and helpful dialogue with the user/client to define perceptual objectives and requirements.

Joint dynamics of operation and development: The idea of APES makes the sequential separation of development and operation – learn and run – rigid, leading to

limited usage domain and low efficiency. Three different issues are proposed to unify those activities: interactive learning involving user/client, long-term monitored integration of experience and information ("curriculum learning") and joint opportunistic exploitation of multiple learning schemes.

Perceptual interaction

A foremost feature of an APES is its ability to interact with the user/client – human or machine – to improve the usability of perceptual production. Introducing explicitly the user/client suggests several research issues necessitating investigation.

Model and evaluation of perception as sign production: Considering perceptual outputs as signs implies that their value and relevance depends on the user/client/recipient needs. The introduction of this third part requires new ways to model and evaluate perceptual processes as *semiosis*.

Specification of perceptual service: Considering perceptual process as a service is a way to structure the various features required to ensure a successful perception, and involves three stages: a contracting phase where the perceptual process and user/client agree on how perception should occur, an actual perceptual process, and a final checkable delivery step. A complete specification should describe these three stages: the question is to define them in a usable formalism.

Clarification of the intelligibility objectives: One dimension of the interaction between the perceptual system and the user/client is to provide insight of its inner behavior in a *meaningful* way. This question of meaningfulness or intelligibility is quite naively addressed by the AI community and would benefit from conceptual tools developed in natural or human sciences to make it clearer.

User centered design: A possible exploitation of intelligibility capacities provided by the perceptual system is for design purposes. Two research objectives are possible: using design success or failure as a key indicator of intelligibility capacity and develop corresponding protocols and benchmarks; integrating intelligible features to a global design process.

Text as pivot computational representation: Natural language is a spontaneous intelligible representation that may refer to multiple contents: outer world description, inner system state, queries, etc. Thanks to deep learning and available databases, techniques to associate digital and language representations have been proposed to solve problems such as captioning, visual question answering, dialog: one proposed research direction is to extend those techniques to rule the interaction between the system and the user/client in a semantic way with a shared vocabulary.

"Grown-up" perception

The maturation of perceptual systems requires new tools and research objectives to make them reliably usable.

Self-assessment: Knowing what they can achieve is a required faculty for perceptual systems to become truly autonomous and responsible. Self-assessment tools must be developed according to two objectives: qualification of their capacity to faithfully report outer-world content and evaluation of their ability to satisfy the user/client needs. The complexity of deep learning and limited understanding of its behavior makes these two objectives difficult to achieve in current state of the art.

Multi-task system safety: Combining and sharing resources to accomplish multiple tasks is a current strategy to improve performance when using machine learning techniques. Multiplying task objectives can be both detrimental and beneficial regarding the trustworthiness of the perceptual process: the contribution of each system component to the overall output is difficult to identify, but the monitoring of each task may lead to more efficient abnormal behavior detection.

Joint design of run-time failure prognosis and recovery: The study of perceptual system safety often separates robustness control to abnormal behavior detection issues. A useful research direction is to jointly handle both sides, and design global strategies able to efficiently recover from the various sources of failure.

Explainability for user centered validation: Explainability is a desired property of a perceptual system potentially capable of helping to assess the safety of their behavior. Two directions of investigations are required to mature this idea: precising the nature of explanations useful for this purpose, and identifying the phase of the system life-cycle (design, run-time or posterior analysis) they refer to.

Certification of APES: The maturity of a technology is revealed by certification concerns, i.e. by the development and usage of methods, tools and protocols able to convince authorities that a proposed system is safe and satisfies an intended behavior. Such a global methodology is still waited for APES, especially because of their dependence on machine learning techniques. Several related problems have been addressed (corner cases or hazards identification, operating domain coverage using for instance data synthesis, machine learning good practice, adversarial defense validation etc.) but require improvement to be transformed in a general approach.

Aligning natural and cognitive science with engineering

A last series of rather speculative research questions address the relation between artificially engineered and natural systems. Regarding perception, the question can be divided in two strategies: reverse engineering of natural and human intelligent skills (as proposed by (Lake et al., 2017)) or development of a general "strong AI",

i.e. an artificial system hosting intelligent skills or behavior comparable to humans, as a proxy to artificial perception.

Representation elimination: There is still room to discuss the relation between perception, cognition and even consciousness. One question is the status and usefulness of perceptual representations, mental or formal, as an essential feature of models. This issue has been debated in natural science and philosophy, but building, exploiting and validating an engineered version of a "perception without representation" remains difficult.

Biologically plausible *and* efficient models: Natural science models can inspire artificial perception approaches. Two possible directions are worth studying: *attention* principle, either as selection or modulation, which has rather recently given rise to a large body of studies in AI, although not really aligned with natural science, and *predictive coding* that makes perceptual prediction a fundamental principle of neural organization and processing but has difficulty to scale and to compete with the problems deep learning currently addresses.

Bibliography

Abdul, Ashraf, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli (2018). "Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, p. 582 (cit. on pp. 119, 135).

Acebrón, Juan A, Luis L Bonilla, Conrad J Pérez Vicente, Félix Ritort, and Renato Spigler (2005). "The Kuramoto model: A simple paradigm for synchronization phenomena". In: *Reviews of modern physics* 77.1, p. 137 (cit. on p. 55).

Adler, Philip, Casey Falk, Sorelle A Friedler, et al. (2018). "Auditing black-box models for indirect influence". In: *Knowledge and Information Systems* 54.1, pp. 95–122 (cit. on p. 124).

Afifi, Mahmoud and Michael S Brown (2019). "What Else Can Fool Deep Learning? Addressing Color Constancy Errors on Deep Neural Network Performance". In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 243–252 (cit. on p. 160).

Akata, Zeynep, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid (2013). "Label-Embedding for Attribute-Based Classification." In: *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)* (cit. on p. 93).

Akhtar, Naveed and Ajmal Mian (2018). "Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey". In: *IEEE Access* 6, pp. 14410–14430 (cit. on p. 156).

Akhtar, Naveed, Jian Liu, and Ajmal Mian (2018). "Defense against universal adversarial perturbations". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3389–3398 (cit. on p. 167).

Akoglu, Leman, Hanghang Tong, and Danai Koutra (2015). "Graph based anomaly detection and description: a survey". In: *Data mining and knowledge discovery* 29.3, pp. 626–688 (cit. on p. 164).

Albalate, Amparo and Wolfgang Minker (2013). *Semi-Supervised and Unsupervised Machine Learning: Novel Strategies*. John Wiley & Sons (cit. on p. 90).

Aloimonos, John, Isaac Weiss, and Amit Bandyopadhyay (1988). "Active vision". In: *International journal of computer vision* 1.4, pp. 333–356 (cit. on p. 82).

Alom, Md Zahangir, Tarek M Taha, Christopher Yakopcic, et al. (2018). "The history began from alexnet: A comprehensive survey on deep learning approaches". In: *arXiv preprint arXiv:1803.01164* (cit. on pp. 1, 88).

Alur, Rajeev (2015). *Principles of cyber-physical systems*. MIT Press (cit. on p. 84).

Amjad, Rana Ali and Bernhard Claus Geiger (2019). "Learning representations for neural network-based classification using the information bottleneck principle". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (cit. on p. 121).

Ammirato, Phil, Patrick Poirson, Eunbyung Park, Jana Košecká, and Alexander C Berg (2017). "A dataset for developing and benchmarking active vision". In: *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, pp. 1378–1385 (cit. on p. 85).

Ammirato, Phil, Alexander C Berg, and Jana Kosecka (2018). "Active Vision Dataset Benchmark". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2046–2049 (cit. on p. 85).

Amodei, Dario, Chris Olah, Jacob Steinhardt, et al. (2016). “Concrete problems in AI safety”. In: *arXiv preprint arXiv:1606.06565* (cit. on p. 151).

Ancona, Marco, Enea Ceolini, Cengiz Öztireli, and Markus Gross (2018). “Towards better understanding of gradient-based attribution methods for Deep Neural Networks”. In: *International Conference on Learning Representations* (cit. on pp. 125, 131).

Anderson, Peter, Basura Fernando, Mark Johnson, and Stephen Gould (2016). “Spice: Semantic propositional image caption evaluation”. In: *European Conference on Computer Vision*. Springer, pp. 382–398 (cit. on p. 126).

Anderson, Peter, Xiaodong He, Chris Buehler, et al. (2018a). “Bottom-up and top-down attention for image captioning and visual question answering”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6077–6086 (cit. on p. 80).

Anderson, Peter, Angel Chang, Devendra Singh Chaplot, et al. (2018b). “On evaluation of embodied navigation agents”. In: *arXiv preprint arXiv:1807.06757* (cit. on p. 85).

Anderson, Peter, Qi Wu, Damien Teney, et al. (2018c). “Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3674–3683 (cit. on p. 85).

Andreas, Jacob, Marcus Rohrbach, Trevor Darrell, and Dan Klein (2016). “Neural module networks”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 39–48 (cit. on p. 52).

Andreopoulos, A. (2009). “Active object recognition in theory and practice”. PhD thesis. Toronto, Ontario: York University (cit. on p. 77).

Andreopoulos, A. and J.K. Tsotsos (2008). “Active Vision for Door Localization and Door Opening using Playbot: A Computer Controlled Wheelchair for People with Mobility Impairments”. In: *Computer and Robot Vision, 2008. CRV ’08. Canadian Conference on*, pp. 3–10 (cit. on p. 77).

– (2009). “A Theory of Active Object Localization”. In: *IEEE International Conference on Computer Vision* (cit. on p. 77).

Andreopoulos, Alexander and John K Tsotsos (2013). “A computational learning theory of active object recognition under uncertainty”. In: *International journal of computer vision* 101.1, pp. 95–142 (cit. on p. 77).

Andreopoulos, Alexander and JohnK. Tsotsos (2012). “A Computational Learning Theory of Active Object Recognition Under Uncertainty”. English. In: *International Journal of Computer Vision*, pp. 1–48 (cit. on p. 77).

Andriluka, Mykhaylo, Jasper RR Uijlings, and Vittorio Ferrari (2018). “Fluid annotation: a human-machine collaboration interface for full image annotation”. In: *2018 ACM Multimedia Conference on Multimedia Conference*. ACM, pp. 1957–1966 (cit. on p. 103).

Angelino, Elaine, Nicholas Larus-Stone, Daniel Alabi, Margo Seltzer, and Cynthia Rudin (2017). “Learning certifiably optimal rule lists”. In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, pp. 35–44 (cit. on p. 130).

Antol, Stanislaw, Aishwarya Agrawal, Jiasen Lu, et al. (2015). “Vqa: Visual question answering”. In: *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433 (cit. on p. 86).

Antoniou, Antreas, Harrison Edwards, and Amos Storkey (2019). “How to train your MAML”. In: *International Conference on Learning Representations* (cit. on p. 92).

Arbel, Tal and Frank P. Ferrie (1996). “Informative Views and Sequential Recognition”. In: *ECCV ’96: Proceedings of the 4th European Conference on Computer Vision-Volume I*. London, UK: Springer-Verlag, pp. 469–481 (cit. on p. 76).

– (2001). “Entropy-based gaze planning”. In: *Image Vision Comput.* 19.11, pp. 779–786 (cit. on p. 68).

Arjovsky, Martin, Soumith Chintala, and Léon Bottou (2017). “Wasserstein gan”. In: *International Conference on Machine Learning* (cit. on p. 95).

Arnab, Anurag, Ondrej Miksik, and Philip HS Torr (2018). “On the robustness of semantic segmentation models to adversarial attacks”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 888–897 (cit. on p. 156).

Arora, Sanjeev, Rong Ge, Behnam Neyshabur, and Yi Zhang (2018). “Stronger Generalization Bounds for Deep Nets via a Compression Approach”. In: *International Conference on Machine Learning*, pp. 254–263 (cit. on p. 108).

Arora, Sanjeev, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang (2019). “Fine-Grained Analysis of Optimization and Generalization for Overparameterized Two-Layer Neural Networks”. In: *International Conference on Machine Learning*, pp. 322–332 (cit. on p. 108).

Arras, Leila, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek (2017). “Explaining Recurrent Neural Network Predictions in Sentiment Analysis”. In: *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp. 159–168 (cit. on p. 126).

Arridge, Simon, Peter Maass, Ozan Öktem, and Carola-Bibiane Schönlieb (2019). “Solving inverse problems using data-driven models”. In: *Acta Numerica* 28, pp. 1–174 (cit. on p. 101).

Arrieta, Alejandro Barredo, Natalia Diaz-Rodriguez, Javier Del Ser, et al. (2019). “Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI”. In: *arXiv preprint arXiv:1910.1004* (cit. on pp. 118, 130).

Arth, C., C. Leistner, and H. Bischof (2007). “Object Reacquisition and Tracking in Large-Scale Smart Camera Networks”. In: *Distributed Smart Cameras, 2007. ICDSC ’07. First ACM/IEEE International Conference on*, pp. 156–163 (cit. on p. 79).

Ashmore, Rob, Radu Calinescu, and Colin Paterson (2019). “Assuring the Machine Learning Lifecycle: Desiderata, Methods, and Challenges”. In: *arXiv preprint arXiv:1905.04223* (cit. on p. 151).

Atanasov, Nikolay, Bharath Sankaran, Jerome Le Ny, et al. (2013). “Hypothesis testing framework for active object detection”. In: *2013 IEEE International Conference on Robotics and Automation*. IEEE, pp. 4216–4222 (cit. on p. 77).

Atanasov, Nikolay, Bharath Sankaran, Jerome Le Ny, George J Pappas, and Kostas Daniilidis (2014). “Nonmyopic view planning for active object classification and pose estimation”. In: *IEEE Transactions on Robotics* 30.5, pp. 1078–1090 (cit. on pp. 68, 77).

Atapour-Abarghouei, Amir and Toby P Breckon (2018). “Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Vol. 18, p. 1 (cit. on p. 153).

Attias, Idan, Aryeh Kontorovich, and Yishay Mansour (2019). “Improved Generalization Bounds for Robust Learning”. In: *Proceedings of the 30th International Conference on Algorithmic Learning Theory*. Ed. by Aurélien Garivier and Satyen Kale. Vol. 98. Proceedings of Machine Learning Research. Chicago, Illinois: PMLR, pp. 162–183 (cit. on p. 158).

Aydemir, A. and P. Jensfelt (2012). “Exploiting and modeling local 3D structure for predicting object locations”. In: *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE (cit. on p. 77).

Aydemir, Alper, Andrzej Pronobis, Moritz Göbelbecker, and Patric Jensfelt (Aug. 2013). “Active Visual Object Search in Unknown Environments Using Uncertain Semantics”. In: *IEEE Transactions on Robotics* 29.4, pp. 986–1002 (cit. on p. 77).

Azevedo, Frederico AC, Ludmila RB Carvalho, Lea T Grinberg, et al. (2009). “Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain”. In: *Journal of Comparative Neurology* 513.5, pp. 532–541 (cit. on p. 35).

Baden, Tom, Timm Schubert, Philipp Berens, and Thomas Euler (2018). “The Functional Organization of Vertebrate Retinal Circuits for Vision”. In: *Oxford Research Encyclopedia of Neuroscience* (cit. on p. 32).

Bagdanov, Andrew D., Alberto del Bimbo, and Federico Pernici (2005). “Acquisition of high-resolution images through on-line saccade sequence planning”. In: *VSSN ’05: Proceedings of the third ACM international workshop on Video surveillance & sensor networks*. Hilton, Singapore: ACM, pp. 121–130 (cit. on p. 78).

Bagdanov, Andrew D., Alberto Del Bimbo, Walter Nunziati, and Federico Pernici (2006). “Learning Foveal Sensing Strategies in Unconstrained Surveillance Environments”. In: *AVSS*, p. 40 (cit. on p. 78).

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2014). “Neural machine translation by jointly learning to align and translate”. In: *arXiv preprint arXiv:1409.0473* (cit. on p. 80).

Bai, Yancheng, Yongqiang Zhang, Mingli Ding, and Bernard Ghanem (2018). “Sod-mtgan: Small object detection via multi-task generative adversarial network”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 206–221 (cit. on p. 94).

Bajcsy, Ruzena, Yiannis Aloimonos, and John K Tsotsos (2018). “Revisiting active perception”. In: *Autonomous Robots* 42.2, pp. 177–196 (cit. on pp. 61, 62, 81, 84, 85).

Bakhtari, Ardevan, Matthew Mackay, and Beno Benhabib (2009). “Active-vision for the autonomous surveillance of dynamic, multi-object environments”. In: *Journal of Intelligent and Robotic Systems* 54.4, p. 567 (cit. on p. 79).

Baldi, Pierre and Kurt Hornik (1989). “Neural networks and principal component analysis: Learning from examples without local minima”. In: *Neural networks* 2.1, pp. 53–58 (cit. on p. 121).

Baltrušaitis, Tadas, Chaitanya Ahuja, and Louis-Philippe Morency (2018). “Multimodal machine learning: A survey and taxonomy”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.2, pp. 423–443 (cit. on p. 136).

Balunovic, Mislav, Maximilian Baader, Gagandeep Singh, Timon Gehr, and Martin Vechev (2019). “Certifying Geometric Robustness of Neural Networks”. In: *Advances in Neural Information Processing Systems*, pp. 15287–15297 (cit. on p. 167).

Banerjee, Satanjeev and Alon Lavie (2005). “METEOR: An automatic metric for MT evaluation with improved correlation with human judgments”. In: *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72 (cit. on p. 126).

Banich, Marie T. and Rebecca J. Compton (2018). *Cognitive Neuroscience*. 4th ed. Cambridge University Press (cit. on p. 29).

Bansal, Aayush, Ali Farhadi, and Devi Parikh (2014). “Towards transparent systems: Semantic characterization of failure modes”. In: *European Conference on Computer Vision*. Springer, pp. 366–381 (cit. on p. 132).

Bansal, Gagan and Daniel S Weld (2018). “A Coverage-Based Utility Model for Identifying Unknown Unknowns”. In: *Proc. of AAAI* (cit. on p. 161).

Barandiaran, Xabier and Alvaro Moreno (2008). “Adaptivity: From metabolism to behavior”. In: *Adaptive Behavior* 16.5, pp. 325–344 (cit. on p. 22).

Barandiaran, Xabier E. (2017). “Autonomy and Enactivism: Towards a Theory of Sensorimotor Autonomous Agency”. In: *Topoi* 36.3, pp. 409–430 (cit. on p. 23).

Barrett, David GT, Ari S Morcos, and Jakob H Macke (2019). “Analyzing biological and artificial neural networks: challenges with opportunities for synergy?” In: *Current opinion in neurobiology* 55, pp. 55–64 (cit. on p. 36).

Bartlett, Peter L, Nick Harvey, Christopher Liaw, and Abbas Mehrabian (2019). “Nearly-tight VC-dimension and Pseudodimension Bounds for Piecewise Linear Neural Networks.” In: *Journal of Machine Learning Research* 20.63, pp. 1–17 (cit. on p. 108).

Bastos, Andre M, W Martin Usrey, Rick A Adams, et al. (2012). “Canonical microcircuits for predictive coding”. In: *Neuron* 76.4, pp. 695–711 (cit. on pp. 37, 40, 56).

Batra, Dhruv, Adarsh Kowdle, Devi Parikh, Jiebo Luo, and Tsuhan Chen (2011). “Interactively co-segmentating topically related images with intelligent scribble guidance”. In: *International journal of computer vision* 93.3, pp. 273–292 (cit. on p. 103).

Battaglia, Peter, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, et al. (2016). “Interaction networks for learning about objects, relations and physics”. In: *Advances in neural information processing systems*, pp. 4502–4510 (cit. on p. 102).

Bau, David, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba (2017). “Network dissection: Quantifying interpretability of deep visual representations”. In: *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. IEEE, pp. 3319–3327 (cit. on pp. 131, 135).

Bau, David, Jun-Yan Zhu, Hendrik Strobelt, et al. (2018). “Gan dissection: Visualizing and understanding generative adversarial networks”. In: *arXiv preprint arXiv:1811.10597* (cit. on p. 135).

Bau, David, Jun-Yan Zhu, Hendrik Strobelt, et al. (2019). “Visualizing and Understanding Generative Adversarial Networks”. In: *International Conference on Learning Representations* (cit. on p. 123).

Beluch, William H, Tim Genewein, Andreas Nürnberger, and Jan M Köhler (2018). “The power of ensembles for active learning in image classification”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9368–9377 (cit. on p. 104).

Ben-younes, Hedi, Remi Cadene, Matthieu Cord, and Nicolas Thome (2017). “MUTAN: Multimodal Tucker Fusion for Visual Question Answering”. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2612–2620 (cit. on p. 127).

Bengio, Yoshua, Jérôme Louradour, Ronan Collobert, and Jason Weston (2009). “Curriculum learning”. In: *Proceedings of the 26th annual international conference on machine learning*. ACM, pp. 41–48 (cit. on p. 114).

Bengio, Yoshua, Li Yao, Guillaume Alain, and Pascal Vincent (2013a). “Generalized denoising auto-encoders as generative models”. In: *Advances in Neural Information Processing Systems* (cit. on p. 95).

Bengio, Yoshua, Aaron Courville, and Pascal Vincent (2013b). “Representation learning: A review and new perspectives”. In: *IEEE transactions on pattern analysis and machine intelligence* 35.8, pp. 1798–1828 (cit. on p. 91).

Bertsekas, Dimitri (1995). *Dynamic programming and optimal control*. Vol. 1. 2. Athena scientific Belmont, MA (cit. on p. 71).

Bertsekas, Dimitri P and John N Tsitsiklis (1996). *Neuro-dynamic programming*. Vol. 5. Athena Scientific Belmont, MA (cit. on p. 71).

Best, Graeme (2019). “Planning Algorithms for Multi-Robot Active Perception”. PhD thesis. University of Sydney (cit. on pp. 79, 81).

Bhatt, Umang, Alice Xiang, Shubham Sharma, et al. (2019). “Explainable Machine Learning in Deployment”. In: *arXiv:1909.06342* (cit. on p. 118).

Bichindaritz, Isabelle and Cindy Marling (2006). “Case-based reasoning in the health sciences: What’s next?” In: *Artificial intelligence in medicine* 36.2, pp. 127–135 (cit. on p. 130).

Bilen, Hakan and Andrea Vedaldi (2016). “Weakly supervised deep detection networks”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2846–2854 (cit. on p. 90).

Bimbo, Alberto Del and Federico Pernici (2006). “Towards on-line saccade planning for high-resolution image sensing”. In: *Pattern Recognition Letters* 27.15, pp. 1826–1834 (cit. on p. 78).

Biran, Or and Courtenay Cotton (2017). “Explanation and justification in machine learning: A survey”. In: *IJCAI-17 Workshop on Explainable AI (XAI)*, p. 8 (cit. on p. 118).

Bishop, Christopher M. (1995). *Neural networks for pattern recognition*. Oxford university press (cit. on p. 121).

Bishop, Christopher M (2006). *Pattern recognition and machine learning*. springer (cit. on p. 87).

Blanchard, Gilles and Donald Geman (2005). “Hierarchical testing designs for pattern recognition”. In: *Annals of Statistics*, pp. 1155–1202 (cit. on p. 72).

Boden, Margaret A. (2014). “GOFAI”. In: *The Cambridge Handbook of Artificial Intelligence*. Ed. by Keith Frankish and William M.Editors Ramsey. Cambridge University Press, 89–107 (cit. on p. 120).

Bookman, Lawrence A and Ron Sun (1994). *Computational Architectures Integrating Neural and Symbolic Processes: A Perspective on the State of the Art*. Kluwer Academic Publishers (cit. on p. 96).

Borji, Ali (2019a). “Pros and cons of GAN evaluation measures”. In: *Computer Vision and Image Understanding* 179, pp. 41–65 (cit. on p. 95).

– (2019b). “Saliency Prediction in the Deep Learning Era: Successes and Limitations”. In: *IEEE transactions on pattern analysis and machine intelligence* (cit. on p. 99).

Borji, Ali and Laurent Itti (2013). “State-of-the-art in visual attention modeling”. In: *IEEE transactions on pattern analysis and machine intelligence* 35.1, pp. 185–207 (cit. on pp. 19, 80).
– (2014). “Defending Yarbus: Eye movements reveal observers’ task”. In: *Journal of vision* 14.3, pp. 29–29 (cit. on pp. 42, 80).
Borji, Ali, Ming-Ming Cheng, Qibin Hou, Huaizu Jiang, and Jia Li (2014a). “Salient object detection: A survey”. In: *arXiv preprint arXiv:1411.5878* (cit. on p. 43).
Borji, Ali, Dicky N Sihite, and Laurent Itti (2014b). “What/where to look next? Modeling top-down visual attention in complex interactive environments”. In: *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 44.5, pp. 523–538 (cit. on p. 43).
Borji, Ali, Ming-Ming Cheng, Huaizu Jiang, and Jia Li (2015). “Salient object detection: A benchmark”. In: *IEEE transactions on image processing* 24.12, pp. 5706–5722 (cit. on pp. 43, 164).
Borotschnig, Hermann, Lucas Paletta, Manfred Prantl, and Axel Pinz (2000). “Appearance-based active object recognition”. In: *Image and Vision Computing* 18.9, pp. 715–727 (cit. on pp. 68, 76).
Boult, TE, S Cruz, AR Dhamija, et al. (2019). “Learning and the unknown: Surveying steps toward open world recognition”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33, pp. 9801–9807 (cit. on p. 161).
Branson, Steve, Grant Van Horn, and Pietro Perona (2017). “Lean crowdsourcing: Combining humans and machines in an online system”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7474–7483 (cit. on p. 171).
Brentano, Franz (2014). *Psychology from an empirical standpoint*. Routledge (cit. on p. 47).
Brette, Romain (2012). “Computing with neural synchrony”. In: *PLoS computational biology* 8.6, e1002561 (cit. on p. 36).
Brodeur, Simon, Ethan Perez, Ankesh Anand, et al. (2018). *HoME: a Household Multimodal Environment* (cit. on p. 85).
Brogaard, Berit (2014). *Does perception have content?* Oxford University Press (cit. on p. 45).
Brooks, Rodney A (1991). “Intelligence without representation”. In: *Artificial intelligence* 47.1-3, pp. 139–159 (cit. on p. 55).
Brown, Tom B., Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer (2017). “Adversarial Patch”. In: *arXiv preprint arXiv:1712.09665,* (cit. on p. 159).
Brust, Clemens-Alexander, Christoph Käding, and Joachim Denzler (2018). “Active learning for deep object detection”. In: *arXiv preprint arXiv:1809.09875* (cit. on p. 103).
Bu, Lucian, Robert Babu, Bart De Schutter, et al. (2008). “A comprehensive survey of multiagent reinforcement learning”. In: *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 38.2, pp. 156–172 (cit. on p. 112).
Bucher, Maxime, Stéphane Herbin, and Frédéric Jurie (2016a). “Hard Negative Mining for Metric Learning Based Zero-Shot Classification”. In: *Computer Vision - ECCV 2016 Workshops - Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III*, pp. 524–531 (cit. on pp. 93, 111, 147, 233).
– (2016b). “Improving Semantic Embedding Consistency by Metric Learning for Zero-Shot Classiffication”. In: *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V*, pp. 730–746 (cit. on pp. 93, 233).
Bucher, Maxime, Stéphane Herbin, and Frédéric Jurie (2017). “Generating Visual Representations for Zero-Shot Classification”. In: *2017 IEEE International Conference on Computer Vision Workshops, ICCV Workshops 2017, Venice, Italy, October 22-29, 2017*, pp. 2666–2673 (cit. on pp. 94, 95, 233).
Bucher, Maxime, Stéphane Herbin, and Frédéric Jurie (2018). “Semantic bottleneck for computer vision tasks”. In: *ACCV 2018* (cit. on pp. 14, 92, 129, 133, 137, 233).
Buck, Christian, Jannis Bulian, Massimiliano Ciaramita, et al. (2018). “Ask the Right Questions: Active Question Reformulation with Reinforcement Learning”. In: *International Conference on Learning Representations* (cit. on p. 86).
Bullier, Jean (2001). “Integrated model of visual processing”. In: *Brain research reviews* 36.2-3, pp. 96–107 (cit. on p. 55).

Bullmore, Ed and Olaf Sporns (2009). “Complex brain networks: graph theoretical analysis of structural and functional systems”. In: *Nature reviews neuroscience* 10.3, p. 186 (cit. on p. 35).

Burns, J Brian, Richard S Weiss, and Edward M Riseman (1992). “The non-existence of general-case view-invariants”. In: *Geometric invariance in computer vision* 1, pp. 554–559 (cit. on p. 41).

Buschman, Timothy J and Sabine Kastner (2015). “From behavior to neural dynamics: an integrated theory of attention”. In: *Neuron* 88.1, pp. 127–144 (cit. on p. 42).

Butko, Nicholas J and Javier R Movellan (2010). “Infomax control of eye movements”. In: *IEEE Transactions on Autonomous Mental Development* 2.2, pp. 91–107 (cit. on p. 68).

Bylinskii, Zoya, Adrià Recasens, Ali Borji, et al. (2016). “Where should saliency models look next?” In: *European Conference on Computer Vision*. Springer, pp. 809–824 (cit. on p. 43).

Callari, Franco and Frank P. Ferrie (2001). “Active Object Recognition: Looking for Differences”. In: *International Journal of Computer Vision* 43.3, pp. 189–204 (cit. on p. 68).

Campos, Guilherme O, Arthur Zimek, Jörg Sander, et al. (2016). “On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study”. In: *Data Mining and Knowledge Discovery* 30.4, pp. 891–927 (cit. on p. 166).

Canevet, Olivier and Francois Fleuret (2015). “Efficient Sample Mining for Object Detection”. In: *Asian Conference on Machine Learning*, pp. 48–63 (cit. on p. 147).

Carlini, Nicholas and David Wagner (2017a). “Adversarial examples are not easily detected: Bypassing ten detection methods”. In: *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*. ACM, pp. 3–14 (cit. on pp. 156, 157).

– (2017b). “Towards Evaluating the Robustness of Neural Networks”. In: *2017 38th IEEE Symposium on Security and Privacy (SP)*. IEEE, pp. 39–57 (cit. on p. 157).

Carrasco, Marisa (2011). “Visual attention: The past 25 years”. In: *Vision research* 51.13, pp. 1484–1525 (cit. on pp. 19, 42).

Caruana, Rich (1997). “Multitask learning”. In: *Machine learning* 28.1, pp. 41–75 (cit. on pp. 92, 177).

Carvalho, Diogo V, Eduardo M Pereira, and Jaime S Cardoso (2019). “Machine Learning Interpretability: A Survey on Methods and Metrics”. In: *Electronics* 8.8, p. 832 (cit. on p. 130).

Casagrande, Vivien A and Jon H Kaas (1994). “The afferent, intrinsic, and efferent connections of primary visual cortex in primates”. In: *Primary visual cortex in primates*. Springer, pp. 201–259 (cit. on p. 16).

Castro, Francisco M, Manuel J Marın-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari (2018). “End-to-end incremental learning”. In: *ECCV 2018-European Conference on Computer Vision* (cit. on p. 106).

Čehovin, Luka, Aleš Leonardis, and Matej Kristan (2016). “Visual object tracking performance measures revisited”. In: *IEEE Transactions on Image Processing* 25.3, pp. 1261–1274 (cit. on p. 149).

Chalapathy, Raghavendra, Aditya Krishna Menon, and Sanjay Chawla (2017). “Robust, deep and inductive anomaly detection”. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, pp. 36–51 (cit. on p. 165).

Chalmond, B., B. Francesconi, and S. Herbin (Sept. 2006). “Using hidden scale for salient object detection”. In: *IEEE Transactions on Image Processing* 15.9, pp. 2644–2656 (cit. on p. 99).

Chandler, Daniel (2007). *Semiotics: the basics*. Routledge (cit. on pp. 10, 134).

Chandola, Varun, Arindam Banerjee, and Vipin Kumar (2009). “Anomaly detection: A survey”. In: *ACM computing surveys (CSUR)* 41.3, p. 15 (cit. on p. 164).

Chang, Michael B, Tomer Ullman, Antonio Torralba, and Joshua B Tenenbaum (2016). “A compositional object-based approach to learning physical dynamics”. In: *arXiv preprint arXiv:1612.00341* (cit. on p. 102).

Chao, Wei-Lun, Soravit Changpinyo, Boqing Gong, and Fei Sha (2016). “An empirical study and analysis of generalized zero-shot learning for object recognition in the wild”. In: *European Conference on Computer Vision*. Springer, pp. 52–68 (cit. on p. 92).

Chapelle, O., B. Scholkopf, and A. Zien, eds. (2006). *Semi-Supervised Learning*. MIT Press (cit. on p. 90).

Chaudhari, Sneha, Gungor Polatkan, Rohan Ramanath, and Varun Mithal (2019). “An attentive survey of attention models”. In: *arXiv preprint arXiv:1904.02874* (cit. on p. 81).

Chaudhry, Arslan, Puneet K Dokania, Thalaiyasingam Ajanthan, and Philip HS Torr (2018). “Riemannian walk for incremental learning: Understanding forgetting and intransigence”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 532–547 (cit. on p. 106).

Chaumette, F. and S. Hutchinson (Dec. 2006). “Visual Servo Control, Part I: Basic Approaches”. In: *IEEE Robotics and Automation Magazine* 13.4, pp. 82–90 (cit. on p. 78).

– (Mar. 2007). “Visual Servo Control, Part II: Advanced Approaches”. In: *IEEE Robotics and Automation Magazine* 14.1, pp. 109–118 (cit. on p. 78).

Chen, Hongshen, Xiaorui Liu, Dawei Yin, and Jiliang Tang (2017). “A survey on dialogue systems: Recent advances and new frontiers”. In: *Acm Sigkdd Explorations Newsletter* 19.2, pp. 25–35 (cit. on p. 86).

Chen, Jianbo, Le Song, Martin J Wainwright, and Michael I Jordan (2018). “Learning to Explain: An Information-Theoretic Perspective on Model Interpretation”. In: *arXiv preprint arXiv:1802.07814* (cit. on p. 124).

Chen, Po-Yi, Alexander H Liu, Yen-Cheng Liu, and Yu-Chiang Frank Wang (2019). “Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2624–2632 (cit. on p. 101).

Chen, Shengyong, Youfu F Li, Wanliang Wang, and Jianwei Zhang (2008). *Active sensor planning for multiview vision tasks*. Vol. 1. Springer (cit. on p. 77).

Chen, Shengyong, Youfu Li, and Ngai Ming Kwok (2011). “Active vision in robotic systems: A survey of recent developments”. In: *International Journal of Robotics Research* 30.11, pp. 1343–1377 (cit. on pp. 80, 81).

Chen, Zhiyuan and Bing Liu (2016). “Lifelong machine learning”. In: *Synthesis Lectures on Artificial Intelligence and Machine Learning* 10.3, pp. 1–145 (cit. on pp. 104, 115).

Cheng, Chih-Hong, Dhiraj Gulati, and Rongjie Yan (2019). “Architecting Dependable Learning-enabled Autonomous Systems: A Survey”. In: *arXiv preprint arXiv:1902.10590* (cit. on p. 151).

Cheng, Ming-Ming, Ziming Zhang, Wen-Yan Lin, and Philip Torr (2014). “BING: Binarized normed gradients for objectness estimation at 300fps”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3286–3293 (cit. on p. 81).

Cheng, Yu, Zhe Gan, Yitong Li, Jingjing Liu, and Jianfeng Gao (2018). “Sequential attention gan for interactive image editing via dialogue”. In: *arXiv preprint arXiv:1812.08352* (cit. on p. 86).

Choi, Jongwon, Hyung Jin Chang, Sangdoo Yun, et al. (2017). “Attentional correlation filter network for adaptive visual tracking”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4807–4816 (cit. on p. 80).

Christman, John (2018). “Autonomy in Moral and Political Philosophy”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Spring 2018. Metaphysics Research Lab, Stanford University (cit. on p. 23).

Chun, Marvin M (2011). “Visual working memory as visual attention sustained internally over time”. In: *Neuropsychologia* 49.6, pp. 1407–1409 (cit. on p. 44).

Chun, Marvin M, Julie D Golomb, and Nicholas B Turk-Browne (2011). “A taxonomy of external and internal attention”. In: *Annual review of psychology* 62, pp. 73–101 (cit. on pp. 19, 43).

Cisse, Moustapha, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier (2017). “Parseval Networks: Improving Robustness to Adversarial Examples”. In: *International Conference on Machine Learning*, pp. 854–863 (cit. on pp. 157, 158).

Clark, Andy (2012). “Dreaming the whole cat: Generative models, predictive processing, and the enactivist conception of perceptual experience”. In: *Mind* 121.483, pp. 753–771 (cit. on p. 47).

– (2013). “Whatever next? Predictive brains, situated agents, and the future of cognitive science”. In: *Behavioral and brain sciences* 36.3, pp. 181–204 (cit. on p. 38).

– (2015a). *Predicting peace: the end of the representation wars*. Open MIND. Frankfurt am Main: MIND Group (cit. on p. 47).
– (2015b). *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press (cit. on p. 38).
Clouard, Régis, Arnaud Renouf, and Marinette Revenu (2010). "An ontology-based model for representing image processing application objectives". In: *International Journal of Pattern Recognition and Artificial Intelligence* 24.08, pp. 1181–1208 (cit. on p. 109).
Cohen, Jeremy, Elan Rosenfeld, and Zico Kolter (2019). "Certified Adversarial Robustness via Randomized Smoothing". In: *International Conference on Machine Learning*, pp. 1310–1320 (cit. on p. 167).
Cohen, Marlene R and Adam Kohn (2011). "Measuring and interpreting neuronal correlations". In: *Nature neuroscience* 14.7, p. 811 (cit. on p. 36).
Cohen, Michael A, Patrick Cavanagh, Marvin M Chun, and Ken Nakayama (2012). "The attentional requirements of consciousness". In: *Trends in cognitive sciences* 16.8, pp. 411–417 (cit. on p. 43).
Cohn, David A, Zoubin Ghahramani, and Michael I Jordan (1996). "Active learning with statistical models". In: *Journal of artificial intelligence research* 4, pp. 129–145 (cit. on p. 103).
committee, ORAD (2018). *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*. Tech. rep. J3016 201806. Society of Automotive Engineers (SAE) (cit. on p. 25).
Coop, Robert, Aaron Mishtal, and Itamar Arel (2013). "Ensemble learning in fixed expansion layer networks for mitigating catastrophic forgetting". In: *IEEE transactions on neural networks and learning systems* 24.10, pp. 1623–1634 (cit. on p. 106).
Costello, Cash J., Christopher P. Diehl, Amit Banerjee, and Hesky Fisher (2004). "Scheduling an active camera to observe people". In: *VSSN '04: Proceedings of the ACM 2nd international workshop on Video surveillance & sensor networks*. New York, NY, USA: ACM, pp. 39–45 (cit. on p. 78).
Courty, Nicolas, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy (2016). "Optimal transport for domain adaptation". In: *IEEE transactions on pattern analysis and machine intelligence* 39.9, pp. 1853–1865 (cit. on p. 90).
Cover, Thomas M and Joy A Thomas (2012). *Elements of information theory*. John Wiley & Sons (cit. on p. 87).
Cox, David Daniel and Thomas Dean (2014). "Neural networks and neuroscience-inspired computer vision". In: *Current Biology* 24.18, R921–R929 (cit. on p. 56).
Crane, Tim and Craig French (2017). "The Problem of Perception". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Spring 2017. Metaphysics Research Lab, Stanford University (cit. on p. 45).
Craye, Céline (2017). "Intrinsic motivation mecanisms for incremental learning of visual saliency". PhD thesis. Paris Saclay (cit. on p. 91).
Craye, Céline, David Filliat, and Jean-François Goudou (2018). "Biovision: a biomimetics platform for intrinsically motivated visual saliency learning". In: *IEEE Transactions on Cognitive and Developmental Systems* (cit. on p. 91).
Creswell, Antonia, Tom White, Vincent Dumoulin, et al. (2018). "Generative adversarial networks: An overview". In: *IEEE Signal Processing Magazine* 35.1, pp. 53–65 (cit. on p. 94).
Croon, G.C.H.E. de, I.G. Sprinkhuizen-Kuyper, and E.O. Postma (2009). "Comparing active vision models". In: *Image and Vision Computing* 27.4, pp. 374 –384 (cit. on p. 68).
Csurka, Gabriela (2017). "Domain adaptation for visual applications: A comprehensive survey". In: *arXiv preprint arXiv:1702.05374* (cit. on pp. 26, 89).
Cui, Yin, Guandao Yang, Andreas Veit, Xun Huang, and Serge Belongie (2018). "Learning to Evaluate Image Captioning". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5804–5812 (cit. on p. 155).
Cullina, Daniel, Arjun Nitin Bhagoji, and Prateek Mittal (2018). "PAC-learning in the presence of evasion adversaries". In: *arXiv preprint arXiv:1806.01471* (cit. on p. 163).

Dalvi, Fahim, Avery Nortonsmith, Anthony Bau, et al. (2019). "NeuroX: A toolkit for analyzing individual neurons in neural networks". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33, pp. 9851–9852 (cit. on p. 135).

Das, Abhishek, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra (2017a). "Human attention in visual question answering: Do humans and deep networks look at the same regions?" In: *Computer Vision and Image Understanding* 163, pp. 90–100 (cit. on pp. 80, 127).

Das, Abhishek, Satwik Kottur, José MF Moura, Stefan Lee, and Dhruv Batra (2017b). "Learning cooperative visual dialog agents with deep reinforcement learning". In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2951–2960 (cit. on p. 112).

Das, Abhishek, Satwik Kottur, Khushi Gupta, et al. (2017c). "Visual dialog". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 326–335 (cit. on p. 86).

Das, Abhishek, Samyak Datta, Georgia Gkioxari, et al. (2018). "Embodied question answering". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2054–2063 (cit. on p. 85).

Das, Srijan, Arpit Chaudhary, Francois Bremond, and Monique Thonnat (2019). "Where to Focus on for Human Action Recognition?" In: *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, pp. 71–80 (cit. on p. 80).

Dasgupta, Sanjoy (2005). "Analysis of a greedy active learning strategy". In: *Advances in neural information processing systems*, pp. 337–344 (cit. on p. 103).

Dathathri, Sumanth, Stephan Zheng, Richard M Murray, and Yisong Yue (2018). "Detecting Adversarial Examples via Neural Fingerprinting". In: *arXiv preprint arXiv:1803.03870* (cit. on p. 167).

Daucé, Emmanuel (2018). "Active Fovea-Based Vision Through Computationally-Effective Model-Based Prediction". In: *Frontiers in neurorobotics* 12, p. 76 (cit. on p. 68).

Day, Oscar and Taghi M Khoshgoftaar (2017). "A survey on heterogeneous transfer learning". In: *Journal of Big Data* 4.1, p. 29 (cit. on p. 89).

De Vries, Harm, Florian Strub, Sarath Chandar, et al. (2017). "Guesswhat?! visual object discovery through multi-modal dialogue". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5503–5512 (cit. on p. 86).

DeAngelus, Marianne and Jeff B Pelz (2009). "Top-down control of eye movements: Yarbus revisited". In: *Visual Cognition* 17.6-7, pp. 790–811 (cit. on p. 80).

DeFelipe, Javier, Lidia Alonso-Nanclares, and Jon I Arellano (2002). "Microstructure of the neocortex: comparative aspects". In: *Journal of neurocytology* 31.3-5, pp. 299–316 (cit. on pp. 33, 35).

Defretin, Joseph (2011). "Stratégies de vision active pour la reconnaissance d'objets". PhD thesis. Ecole normale supérieure de Cachan (cit. on p. 77).

Defretin, Joseph, Stéphane Herbin, Guy Le Besnerais, and Nicolas Vayatis (2010). "Adaptive Planification in Active 3D Object Recognition for Many Classes of Objects". In: *Workshop "Towards Closing the Loop: Active Learning for Robotics", RSS Robotics: Science and Systems Conference* (cit. on p. 68).

Deinzer, F., C. Derichs, H. Niemann, and J. Denzler (2006). "Integrated Viewpoint Fusion and Viewpoint Selection for Optimal Object Recognition". In: *BMVC'06*, I:287 (cit. on p. 77).

Deinzer, Frank, Christian Derichs, Heinrich Niemann, and Joachim Denzler (2009). "A Framework for Actively Selecting Viewpoints in Object Recognition". In: *IJPRAI* 23.4, pp. 765–799 (cit. on p. 77).

Dembo, Amir and Ofer Zeitouni (1998). *Large Deviations Techniques and Applications*. Springer (cit. on p. 73).

Deng, Chaorui, Qi Wu, Qingyao Wu, et al. (2018). "Visual grounding via accumulated attention". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7746–7755 (cit. on p. 80).

Dennett, Daniel C. (2017). "A History of Qualia". In: *Topoi* (cit. on p. 48).

Denzler, Joachim and Christopher M. Brown (2002). "Information Theoretic Sensor Data Selection for Active Object Recognition and State Estimation". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 24.2, pp. 145–157 (cit. on pp. 68, 77).

Der Kiureghian, Armen and Ove Ditlevsen (2009). “Aleatory or epistemic? Does it matter?” In: *Structural Safety* 31.2, pp. 105–112 (cit. on p. 83).

Deselaers, Thomas, Bogdan Alexe, and Vittorio Ferrari (2012). “Weakly supervised localization and learning with generic knowledge”. In: *International journal of computer vision* 100.3, pp. 275–293 (cit. on p. 90).

DeVries, Terrance and Graham W Taylor (2018). “Learning Confidence for Out-of-Distribution Detection in Neural Networks”. In: *arXiv preprint arXiv:1802.04865* (cit. on p. 165).

Devrim Kaba, Mustafa, Mustafa Gokhan Uzunbas, and Ser Nam Lim (2017). “A reinforcement learning approach to the view planning problem”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6933–6941 (cit. on p. 77).

Dhurandhar, Amit, Vijay Iyengar, Ronny Luss, and Karthikeyan Shanmugam (2017). “A formal framework to characterize interpretability of procedures”. In: *ICML Workshop on Human Interpretability in Machine Learning* (cit. on p. 119).

Di Paolo, Ezequiel, Thomas Buhrmann, and Xabier Barandiaran (2017). *Sensorimotor life: An enactive proposal*. Oxford University Press (cit. on p. 55).

Díaz-Rodríguez, Natalia, Vincenzo Lomonaco, David Filliat, and Davide Maltoni (2018). “Don’t forget, there is more than forgetting: new metrics for Continual Learning”. In: *arXiv preprint arXiv:1810.13166* (cit. on p. 106).

DiCarlo, James J, Davide Zoccolan, and Nicole C Rust (2012). “How does the brain solve visual object recognition?” In: *Neuron* 73.3, pp. 415–434 (cit. on pp. 15, 16).

Dickinson, Sven J., Henrik I. Christensen, John K. Tsotsos, and Göran Olofsson (1997). “Active Object Recognition Integrating Attention and Viewpoint Control”. In: *Computer Vision and Image Understanding* 67.3, pp. 239–260 (cit. on p. 76).

Dietterich, Thomas G (2017). “AAAI Presidential Address: Steps Toward Robust Artificial Intelligence”. In: *AI Magazine* 38.3, pp. 3–24 (cit. on p. 161).

Diochnos, Dimitrios I, Saeed Mahloujifar, and Mohammad Mahmoody (2019). “Lower Bounds for Adversarially Robust PAC Learning”. In: *arXiv preprint arXiv:1906.05815* (cit. on p. 158).

Doersch, Carl and Andrew Zisserman (2017). “Multi-task self-supervised visual learning”. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2051–2060 (cit. on p. 21).

Dominique Luzeaux, Jean-René Ruault (2011). *Large scale Complex Systems and Systems of Systems Engineering: Case Studies*. Wiley-ISTE (cit. on p. 51).

Donahue, Jeff, Philipp Krähenbühl, and Trevor Darrell (2016). “Adversarial feature learning”. In: *arXiv preprint arXiv:1605.09782* (cit. on p. 91).

Donnarumma, Francesco, Marcello Costantini, Ettore Ambrosini, Karl Friston, and Giovanni Pezzulo (2017). “Action perception as hypothesis testing”. In: *Cortex* 89, pp. 45–60 (cit. on p. 39).

Doran, Derek, Sarah Schulz, and Tarek R Besold (2017). “What Does Explainable AI Really Mean? A New Conceptualization of Perspectives”. In: *arXiv preprint arXiv:1710.00794* (cit. on p. 118).

Doshi-Velez, Finale and Been Kim (2017). “A roadmap for a rigorous science of interpretability”. In: *arXiv preprint arXiv:1702.08608* (cit. on pp. 117, 118).

Dosovitskiy, Alexey and Thomas Brox (2016). “Inverting visual representations with convolutional networks”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4829–4837 (cit. on p. 123).

Draper, Bruce A, Robert T Collins, John Brolio, Allen R Hanson, and Edward M Riseman (1989). “The schema system”. In: *International Journal of Computer Vision* 2.3, pp. 209–250 (cit. on p. 13).

Draper, Bruce A, Jose Bins, and Kyungim Baek (1999). “ADORE: adaptive object recognition”. In: *International Conference on Computer Vision Systems*. Springer, pp. 522–537 (cit. on p. 13).

Drayson, Zoe (2017). “Modularity and the Predictive Mind”. In: *Philosophy and Predictive Processing*. Ed. by Thomas K. Metzinger and Wanja Wiese. Frankfurt am Main: MIND Group. Chap. 12 (cit. on p. 40).

Dreossi, Tommaso, Shromona Ghosh, Alberto Sangiovanni-Vincentelli, and Sanjit A Seshia (2017). "Systematic testing of convolutional neural networks for autonomous driving". In: *arXiv preprint arXiv:1708.03309* (cit. on p. 162).

Drummond, John J (1979). "On seeing a material thing in space: The role of kinaesthesis in visual perception". In: *Philosophy and Phenomenological Research* 40.1, pp. 19–32 (cit. on p. 48).

– (2012). "Intentionality without representationalism". In: *The Oxford Handbook of Contemporary Phenomenology*. Ed. by Dan Zahavi. Oxford University Press, pp. 115–133 (cit. on pp. 46, 48).

Duchi, John and Yoram Singer (2009). "Efficient online and batch learning using forward backward splitting". In: *Journal of Machine Learning Research* 10.Dec, pp. 2899–2934 (cit. on p. 113).

Duclos, Daniel, Jacques Lonnoy, Quentin Guillerm, et al. (2008a). "ROBIN: a platform for evaluating, Automatic Target Recognition algorithms., Part 1: Overview of the project and presentation of the SAGEM DS competition". In: *Automatic Target Recognition XVIII, , 2008*. Ed. by Firooz A. Sadjadi and Abhijit Mahalanobis. Vol. 6967. Proceedings of SPIE. France: SPIE (cit. on p. 155).

– (2008b). "ROBIN: a platform for evaluating Automatic Target Recognition algorithms. Part 2: protocols used for evaluating algorithms and results obtained on the SAGEM DS database". In: *Automatic target recognition XVIII, March, 2008*. Ed. by Firooz A. Sadjadi and Abhijit Mahalanobis. Vol. 6967. Proceedings of SPIE, the International Society for Optical Engineering. Orlando, FL, Etats-Unis: SPIE, pp. 1–10 (cit. on p. 155).

Dulac-Arnold, Gabriel, Ludovic Denoyer, Nicolas Thome, Matthieu Cord, and Patrick Gallinari (2013). "Sequentially generated instance-dependent image representations for classification". In: *arXiv preprint arXiv:1312.6594* (cit. on p. 71).

Dumoulin, Vincent, Ethan Perez, Nathan Schucher, et al. (2018). "Feature-wise transformations". In: *Distill*. https://distill.pub/2018/feature-wise-transformations (cit. on p. 81).

Durand, Thibaut, Nicolas Thome, and Matthieu Cord (2016). "Weldon: Weakly supervised learning of deep convolutional neural networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4743–4752 (cit. on p. 90).

Durand, Thibaut, Taylor Mordan, Nicolas Thome, and Matthieu Cord (2017). "Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 642–651 (cit. on p. 90).

Elsken, Thomas, Jan Hendrik Metzen, and Frank Hutter (2019). "Neural Architecture Search: A Survey." In: *Journal of Machine Learning Research* 20.55, pp. 1–21 (cit. on pp. 88, 111).

Engel, Andreas K, Pascal Fries, and Wolf Singer (2001). "Dynamic predictions: oscillations and synchrony in top–down processing". In: *Nature Reviews Neuroscience* 2.10, p. 704 (cit. on p. 37).

Engelmann, Jan B, Eswar Damaraju, Srikanth Padmala, and Luiz Pessoa (2009). "Combined effects of attention and motivation on visual task performance: transient and sustained motivational effects". In: *Frontiers in human neuroscience* 3, p. 4 (cit. on p. 44).

Erfani, Sarah M, Sutharshan Rajasegarar, Shanika Karunasekera, and Christopher Leckie (2016). "High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning". In: *Pattern Recognition* 58, pp. 121–134 (cit. on p. 165).

Evgeniou, Theodoros and Massimiliano Pontil (2004). "Regularized multi–task learning". In: *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 109–117 (cit. on p. 177).

Evtimov, Ivan, Kevin Eykholt, Earlence Fernandes, et al. (2017). "Robust physical-world attacks on machine learning models". In: *arXiv preprint arXiv:1707.08945* (cit. on p. 159).

Farquhar, Sebastian and Yarin Gal (2018). "Towards Robust Evaluations of Continual Learning". In: *arXiv preprint arXiv:1805.09733* (cit. on p. 106).

Faugeras, Olivier (1993). *Three-dimensional computer vision: a geometric viewpoint*. MIT press (cit. on p. 97).

Fazlyab, Mahyar, Alexander Robey, Hamed Hassani, Manfred Morari, and George Pappas (2019). "Efficient and accurate estimation of lipschitz constants for deep neural networks". In: *Advances in Neural Information Processing Systems*, pp. 11423–11434 (cit. on p. 163).

Fei-Fei, Li, Rob Fergus, and Pietro Perona (2006). "One-shot learning of object categories". In: *IEEE transactions on pattern analysis and machine intelligence* 28.4, pp. 594–611 (cit. on p. 92).

Feinman, Reuben, Ryan R Curtin, Saurabh Shintre, and Andrew B Gardner (2017). "Detecting adversarial samples from artifacts". In: *arXiv preprint arXiv:1703.00410* (cit. on p. 167).

Feldman, Harriet and Karl Friston (2010). "Attention, uncertainty, and free-energy". In: *Frontiers in human neuroscience* 4, p. 215 (cit. on pp. 37, 39).

Felleman, Daniel J and DC Essen Van (1991). "Distributed hierarchical processing in the primate cerebral cortex." In: *Cerebral cortex (New York, NY: 1991)* 1.1, pp. 1–47 (cit. on p. 13).

Felzenszwalb, Pedro F, Ross B Girshick, David McAllester, and Deva Ramanan (2009). "Object detection with discriminatively trained part-based models". In: *IEEE transactions on pattern analysis and machine intelligence* 32.9, pp. 1627–1645 (cit. on p. 97).

Fernando, Chrisantha, Dylan Banarse, Charles Blundell, et al. (2017). "Pathnet: Evolution channels gradient descent in super neural networks". In: *arXiv preprint arXiv:1701.08734* (cit. on p. 106).

Fidler, Sanja, Marko Boben, and Aleš Leonardis (2010). "A coarse-to-fine taxonomy of constellations for fast multi-class object detection". In: *Computer Vision–ECCV 2010*. Springer, pp. 687–700 (cit. on p. 72).

Finn, Chelsea, Pieter Abbeel, and Sergey Levine (2017). "Model-agnostic meta-learning for fast adaptation of deep networks". In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, pp. 1126–1135 (cit. on p. 92).

Firestone, Chaz and Brian J Scholl (2016). "Cognition does not affect perception: Evaluating the evidence for" top-down" effects". In: *Behavioral and brain sciences* 39 (cit. on pp. 17, 41, 44).

Fischer, Volker, Mummadi Chaithanya Kumar, Jan Hendrik Metzen, and Thomas Brox (2017). "Adversarial Examples for Semantic Image Segmentation". In: *International Conference on Learning Representations* (cit. on p. 156).

Fish, William (2009). *Perception, hallucination, and illusion*. OUP USA (cit. on p. 46).

– (2010). *Philosophy of perception: A contemporary introduction*. Routledge (cit. on p. 45).

Foerster, Jakob, Ioannis Alexandros Assael, Nando de Freitas, and Shimon Whiteson (2016). "Learning to communicate with deep multi-agent reinforcement learning". In: *Advances in Neural Information Processing Systems*, pp. 2137–2145 (cit. on p. 112).

Forestier, Sébastien, Yoan Mollard, and Pierre-Yves Oudeyer (2017). "Intrinsically motivated goal exploration processes with automatic curriculum learning". In: *arXiv preprint arXiv:1708.02190* (cit. on p. 114).

Forsyth, David A and Jean Ponce (2002). *Computer vision: a modern approach*. Prentice Hall Professional Technical Reference (cit. on p. 97).

François-Lavet, Vincent, Peter Henderson, Riashat Islam, Marc G Bellemare, Joelle Pineau, et al. (2018). "An introduction to deep reinforcement learning". In: *Foundations and Trends® in Machine Learning* 11.3-4, pp. 219–354 (cit. on pp. 68, 71).

Franklin, Stan and Art Graesser (1996). "Is it an Agent, or just a Program?: A Taxonomy for Autonomous Agents". In: *International Workshop on Agent Theories, Architectures, and Languages*. Springer, pp. 21–35 (cit. on p. 18).

French, Robert M. (1999). "Catastrophic forgetting in connectionist networks". In: *Trends in Cognitive Sciences* 3.4, pp. 128 –135 (cit. on p. 105).

Frintrop, Simone, Erich Rome, and Henrik I. Christensen (2010). "Computational visual attention systems and their cognitive foundations: A survey". In: *TAP* 7.1 (cit. on p. 80).

Friston, Karl (2010). "The free-energy principle: a unified brain theory?" In: *Nature reviews neuroscience* 11.2, p. 127 (cit. on pp. 36, 40).

– (2012). "The history of the future of the Bayesian brain". In: *NeuroImage* 62.2, pp. 1230–1233 (cit. on p. 36).

Friston, Karl, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, and Giovanni Pezzulo (2017). "Active inference: a process theory". In: *Neural computation* 29.1, pp. 1–49 (cit. on p. 68).

Friston, Karl J, Klaas Enno Stephan, Read Montague, and Raymond J Dolan (2014). "Computational psychiatry: the brain as a phantastic organ". In: *The Lancet Psychiatry* 1.2, pp. 148–158 (cit. on p. 39).

Frith, Chris (2007). *Making up the mind: How the brain creates our mental world*. Blackwell Publishing (cit. on pp. 31, 40).

Fritsch, Jannik, Tobias Kuhnl, and Andreas Geiger (2013). "A new performance measure and evaluation benchmark for road detection algorithms". In: *Intelligent Transportation Systems-(ITSC), 2013 16th International IEEE Conference on*. IEEE, pp. 1693–1700 (cit. on p. 155).

Fu, Jianlong, Heliang Zheng, and Tao Mei (2017). "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4438–4446 (cit. on p. 80).

Fu, Yanwei, Timothy M Hospedales, Tao Xiang, Zhenyong Fu, and Shaogang Gong (2014). "Transductive multi-view embedding for zero-shot recognition and annotation". In: *European Conference on Computer Vision* (cit. on p. 92).

Fu, Yifan, Xingquan Zhu, and Bin Li (2013). "A survey on instance selection for active learning". In: *Knowledge and information systems* 35.2, pp. 249–283 (cit. on pp. 103, 146).

Fukushima, Kunihiko (1980). "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position". In: *Biological cybernetics* 36.4, pp. 193–202 (cit. on p. 88).

Gabriel, Markus (2017). *I am not a brain.* Polity Press (cit. on p. 48).

Gal, Yarin and Zoubin Ghahramani (2016). "Dropout as a bayesian approximation: Representing model uncertainty in deep learning". In: *international conference on machine learning*, pp. 1050–1059 (cit. on p. 83).

Gal, Yarin, Riashat Islam, and Zoubin Ghahramani (2017). "Deep bayesian active learning with image data". In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, pp. 1183–1192 (cit. on p. 104).

Galassi, Andrea, Marco Lippi, and Paolo Torroni (2019). "Attention, please! a critical review of neural attention models in natural language processing". In: *arXiv preprint arXiv:1902.02181* (cit. on p. 81).

Gallagher, Shaun (2012). "On the possibility of naturalizing phenomenology". In: *The Oxford Handbook of Contemporary Phenomenology*, pp. 70–93 (cit. on p. 49).

– (2017). *Enactivist interventions: Rethinking the mind*. Oxford University Press (cit. on pp. 19, 47).

Gallagher, Shaun and Daniel Schmicking (2010). *Handbook of phenomenology and cognitive science*. Springer (cit. on p. 47).

Gallinari, Patrick, Sylvie Thiria, Fouad Badran, and Françoise Fogelman-Soulie (1991). "On the relations between discriminant analysis and multilayer perceptrons". In: *neural networks* 4.3, pp. 349–360 (cit. on p. 121).

Gallos, Dimitrios and Frank Ferrie (2019). "Active Vision in the Era of Convolutional Neural Networks". In: *2019 16th Conference on Computer and Robot Vision (CRV)*. IEEE, pp. 81–88 (cit. on p. 83).

Ganeri, Jonardon (2017). *Attention, not self*. Oxford University Press (cit. on p. 42).

Gangaputra, Sachin and Donald Geman (2006). "A design principle for coarse-to-fine classification". In: *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*. Vol. 2. IEEE, pp. 1877–1884 (cit. on p. 72).

Gangopadhyay, Nivedita and Julian Kiverstein (2009). "Enactivism and the unity of perception and action". In: *Topoi* 28.1, pp. 63–73 (cit. on p. 47).

Gao, Jianfeng, Michel Galley, Lihong Li, et al. (2019). "Neural approaches to conversational AI". In: *Foundations and Trends® in Information Retrieval* 13.2-3, pp. 127–298 (cit. on p. 86).

Gao, Mingfei, Ruichi Yu, Ang Li, Vlad I Morariu, and Larry S Davis (2018). “Dynamic zoom-in network for fast object detection in large images”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6926–6935 (cit. on p. 81).

Gardella, Christophe, Olivier Marre, and Thierry Mora (2019). “Modeling the correlated activity of neural populations: A review”. In: *Neural computation* 31.2, pp. 233–269 (cit. on p. 36).

Gazzaley, Adam and Anna C Nobre (2012). “Top-down modulation: bridging selective attention and working memory”. In: *Trends in cognitive sciences* 16.2, pp. 129–135 (cit. on p. 44).

Gazzaniga, Michael and Richard B Ivry (2013). *Cognitive Neuroscience: The Biology of the Mind: Fourth International Student Edition*. WW Norton (cit. on pp. 31, 42, 43).

Geifman, Yonatan and Ran El-Yaniv (2017). “Selective classification for deep neural networks”. In: *Advances in neural information processing systems*, pp. 4878–4887 (cit. on p. 164).

– (2019). “SelectiveNet: A Deep Neural Network with an Integrated Reject Option”. In: *International Conference on Machine Learning*, pp. 2151–2159 (cit. on p. 164).

Genesereth, Michael R and Nils J Nilsson (2012). *Logical foundations of artificial intelligence*. Morgan Kaufmann (cit. on p. 96).

Gepperth, Alexander and Barbara Hammer (2016). “Incremental learning algorithms and applications”. In: *European Symposium on Artificial Neural Networks (ESANN)* (cit. on pp. 104, 106).

Gepperth, Alexander and Cem Karaoguz (2016). “A bio-inspired incremental learning architecture for applied perceptual problems”. In: *Cognitive Computation* 8.5, pp. 924–934 (cit. on p. 106).

Ghallab, Malik, Dana Nau, and Paolo Traverso (2004). *Automated Planning: theory and practice*. Elsevier (cit. on p. 71).

Gibson, J.J. (2019). *The Ecological Approach to Visual Perception*. Boston: Houghton Mifflin (cit. on pp. 19, 41, 104).

Gilbert, Charles D and Wu Li (2013). “Top-down influences on visual processing”. In: *Nature Reviews Neuroscience* 14.5, p. 350 (cit. on p. 16).

Gillebert, Céline R and Glyn W Humphreys (2013). “Mutual interplay between perceptual organization and attention”. In: *The Oxford Handbook of Perceptual Organization* (cit. on p. 42).

Gilmer, Justin, Ryan P Adams, Ian Goodfellow, David Andersen, and George E Dahl (2018). “Motivating the Rules of the Game for Adversarial Example Research”. In: *arXiv preprint arXiv:1807.06732* (cit. on p. 160).

Gilpin, Leilani H, David Bau, Ben Z Yuan, et al. (2018). “Explaining Explanations: An Approach to Evaluating Interpretability of Machine Learning”. In: *arXiv preprint arXiv:1806.00069* (cit. on pp. 118, 119, 131).

Gilton, Davis, Greg Ongie, and Rebecca Willett (2019). “Neumann networks for inverse problems in imaging”. In: *arXiv preprint arXiv:1901.03707* (cit. on p. 101).

Girard, BetABERTHOZ and Alain Berthoz (2005). “From brainstem to cortex: computational models of saccade generation circuitry”. In: *Progress in neurobiology* 77.4, pp. 215–251 (cit. on p. 55).

Godard, Clément, Oisin Mac Aodha, and Gabriel J Brostow (2017). “Unsupervised monocular depth estimation with left-right consistency”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 270–279 (cit. on p. 101).

Goldstein, E Bruce (2010). *Encyclopedia of Perception*. SAGE (cit. on p. 29).

Gollisch, Tim and Markus Meister (2008). “Rapid neural coding in the retina with relative spike latencies”. In: *science* 319.5866, pp. 1108–1111 (cit. on p. 32).

– (2010). “Eye smarter than scientists believed: neural computations in circuits of the retina”. In: *Neuron* 65.2, pp. 150–164 (cit. on p. 32).

Gong, Zhitao, Wenlu Wang, and Wei-Shinn Ku (2017). “Adversarial and clean data are not twins”. In: *arXiv preprint arXiv:1704.04960* (cit. on p. 166).

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, et al. (2014a). “Generative adversarial nets”. In: *Advances in neural information processing systems*, pp. 2672–2680 (cit. on p. 94).

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep learning*. MIT press (cit. on p. 87).

Goodfellow, Ian, Patrick McDaniel, and Nicolas Papernot (2018). “Making Machine Learning Robust Against Adversarial Inputs”. In: *Commun. ACM* 61.7, pp. 56–66 (cit. on pp. 159, 160).

Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy (2014b). “Explaining and Harnessing Adversarial Examples”. In: *CoRR* abs/1412.6572 (cit. on p. 156).

Gopinath, Divya, Guy Katz, Corina S Pasareanu, and Clark Barrett (2017). “Deepsafe: A data-driven approach for checking adversarial robustness in neural networks”. In: *arXiv preprint arXiv:1710.00486* (cit. on p. 162).

Gordon, Daniel, Aniruddha Kembhavi, Mohammad Rastegari, et al. (2018). “Iqa: Visual question answering in interactive environments”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4089–4098 (cit. on p. 85).

Gosselin, Philippe Henri and Matthieu Cord (2008). “Active learning methods for interactive image retrieval”. In: *IEEE Transactions on Image Processing* 17.7, pp. 1200–1211 (cit. on p. 103).

Goswami, Gaurav, Akshay Agarwal, Nalini Ratha, Richa Singh, and Mayank Vatsa (2019). “Detecting and mitigating adversarial perturbations for robust face recognition”. In: *International Journal of Computer Vision* 127.6-7, pp. 719–742 (cit. on p. 156).

Goyal, Yash, Akrit Mohapatra, Devi Parikh, and Dhruv Batra (2016). “Towards transparent ai systems: Interpreting visual question answering models”. In: *arXiv preprint arXiv:1608.08974* (cit. on p. 125).

Grauman, Kristen and Bastian Leibe (2011). “Visual object recognition”. In: *Synthesis lectures on artificial intelligence and machine learning* 5.2, pp. 1–181 (cit. on p. 97).

Graves, Alex, Greg Wayne, and Ivo Danihelka (2014). “Neural turing machines”. In: *arXiv preprint arXiv:1410.5401* (cit. on p. 52).

Greene, Michelle R, Tommy Liu, and Jeremy M Wolfe (2012). “Reconsidering Yarbus: A failure to predict observers’ task from eye movement patterns”. In: *Vision research* 62, pp. 1–8 (cit. on pp. 42, 80).

Gretton, Arthur, Dino Sejdinovic, Heiko Strathmann, et al. (2012). “Optimal kernel choice for large-scale two-sample tests”. In: *Advances in neural information processing systems*, pp. 1205–1213 (cit. on p. 95).

Gribonval, Rémi, Gitta Kutyniok, Morten Nielsen, and Felix Voigtlaender (2019). “Approximation spaces of deep neural networks”. In: *arXiv preprint arXiv:1905.01208* (cit. on p. 108).

Grosse, Kathrin, Praveen Manoharan, Nicolas Papernot, Michael Backes, and Patrick McDaniel (2017). “On the (statistical) detection of adversarial examples”. In: *arXiv preprint arXiv:1702.06280* (cit. on p. 166).

Guidotti, Riccardo, Anna Monreale, Salvatore Ruggieri, et al. (2018). “A survey of methods for explaining black box models”. In: *ACM computing surveys (CSUR)* 51.5, p. 93 (cit. on p. 118).

Guilmart, Christophe, Stéphane Herbin, and Patrick Pérez (2011). “Context-driven moving object detection in aerial scenes with user input”. In: *18th IEEE International Conference on Image Processing, ICIP 2011, Brussels, Belgium, September 11-14, 2011*, pp. 1781–1784 (cit. on p. 100).

Guinet, J. (2008). “Reconnaissance Multi-vues de véhicules sur séquences d’images”. PhD thesis. Université de Cergy-Pontoise (cit. on pp. 79, 97, 98, 233).

Guinet, J., S. Herbin, G. Le Besnerais, and S. Philipp-Foliguet (2007). “Extrapolation d’aspect pour l’acquisition de cibles sur séquences aériennes”. In: *actes du GRETSI, Troyes* (cit. on p. 98).

Gunning, David (2017). “Explainable artificial intelligence (xai)”. In: *Defense Advanced Research Projects Agency (DARPA)* (cit. on p. 120).

Guo, Chuan, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger (2017). “On calibration of modern neural networks”. In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, pp. 1321–1330 (cit. on p. 83).

Guo, Michelle, Albert Haque, De-An Huang, Serena Yeung, and Li Fei-Fei (2018a). “Dynamic task prioritization for multitask learning”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 270–287 (cit. on p. 109).

Guo, Pei, Connor Anderson, Kolten Pearson, and Ryan Farrell (2018b). “Neural Network Interpretation via Fine Grained Textual Summarization”. In: *arXiv preprint arXiv:1805.08969* (cit. on p. 126).

Guo, Sheng, Weilin Huang, Haozhi Zhang, et al. (2018c). “Curriculumnet: Weakly supervised learning from large-scale web images”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 135–150 (cit. on p. 114).

Guo, W., J. Wang, and S. Wang (2019). “Deep Multimodal Representation Learning: A Survey”. In: *IEEE Access* 7, pp. 63373–63394 (cit. on p. 136).

Guo, Yanlin, Cen Rao, S. Samarasekera, et al. (2008). “Matching vehicles under large pose transformations using approximate 3D models and piecewise MRF model”. In: *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–8 (cit. on p. 79).

Hains, G., A. Jakobsson, and Y. Khmelevsky (2018). “Towards formal methods and software engineering for deep learning: Security, safety and productivity for dl systems development”. In: *2018 Annual IEEE International Systems Conference (SysCon)*, pp. 1–5 (cit. on p. 163).

Hamdoun, O., F. Moutarde, B. Stanciulescu, and B. Steux (2008). “Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences”. In: *Distributed Smart Cameras, 2008. ICDSC 2008. Second ACM/IEEE International Conference on*, pp. 1–6 (cit. on p. 79).

Harris, Kenneth D. and Thomas D. Mrsic-Flogel (2013). “Cortical connectivity and sensory coding”. In: *Nature* 503.7474, p. 51 (cit. on pp. 33, 34).

Hartley, Richard and Andrew Zisserman (2003). *Multiple view geometry in computer vision*. Cambridge university press (cit. on p. 97).

Hassabis, Demis, Dharshan Kumaran, Christopher Summerfield, and Matthew Botvinick (2017). “Neuroscience-inspired artificial intelligence”. In: *Neuron* 95.2, pp. 245–258 (cit. on p. 55).

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (2016). “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778 (cit. on pp. 33, 88).

Hecker, Simon, Dengxin Dai, and Luc Van Gool (2018). “Failure Prediction for Autonomous Driving”. In: *arXiv preprint arXiv:1805.01811* (cit. on p. 169).

Helmholtz, Hermann von and James P. C. Southall (1924). *Helmholtz’s treatise on physiological optics, Vol. 1, Trans*. Optical Society of America (cit. on pp. 8, 38).

Hendricks, Lisa Anne, Zeynep Akata, Marcus Rohrbach, et al. (2016). “Generating visual explanations”. In: *European Conference on Computer Vision*. Springer, Cham, pp. 3–19 (cit. on p. 126).

Hendrik Metzen, Jan, Mummadi Chaithanya Kumar, Thomas Brox, and Volker Fischer (2017). “Universal adversarial perturbations against semantic image segmentation”. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2755–2764 (cit. on p. 156).

Hendrycks, Dan and Kevin Gimpel (2017). “A baseline for detecting misclassified and out-of-distribution examples in neural networks”. In: *International Conference on Learning Representations* (cit. on p. 165).

Herbin, S., F. Champagnat, J. Israel, et al. (2012). “Scene Understanding from Aerospace Sensors: What can be Expected?” In: *AerospaceLab* AL04-06 (cit. on p. 155).

Herbin, Stéphane (1996). “Recognizing 3D Objects by Generating Random Actions”. In: *1996 Conference on Computer Vision and Pattern Recognition (CVPR ’96), June 18-20, 1996 San Francisco, CA, USA*, pp. 35–40 (cit. on pp. 73, 75, 76).

– (1997a). “Elements pour la formalisation d’une reconnaissance active. Application a la vision tri-dimensionnelle”. PhD thesis. École normale supérieure de Cachan - ENS Cachan (cit. on p. 49).

– (1997b). “Graphes d’aspects probabilisés”. In: *Actes du 11ème Congrés AFCET-AFIA, RFIA’98*. Vol. I, pp. 87–96 (cit. on pp. 73, 75).

– (1998). “Combining Geometric and Probabilistic Structure for Active Recognition of 3D Objects”. In: *Computer Vision - ECCV’98, 5th European Conference on Computer Vision, Freiburg, Germany, June 2-6, 1998, Proceedings, Volume II*, pp. 748–764 (cit. on pp. 73, 75).

– (2002). “Similarity Measures Between Feature Maps - Application to texture comparison”. In: *Proceedings of the Texture 2002 workshop*, pp. 67–72 (cit. on pp. 73, 75).

Herbin, Stéphane (2003). “Active Sampling Strategies for Multihyposthesis Testing”. In: *Energy Minimization Methods in Computer Vision and Pattern Recognition, 4th International Workshop, EMMCVPR 2003, Lisbon, Portugal, July 7-9, 2003, Proceedings*, pp. 97–112 (cit. on p. 73).
– (2004). “Robust Multihypothesis Discrimination of Controlled I.I.D. Processes”. In: *17th International Conference on Pattern Recognition, ICPR 2004, Cambridge, UK, August 23-26, 2004*. Pp. 200–203 (cit. on p. 73).
– (2014). “Fine-grained recognition by sequential hypothesis rejection and foveated vision on parts”. In: *ECCV 2014 workshop on Parts and Attributes* (cit. on pp. 68, 69, 233).
Hero, Alfred O and Douglas Cochran (2011). “Sensor management: Past, present, and future”. In: *IEEE Sensors Journal* 11.12, pp. 3064–3075 (cit. on p. 79).
Hero, Alfred Olivier, David Castañón, Doug Cochran, and Keith Kastella (2007). *Foundations and applications of sensor management*. Springer Science & Business Media (cit. on p. 79).
Higgins, Irina, David Amos, David Pfau, et al. (2018). “Towards a definition of disentangled representations”. In: *arXiv preprint arXiv:1812.02230* (cit. on p. 91).
Hildreth, Ellen C. and Shimon Ullman (1988). “The computational study of vision”. In: *Foundations of cognitive science*. Ed. by Michael I Posner. MIT Press, pp. 581–630 (cit. on p. 17).
Hinton, Geoffrey, Oriol Vinyals, and Jeffrey Dean (2015). “Distilling the Knowledge in a Neural Network”. In: *NIPS Deep Learning and Representation Learning Workshop* (cit. on p. 105).
Hoffman, Robert R, Shane T Mueller, Gary Klein, and Jordan Litman (2018). “Metrics for explainable ai: Challenges and prospects”. In: *arXiv preprint arXiv:1812.04608* (cit. on p. 131).
Hohman, Fred Matthew, Minsuk Kahng, Robert Pienta, and Duen Horng Chau (2018). “Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers”. In: *IEEE Transactions on Visualization and Computer Graphics* (cit. on pp. 118, 123).
Hohwy, Jakob (2013). *The predictive mind*. Oxford University Press (cit. on p. 38).
– (2016). “The self-evidencing brain”. In: *Noûs* 50.2, pp. 259–285 (cit. on p. 38).
Hong, Seunghoon, Suha Kwak, and Bohyung Han (2017). “Weakly supervised learning with deep convolutional neural networks for semantic segmentation: Understanding semantic layout of images with minimum human supervision”. In: *IEEE Signal Processing Magazine* 34.6, pp. 39–49 (cit. on p. 90).
Hong, Yongjun, Uiwon Hwang, Jaeyoon Yoo, and Sungroh Yoon (2019). “How Generative Adversarial Networks and Their Variants Work: An Overview”. In: *ACM Computing Surveys (CSUR)* 52.1, p. 10 (cit. on p. 94).
Hossain, MD, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga (2019). “A comprehensive survey of deep learning for image captioning”. In: *ACM Computing Surveys (CSUR)* 51.6, p. 118 (cit. on p. 80).
Howard, Andrew G, Menglong Zhu, Bo Chen, et al. (2017). “Mobilenets: Efficient convolutional neural networks for mobile vision applications”. In: *arXiv preprint arXiv:1704.04861* (cit. on p. 89).
Hsu, Yen-Chang, Yen-Cheng Liu, and Zsolt Kira (2018). “Re-evaluating Continual Learning Scenarios: A Categorization and Case for Strong Baselines”. In: *arXiv preprint arXiv:1810.12488* (cit. on p. 105).
Hu, Ronghang, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko (2017). “Learning to Reason: End-To-End Module Networks for Visual Question Answering”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 804–813 (cit. on p. 128).
Huang, Gao, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger (2017a). “Densely connected convolutional networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708 (cit. on pp. 33, 88).
Huang, Gao, Danlu Chen, Tianhong Li, et al. (2017b). “Multi-scale dense networks for resource efficient image classification”. In: *arXiv preprint arXiv:1703.09844* (cit. on p. 71).
Huang, Jonathan, Vivek Rathod, Chen Sun, et al. (2017c). “Speed/Accuracy Trade-Offs for Modern Convolutional Object Detectors”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7310–7311 (cit. on p. 89).

Huang, Xiaowei, Marta Kwiatkowska, Sen Wang, and Min Wu (2017d). “Safety verification of deep neural networks”. In: *International Conference on Computer Aided Verification*. Springer, pp. 3–29 (cit. on p. 162).

Huang, Xiaowei, Daniel Kroening, Marta Kwiatkowska, et al. (2018). “Safety and Trustworthiness of Deep Neural Networks: A Survey”. In: *arXiv preprint arXiv:1812.08342* (cit. on p. 151).

Hudelot, Céline (2005). “Towards a cognitive vision platform for semantic image interpretation; application to the recognition of biological organisms”. PhD thesis. Université Nice Sophia Antipolis (cit. on p. 14).

Huemer, Michael (2016). “Sense-Data”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2016. Metaphysics Research Lab, Stanford University (cit. on p. 46).

Husserl, Edmund (1931). *Ideas pertaining to a pure phenomenology and to a phenomenological philosophy: First book: General introduction to a pure phenomenology*. Tr. by F. Kersten. Kluwer Academic Publishers (cit. on p. 47).

– (1997). *Thing and space: Lectures of 1907*. Vol. 7. tr. R. Rojcewicz. Springer Science & Business Media (cit. on pp. 47, 48).

Hülse, M, S McBride, J Law, and M Lee (2010). “Integration of Active Vision and Reaching From a Developmental Robotics Perspective”. In: *IEEE Transactions on Autonomous Mental Development* 2.4, pp. 355–367 (cit. on p. 13).

Idier, Jérôme (2013). *Bayesian approach to inverse problems*. John Wiley & Sons (cit. on p. 100).

Ilievski, Ilija and Jiashi Feng (2017). “Multimodal learning and reasoning for visual question answering”. In: *Advances in Neural Information Processing Systems*, pp. 551–562 (cit. on p. 128).

Ilyas, Andrew, Shibani Santurkar, Dimitris Tsipras, et al. (2019). “Adversarial examples are not bugs, they are features”. In: pp. 125–136 (cit. on p. 160).

Iscen, Ahmet, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum (2019). “Label propagation for deep semi-supervised learning”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5070–5079 (cit. on p. 90).

Ishii, Shin, Wako Yoshida, and Junichiro Yoshimoto (2002). “Control of exploitation–exploration meta-parameter in reinforcement learning”. In: *Neural networks* 15.4-6, pp. 665–687 (cit. on p. 111).

Isola, Phillip, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros (2017). “Image-to-image translation with conditional adversarial networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134 (cit. on p. 94).

Ivaldi, Serena, Natalia Lyubova, Alain Droniou, et al. (2013). “Object learning through active exploration”. In: *IEEE Transactions on Autonomous Mental Development* 6.1, pp. 56–72 (cit. on pp. 62, 91).

Jaderberg, Max, Karen Simonyan, Andrew Zisserman, et al. (2015). “Spatial transformer networks”. In: *Advances in neural information processing systems*, pp. 2017–2025 (cit. on p. 81).

Jain, Prateek and Ashish Kapoor (2009). “Active learning for large multi-class problems”. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 762–769 (cit. on p. 103).

Jain, Unnat, Svetlana Lazebnik, and Alexander G Schwing (2018). “Two can play this game: visual dialog with discriminative question generation and answering”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5754–5763 (cit. on p. 86).

Jan, Steve TK, Joseph Messou, Yen-Chen Lin, Jia-Bin Huang, and Gang Wang (2019). “Connecting the Digital and Physical World: Improving the Robustness of Adversarial Attacks”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33, pp. 962–969 (cit. on p. 159).

Janai, Joel, Fatma Güney, Aseem Behl, and Andreas Geiger (2017). “Computer vision for autonomous vehicles: Problems, datasets and state-of-the-art”. In: *arXiv preprint arXiv:1704.05519* (cit. on p. 140).

Janisch, Jaromír, Tomáš Pevnỳ, and Viliam Lisỳ (2019). “Classification with costly features using deep reinforcement learning”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33, pp. 3959–3966 (cit. on p. 71).

Jayaraman, Dinesh and Kristen Grauman (2018). “End-to-end policy learning for active visual categorization”. In: *IEEE transactions on pattern analysis and machine intelligence* (cit. on p. 84).

Jing, Longlong and Yingli Tian (2019). “Self-supervised Visual Feature Learning with Deep Neural Networks: A Survey”. In: *arXiv preprint arXiv:1902.06162* (cit. on pp. 26, 91).

Joe, Byunggill, Sung Ju Hwang, and Insik Shin (2019). “Learning to Disentangle Robust and Vulnerable Features for Adversarial Detection”. In: *arXiv preprint arXiv:1909.04311* (cit. on p. 160).

Johnson, Justin, Bharath Hariharan, Laurens van der Maaten, et al. (2017a). “CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning”. In: *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. IEEE, pp. 1988–1997 (cit. on p. 128).

Johnson, Justin, Bharath Hariharan, Laurens van der Maaten, et al. (2017b). “Inferring and Executing Programs for Visual Reasoning”. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, pp. 3008–3017 (cit. on p. 128).

Johnson, Kenneth O (2000). “Neural coding”. In: *Neuron* 26.3, pp. 563–566 (cit. on p. 36).

Jordan, Matt, Justin Lewis, and Alexandros G Dimakis (2019). “Provable certificates for adversarial examples: Fitting a ball in the union of polytopes”. In: *Advances in Neural Information Processing Systems*, pp. 14059–14069 (cit. on p. 167).

Joshi, Ajay J, Fatih Porikli, and Nikolaos Papanikolopoulos (2009). “Multi-class active learning for image classification”. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 2372–2379 (cit. on p. 103).

Kandel, Eric R, James H Schwartz, Thomas M Jessell, et al. (2013). *Principles of neural science*. 5th. McGraw-Hill Medical (cit. on pp. 31, 34, 36).

Karasev, Vasiliy, Alessandro Chiuso, and Stefano Soatto (2012). “Controlled recognition bounds for visual learning and exploration”. In: *Advances in neural information processing systems*, pp. 2915–2923 (cit. on p. 77).

Karayev, Sergey, Tobias Baumgartner, Mario Fritz, and Trevor Darrell (2012). “Timely Object Recognition.” In: *NIPS*, pp. 899–907 (cit. on p. 71).

Katz, Guy, Clark Barrett, David L Dill, Kyle Julian, and Mykel J Kochenderfer (2017). “Reluplex: An efficient SMT solver for verifying deep neural networks”. In: *International Conference on Computer Aided Verification*. Springer, pp. 97–117 (cit. on p. 162).

Kawaguchi, Kenji, Leslie Pack Kaelbling, and Yoshua Bengio (2017). “Generalization in deep learning”. In: *arXiv preprint arXiv:1710.05468* (cit. on p. 108).

Kemker, Ronald, Marc McClure, Angelina Abitino, Tyler L. Hayes, and Christopher Kanan (2018). “Measuring Catastrophic Forgetting in Neural Networks”. In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pp. 3390–3398 (cit. on p. 106).

Kendall, Alex and Yarin Gal (2017). “What uncertainties do we need in bayesian deep learning for computer vision?” In: *Advances in neural information processing systems*, pp. 5574–5584 (cit. on p. 83).

Ker, Justin, Lipo Wang, Jai Rao, and Tchoyoson Lim (2018). “Deep Learning Applications in Medical Image Analysis”. In: *IEEE Access* 6, pp. 9375–9389 (cit. on p. 89).

Khan, A., B. Rinner, and A. Cavallaro (2018). “Cooperative Robots to Observe Moving Targets: Review”. In: *IEEE Transactions on Cybernetics* 48.1, pp. 187–198 (cit. on p. 79).

Khan, Arbaaz and Martial Hebert (2018). “Learning safe recovery trajectories with deep neural networks for unmanned aerial vehicles”. In: *2018 IEEE Aerospace Conference*. IEEE, pp. 1–9 (cit. on p. 169).

Khoury, Marc and Dylan Hadfield-Menell (2018). “On the geometry of adversarial examples”. In: *arXiv preprint arXiv:1811.00525* (cit. on p. 161).

– (2019). “Adversarial Training with Voronoi Constraints”. In: *arXiv preprint arXiv:1905.01019* (cit. on p. 161).

Kietzmann, Tim Christian, Patrick McClure, and Nikolaus Kriegeskorte (2018). “Deep neural networks in computational neuroscience”. In: *bioRxiv*, p. 133504 (cit. on p. 36).

Kim, Jin-Hwa, Devi Parikh, Dhruv Batra, Byoung-Tak Zhang, and Yuandong Tian (2017). "Codraw: Visual dialog for collaborative drawing". In: *arXiv preprint arXiv:1712.05558* (cit. on p. 86).

Kindermans, Pieter-Jan, Kristof Schütt, Klaus-Robert Müller, and Sven Dähne (2016). "Investigating the influence of noise and distractors on the interpretation of neural networks". In: *arXiv preprint usesarXiv:1611.07270* (cit. on p. 131).

Kindermans, Pieter-Jan, Sara Hooker, Julius Adebayo, et al. (2017). "The (Un) reliability of saliency methods". In: *arXiv preprint arXiv:1711.00867* (cit. on p. 131).

Kirkpatrick, James, Razvan Pascanu, Neil Rabinowitz, et al. (2017). "Overcoming catastrophic forgetting in neural networks". In: *Proceedings of the national academy of sciences*, p. 201611835 (cit. on pp. 105, 106).

Knill, David C and Alexandre Pouget (2004). "The Bayesian brain: the role of uncertainty in neural coding and computation". In: *TRENDS in Neurosciences* 27.12, pp. 712–719 (cit. on p. 36).

Koffka, Kurt (2013). *Principles of Gestalt psychology*. Routledge (cit. on p. 41).

Kojima, Noriyuki and Jia Deng (2019). "To Learn or Not to Learn: Analyzing the Role of Learning for Navigation in Virtual Environments". In: *arXiv preprint arXiv:1907.11770* (cit. on p. 84).

Kokkinos, Iasonas (2017). "Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6129–6138 (cit. on pp. 21, 109).

Kolter, J Zico and Eric Wong (2017). "Provable defenses against adversarial examples via the convex outer adversarial polytope". In: *arXiv preprint arXiv:1711.00851* (cit. on p. 167).

Kounev, Samuel, Peter Lewis, Kirstie L Bellman, et al. (2017). "The notion of self-aware computing". In: *Self-Aware Computing Systems*. Springer, pp. 3–16 (cit. on p. 57).

Kovashka, Adriana, Olga Russakovsky, Li Fei-Fei, Kristen Grauman, et al. (2016). "Crowdsourcing in computer vision". In: *Foundations and Trends® in Computer Graphics and Vision* 10.3, pp. 177–243 (cit. on pp. 146, 147, 171).

Kragic, Danica (2018). *From active perception to deep learning* (cit. on p. 62).

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton (2012). "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems*, pp. 1097–1105 (cit. on pp. 1, 88).

Krotkov, Eric and Ruzena Bajcsy (1993). "Active vision for reliable ranging: Cooperating focus, stereo, and vergence". In: *International Journal of computer vision* 11.2, pp. 187–203 (cit. on p. 80).

Kubilius, Jonas, Martin Schrimpf, Aran Nayebi, et al. (2018). "CORnet: Modeling the Neural Mechanisms of Core Object Recognition". In: *bioRxiv* (cit. on p. 16).

Kumar, Arvind, Stefan Rotter, and Ad Aertsen (2010). "Spiking activity propagation in neuronal networks: reconciling different perspectives on neural coding". In: *Nature reviews neuroscience* 11.9, p. 615 (cit. on p. 36).

Kumaran, Dharshan, Demis Hassabis, and James L McClelland (2016). "What learning systems do intelligent agents need? Complementary learning systems theory updated". In: *Trends in cognitive sciences* 20.7, pp. 512–534 (cit. on p. 105).

Kurakin, Alexey, Ian J Goodfellow, and Samy Bengio (2017). "Adversarial Machine Learning at Scale". In: *International Conference on Learning Representations* (cit. on p. 158).

Lake, Brenden M., Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman (2017). "Building machines that learn and think like people". In: *Behavioral and Brain Sciences* 40, e253 (cit. on pp. 29, 184).

Lakkaraju, Himabindu, Ece Kamar, Rich Caruana, and Eric Horvitz (2017a). "Identifying Unknown Unknowns in the Open World: Representations and Policies for Guided Exploration." In: *AAAI*. Vol. 1, p. 2 (cit. on p. 161).

Lakkaraju, Himabindu, Ece Kamar, Rich Caruana, and Jure Leskovec (2017b). "Interpretable & Explorable Approximations of Black Box Models". In: *arXiv preprint arXiv:1707.01154* (cit. on pp. 124, 130).

Lakshminarayanan, Balaji, Alexander Pritzel, and Charles Blundell (2017). “Simple and scalable predictive uncertainty estimation using deep ensembles”. In: *Advances in neural information processing systems*, pp. 6402–6413 (cit. on p. 83).

Lala, Sayeri, Maha Shady, Anastasiya Belyaeva, and Molei Liu (2018). “Evaluation of mode collapse in generative adversarial networks”. In: *High Performance Extreme Computing, IEEE* (cit. on p. 95).

Lampert, Christoph H, Hannes Nickisch, and Stefan Harmeling (2009). “Learning to detect unseen object classes by between-class attribute transfer”. In: *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)* (cit. on p. 92).

– (2014). “Attribute-Based Classification for Zero-Shot Visual Object Categorization.” In: *IEEE Trans. on Pattern Analysis and Machine Intelligence* 36.3, pp. 453–465 (cit. on p. 92).

Lan, Xu, Hangxiao Wang, Shaogang Gong, and Xiatian Zhu (2017). “Deep Reinforcement Learning Attention Selection For Person Re-Identification.” In: *BMVC* (cit. on p. 80).

Lanctot, Marc, Vinicius Zambaldi, Audrunas Gruslys, et al. (2017). “A unified game-theoretic approach to multiagent reinforcement learning”. In: *Advances in Neural Information Processing Systems*, pp. 4190–4203 (cit. on p. 112).

Laporte, Catherine and Tal Arbel (2006). “Efficient Discriminant Viewpoint Selection for Active Bayesian Recognition”. In: *International Journal of Computer Vision* 68.3, pp. 267–287 (cit. on pp. 68, 77).

LaValle, Steven M (2006). *Planning algorithms*. Cambridge university press (cit. on p. 71).

Laversanne-Finot, Adrien, Alexandre Pere, and Pierre-Yves Oudeyer (2018). “Curiosity Driven Exploration of Learned Disentangled Goal Spaces”. In: *Conference on Robot Learning*, pp. 487–504 (cit. on pp. 91, 114).

Le Barz, Cédric, Nicolas Thome, Matthieu Cord, Stéphane Herbin, and Martial Sanfourche (Sept. 2014). “Global Robot Ego-localization Combining Image Retrieval and HMM-based Filtering”. In: *6th Workshop on Planning, Perception and Navigation for Intelligent Vehicles*. Chicago, United States, 6 p. (Cit. on p. 99).

Le Barz, Cedric, Nicolas Thome, Matthieu Cord, Stéphane Herbin, and Martial Sanfourche (2015a). “Absolute geo-localization thanks to Hidden Markov Model and exemplar-based metric learning”. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops, Boston, MA, USA, June 7-12, 2015*, pp. 9–17 (cit. on p. 99).

– (2015b). “Exemplar based metric learning for robust visual localization”. In: *2015 IEEE International Conference on Image Processing, ICIP 2015, Quebec City, QC, Canada, September 27-30, 2015*, pp. 4342–4346 (cit. on p. 99).

Le Barz, Cédric (2015). “Navigation visuelle pour les missions autonomes des petits drones”. PhD thesis. niversité Pierre et Marie Curie - Paris VI (cit. on pp. 97, 98, 233).

Le Cacheux, Yannick, Hervé Le Borgne, and Michel Crucianu (2019a). “From Classical to Generalized Zero-Shot Learning: A Simple Adaptation Process”. In: *International Conference on Multimedia Modeling*. Springer, pp. 465–477 (cit. on p. 92).

Le Cacheux, Yannick, Herve Le Borgne, and Michel Crucianu (2019b). “Modeling Inter and Intra-Class Relations in the Triplet Loss for Zero-Shot Learning”. In: *The IEEE International Conference on Computer Vision (ICCV)* (cit. on p. 92).

Leang, Isabelle, Stéphane Herbin, Benoît Girard, and Jacques Droulez (2015). “Robust Fusion of Trackers Using Online Drift Prediction”. In: *Advanced Concepts for Intelligent Vision Systems - 16th International Conference, ACIVS 2015, Catania, Italy, October 26-29, 2015, Proceedings*, pp. 229–240 (cit. on pp. 169, 233).

Leang, Isabelle, Stéphane Herbin, Benoît Girard, and Jacques Droulez (2018). “On-line fusion of trackers for single-object tracking”. In: *Pattern Recognition* 74, pp. 459–473 (cit. on pp. 169, 233).

Lechat, Alexis, Stéphane Herbin, and Frédéric Jurie (2019). “Adaptation du problème de questions-réponses visuelles à un contexte d’apprentissage continu”. In: *actes du GRETSI, Lille* (cit. on p. 107).

LeCun, Yann, Bernhard Boser, John S Denker, et al. (1989). “Backpropagation applied to handwritten zip code recognition”. In: *Neural computation* 1.4, pp. 541–551 (cit. on p. 88).

Lee, Kibok, Kimin Lee, Kyle Min, et al. (2018a). “Hierarchical Novelty Detection for Visual Object Recognition”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1034–1042 (cit. on p. 165).

Lee, Kimin, Kibok Lee, Honglak Lee, and Jinwoo Shin (2018b). “A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks”. In: *NIPS* (cit. on p. 167).

Lee, Sang-Woo, Yu-Jung Heo, and Byoung-Tak Zhang (2018c). “Answerer in questioner’s mind: information theoretic approach to goal-oriented visual dialog”. In: *Advances in Neural Information Processing Systems*, pp. 2579–2589 (cit. on p. 86).

Leibe, Bastian, Ales Leonardis, and Bernt Schiele (2004). “Combined object categorization and segmentation with an implicit shape model”. In: *Workshop on statistical learning in computer vision, ECCV*. Vol. 2. 5, p. 7 (cit. on p. 97).

Leotta, M.J. and J.L. Mundy (2009). “Predicting high resolution image edges with a generic, adaptive, 3-D vehicle model”. In: *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 1311–1318 (cit. on p. 79).

Lesort, Timothée, Hugo Caselles-Dupré, Michael Garcia-Ortiz, Andrei Stoian, and David Filliat (2018). “Generative Models from the perspective of Continual Learning”. In: *arXiv preprint arXiv:1812.09111* (cit. on pp. 94, 105).

Leung, Valerie and Stéphane Herbin (2011). “Flexible tracklet association for complex scenarios using a Markov Logic Network”. In: *IEEE International Conference on Computer Vision Workshops, ICCV 2011 Workshops, Barcelona, Spain, November 6-13, 2011*, pp. 1870–1875 (cit. on pp. 101, 102, 233).

Lewis, Mike, Denis Yarats, Yann Dauphin, Devi Parikh, and Dhruv Batra (2017). “Deal or No Deal? End-to-End Learning of Negotiation Dialogues”. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2443–2453 (cit. on p. 86).

Lewis, Peter R (2017). “Self-aware computing systems: From psychology to engineering”. In: *Proceedings of the Conference on Design, Automation & Test in Europe*. European Design and Automation Association, pp. 1044–1049 (cit. on p. 57).

Lewis, Peter R, Marco Platzner, Bernhard Rinner, Jim Tørresen, and Xin Yao (2016). *Self-Aware Computing Systems*. Springer (cit. on p. 57).

Li, Hongyang, Yu Liu, Wanli Ouyang, and Xiaogang Wang (2019). “Zoom out-and-in network with map attention decision for region proposal and object detection”. In: *International Journal of Computer Vision* 127.3, pp. 225–238 (cit. on p. 81).

Li, Jun, José M Bioucas-Dias, and Antonio Plaza (2011). “Hyperspectral image segmentation using a new Bayesian approach with active learning”. In: *IEEE Transactions on Geoscience and Remote Sensing* 49.10, pp. 3947–3960 (cit. on p. 103).

Li, Qing, Qingyi Tao, Shafiq Joty, Jianfei Cai, and Jiebo Luo (2018). “VQA-E: Explaining, Elaborating, and Enhancing Your Answers for Visual Questions”. In: *arXiv preprint arXiv:1803.07464* (cit. on p. 126).

Li, Xin and Fuxin Li (2017). “Adversarial Examples Detection in Deep Networks with Convolutional Filter Statistics.” In: *ICCV*, pp. 5775–5783 (cit. on p. 166).

Li, Xirong, Cees GM Snoek, Marcel Worring, Dennis Koelma, Arnold WM Smeulders, et al. (2013). “Bootstrapping visual categorization with relevant negatives”. In: *IEEE Transactions on Multimedia* 15.4, pp. 933–945 (cit. on p. 147).

Li, Yujia, Kevin Swersky, and Rich Zemel (2015). “Generative moment matching networks”. In: *International Conference on Machine Learning*, pp. 1718–1727 (cit. on p. 95).

Li, Zhizhong and Derek Hoiem (2017). “Learning without forgetting”. In: *IEEE transactions on pattern analysis and machine intelligence* 40.12, pp. 2935–2947 (cit. on p. 105).

Liang, Shiyu, Yixuan Li, and R. Srikant (2018). “Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks”. In: *International Conference on Learning Representations* (cit. on p. 165).

Lin, Chin-Yew (2004). “ROUGE: a Package for Automatic Evaluation of Summaries”. In: *Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004* (cit. on p. 126).

Lin, Lan, Huan Luo, Renjie Huang, and Mao Ye (2019). "Recurrent models of visual co-attention for person re-identification". In: *IEEE Access* 7, pp. 8865–8875 (cit. on p. 80).

Lin, Tsung-Yi, Piotr Dollár, Ross Girshick, et al. (2017). "Feature pyramid networks for object detection". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125 (cit. on p. 88).

Lindsay, G. (2018). *Deep Convolutional Neural Networks as Models of the Visual System: Q&A*. Neurdiness (Blog) (cit. on p. 33).

Linsker, Ralph (1986a). "From basic network principles to neural architecture: Emergence of orientation columns". In: *Proceedings of the National Academy of Sciences* 83.22, pp. 8779–8783 (cit. on p. 121).

– (1986b). "From basic network principles to neural architecture: Emergence of orientation-selective cells". In: *Proceedings of the National Academy of Sciences* 83.21, pp. 8390–8394 (cit. on p. 121).

Lipton, Zachary C (2016). "The mythos of model interpretability". In: *arXiv preprint arXiv:1606.03490* (cit. on pp. 118, 124).

Liu, Changliu, Tomer Arnon, Christopher Lazarus, Clark Barrett, and Mykel J Kochenderfer (2019). "Algorithms for verifying deep neural networks". In: *arXiv preprint arXiv:1903.06758* (cit. on p. 163).

Liu, Chenxi, Junhua Mao, Fei Sha, and Alan L Yuille (2017). "Attention Correctness in Neural Image Captioning." In: *AAAI*, pp. 4176–4182 (cit. on p. 127).

Liu, Chunming, Xin Xu, and Dewen Hu (2014). "Multiobjective reinforcement learning: A comprehensive overview". In: *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 45.3, pp. 385–398 (cit. on p. 71).

Liu, Junbin, Sridha Sridharan, and Clinton Fookes (2016a). "Recent advances in camera planning for large area surveillance: A comprehensive review". In: *ACM Computing Surveys (CSUR)* 49.1, p. 6 (cit. on p. 79).

Liu, Li, Wanli Ouyang, Xiaogang Wang, et al. (2018). "Deep learning for generic object detection: A survey". In: *arXiv preprint arXiv:1809.02165* (cit. on p. 88).

Liu, Wei, Dragomir Anguelov, Dumitru Erhan, et al. (2016b). "Ssd: Single shot multibox detector". In: *European conference on computer vision*. Springer, pp. 21–37 (cit. on p. 88).

Liu, Xiaodong, Jianfeng Gao, Xiaodong He, et al. (2015). "Representation Learning Using Multi-Task Deep Neural Networks for Semantic Classification and Information Retrieval". In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 912–921 (cit. on p. 91).

Liu, Yanpei, Xinyun Chen, Chang Liu, and Dawn Song (2016c). "Delving into transferable adversarial examples and black-box attacks". In: *ICLR* (cit. on p. 157).

Locatelli, Roberta and Keith A Wilson (2017). "Introduction: Perception without representation". In: *Topoi* 36.2, pp. 197–212 (cit. on p. 46).

Locatello, Francesco, Stefan Bauer, Mario Lucic, et al. (2019). "Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations". In: *International Conference on Machine Learning*, pp. 4114–4124 (cit. on p. 91).

Long, Mingsheng, Yue Cao, Jianmin Wang, and Michael I Jordan (2015). "Learning transferable features with deep adaptation networks". In: *Proceedings of the 32nd International Conference on International Conference on Machine Learning-Volume 37*. JMLR. org, pp. 97–105 (cit. on p. 89).

Lopez-Paz, David et al. (2017). "Gradient episodic memory for continual learning". In: *Advances in Neural Information Processing Systems*, pp. 6467–6476 (cit. on p. 106).

Lotter, William, Gabriel Kreiman, and David Cox (2017). "Deep predictive coding networks for video prediction and unsupervised learning". In: *ICLR* (cit. on p. 56).

Lowe, Ryan, Yi Wu, Aviv Tamar, et al. (2017). "Multi-agent actor-critic for mixed cooperative-competitive environments". In: *Advances in Neural Information Processing Systems*, pp. 6379–6390 (cit. on p. 113).

Lu, Jiajun, Hussein Sibai, Evan Fabry, and David Forsyth (2017a). "No need to worry about adversarial examples in object detection in autonomous vehicles". In: *CVPRW* (cit. on p. 159).

Lu, Jiajun, Theerasit Issaranon, and David A Forsyth (2017b). "SafetyNet: Detecting and Rejecting Adversarial Examples Robustly." In: *ICCV*, pp. 446–454 (cit. on p. 166).

Lu, Jiajun, Hussein Sibai, Evan Fabry, and David Forsyth (2017c). "Standard detectors aren't (currently) fooled by physical adversarial stop signs". In: *arXiv preprint arXiv:1710.03337* (cit. on p. 159).

Lu, Jiasen, Jianwei Yang, Dhruv Batra, and Devi Parikh (2016). "Hierarchical question-image co-attention for visual question answering". In: *Advances In Neural Information Processing Systems*, pp. 289–297 (cit. on p. 80).

Lu, Jiasen, Anitha Kannan, Jianwei Yang, Devi Parikh, and Dhruv Batra (2017d). "Best of both worlds: Transferring knowledge from discriminative learning to a generative visual dialog model". In: *Advances in Neural Information Processing Systems*, pp. 314–324 (cit. on p. 86).

Lucas, Alice, Michael Iliadis, Rafael Molina, and Aggelos K Katsaggelos (2018). "Using deep neural networks for inverse problems in imaging: beyond analytical methods". In: *IEEE Signal Processing Magazine* 35.1, pp. 20–36 (cit. on p. 101).

Lundberg, Scott M and Su-In Lee (2017). "A unified approach to interpreting model predictions". In: *Advances in Neural Information Processing Systems*, pp. 4765–4774 (cit. on p. 124).

Lunz, Sebastian, Ozan Öktem, and Carola-Bibiane Schönlieb (2018). "Adversarial regularizers in inverse problems". In: *Advances in Neural Information Processing Systems*, pp. 8507–8516 (cit. on p. 101).

Luo, Wenhan, Peng Sun, Fangwei Zhong, et al. (2018). "End-to-end Active Object Tracking via Reinforcement Learning". In: *International Conference on Machine Learning*, pp. 3292–3301 (cit. on p. 80).

– (2019). "End-to-end Active Object Tracking and Its Real-world Deployment via Reinforcement Learning". In: *IEEE transactions on pattern analysis and machine intelligence* (cit. on p. 80).

Lupyan, Gary (2015). "Cognitive penetrability of perception in the age of prediction: Predictive systems are penetrable systems". In: *Review of philosophy and psychology* 6.4, pp. 547–569 (cit. on p. 39).

Lyons, Jack (2017). "Epistemological Problems of Perception". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Spring 2017. Metaphysics Research Lab, Stanford University (cit. on p. 45).

Ma, Lei, Felix Juefei-Xu, Fuyuan Zhang, et al. (2018a). "DeepGauge: multi-granularity testing criteria for deep learning systems". In: *Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering*. ACM, pp. 120–131 (cit. on p. 162).

Ma, Xingjun, Bo Li, Yisen Wang, et al. (2018b). "Characterizing Adversarial Subspaces Using Local Intrinsic Dimensionality". In: *International Conference on Learning Representations* (cit. on p. 167).

Ma, Yukun, Haiyun Peng, and Erik Cambria (2018c). "Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM". In: *Thirty-Second AAAI Conference on Artificial Intelligence* (cit. on p. 80).

Mack, Arien (2003). "Inattentional blindness: Looking without seeing". In: *Current Directions in Psychological Science* 12.5, pp. 180–184 (cit. on p. 44).

Mack, Arien, Irvin Rock, et al. (1998). *Inattentional blindness*. MIT press (cit. on pp. 24, 44).

Madary, Michael (2016). *Visual Phenomenology*. MIT Press (cit. on p. 47).

Madry, Aleksander, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu (2018). "Towards Deep Learning Models Resistant to Adversarial Attacks". In: *International Conference on Learning Representations* (cit. on p. 158).

Mahendran, Aravindh and Andrea Vedaldi (2015). "Understanding deep image representations by inverting them". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5188–5196 (cit. on p. 123).

Mahloujifar, Saeed, Xiao Zhang, Mohammad Mahmoody, and David Evans (2019). "Empirically Measuring Concentration: Fundamental Limits on Intrinsic Robustness". In: *arXiv preprint arXiv:1905.12202* (cit. on p. 161).

Mairal, Julien (2015). "Incremental majorization-minimization optimization with application to large-scale machine learning". In: *SIAM Journal on Optimization* 25.2, pp. 829–855 (cit. on p. 113).

Makhzani, Alireza, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey (2016). "Adversarial autoencoders". In: *ICLR* (cit. on p. 95).

Malinin, Andrey and Mark Gales (2018). "Predictive uncertainty estimation via prior networks". In: *Advances in Neural Information Processing Systems*, pp. 7047–7058 (cit. on p. 83).

Malmir, Mohsen, Karan Sikka, Deborah Forster, et al. (2017). "Deep active object recognition by joint label and action prediction". In: *Computer Vision and Image Understanding* 156, pp. 128–137 (cit. on p. 84).

Mandelbaum, Amit and Daphna Weinshall (2017). "Distance-based Confidence Score for Neural Network Classifiers". In: *arXiv preprint arXiv:1709.09844* (cit. on p. 165).

Maninis, Kevis-Kokitsi, Ilija Radosavovic, and Iasonas Kokkinos (2019). "Attentive Single-Tasking of Multiple Tasks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1851–1860 (cit. on p. 109).

Marblestone, Adam H, Greg Wayne, and Konrad P Kording (2016). "Toward an integration of deep learning and neuroscience". In: *Frontiers in Computational Neuroscience* 10, p. 94 (cit. on p. 55).

Marcus, Gary (2018). "Deep learning: A critical appraisal". In: *arXiv preprint arXiv:1801.00631* (cit. on p. 108).

Markou, Markos and Sameer Singh (2003). "Novelty detection: a review—part 1: statistical approaches". In: *Signal processing* 83.12, pp. 2481–2497 (cit. on p. 164).

Marr, David (1982). *Vision: A computational investigation into the human representation and processing of visual information*. Cambridge, Massachusetts: MIT Press (cit. on p. 143).

Mascharka, David, Philip Tran, Ryan Soklaski, and Arjun Majumdar (2018). "Transparency by Design: Closing the Gap Between Performance and Interpretability in Visual Reasoning". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4942–4950 (cit. on p. 128).

Masland, Richard H (2012). "The neuronal organization of the retina". In: *Neuron* 76.2, pp. 266–280 (cit. on p. 32).

Mathieu, Emile, Tom Rainforth, N Siddharth, and Yee Whye Teh (2019). "Disentangling Disentanglement in Variational Autoencoders". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. Long Beach, California, USA: PMLR, pp. 4402–4412 (cit. on p. 91).

Matthen, Mohan (2015). *The Oxford handbook of philosophy of perception*. Oxford Handbooks (cit. on pp. 29, 45).

Mavrinac, Aaron and Xiang Chen (2013). "Modeling coverage in camera networks: A survey". In: *International journal of computer vision* 101.1, pp. 205–226 (cit. on p. 79).

McClelland, James L, David E Rumelhart, et al., eds. (1987). *Parallel distributed processing*. MIT Press (cit. on p. 96).

McClelland, James L., Bruce L. McNaughton, and Randall C. OReilly (1994). "Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory." In: *Psychological review* 102 3, pp. 419–457 (cit. on p. 105).

McCloskey, Michael and Neal J. Cohen (1989). "Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem". In: ed. by Gordon H. Bower. Vol. 24. Psychology of Learning and Motivation. Academic Press, pp. 109 –165 (cit. on p. 105).

McCulloch, Warren S and Walter Pitts (1943). "A logical calculus of the ideas immanent in nervous activity". In: *The bulletin of mathematical biophysics* 5.4, pp. 115–133 (cit. on p. 30).

McGuinness, Kevin and Noel E O'connor (2010). "A comparative evaluation of interactive segmentation algorithms". In: *Pattern Recognition* 43.2, pp. 434–444 (cit. on p. 171).

McIntosh, Lane, Niru Maheswaranathan, Aran Nayebi, Surya Ganguli, and Stephen Baccus (2016). "Deep learning models of the retinal response to natural scenes". In: *Advances in neural information processing systems*, pp. 1369–1377 (cit. on p. 32).

Medathati, NV Kartheek, Heiko Neumann, Guillaume S Masson, and Pierre Kornprobst (2016). "Bio-inspired computer vision: Towards a synergistic approach of artificial and biological vision". In: *Computer Vision and Image Understanding* 150, pp. 1–30 (cit. on p. 55).

Meincke, Anne Sophie (2018). "Autopoiesis, biological autonomy and the process view of life". In: *European Journal for Philosophy of Science* 9.1, p. 5 (cit. on p. 23).

Meinhardt, Tim, Michael Moller, Caner Hazirbas, and Daniel Cremers (2017). "Learning proximal operators: Using denoising networks for regularizing inverse imaging problems". In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1781–1790 (cit. on p. 101).

Meister, Markus and Michael J Berry (1999). "The neural code of the retina". In: *Neuron* 22.3, pp. 435–450 (cit. on p. 32).

Melman, Shachaf, Yael Moses, Gérard Medioni, and Yinghao Cai (2018). "The Multi-strand Graph for a PTZ Tracker". In: *Journal of Mathematical Imaging and Vision* 60.4, pp. 594–608 (cit. on p. 78).

Meng, Dongyu and Hao Chen (2017). "Magnet: a two-pronged defense against adversarial examples". In: *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. ACM, pp. 135–147 (cit. on p. 167).

Menzies, Tim and Charles Pecheur (2005). "Verification and validation and artificial intelligence". In: *Advances in computers* 65, pp. 153–201 (cit. on p. 163).

Merleau-Ponty, Maurice (1945). *Phénoménologie de la perception*. Gallimard (cit. on p. 47).

Metzen, Jan Hendrik, Tim Genewein, Volker Fischer, and Bastian Bischoff (2017). "On Detecting Adversarial Perturbations". In: *International Conference on Learning Representations* (cit. on p. 166).

Miller, Tim (2017). "Explanation in artificial intelligence: insights from the social sciences". In: *arXiv preprint arXiv:1706.07269* (cit. on p. 134).

Min, Huaqing, Chang'an Yi, Ronghua Luo, Jinhui Zhu, and Sheng Bi (2016). "Affordance research in developmental robotics: A survey". In: *IEEE Transactions on Cognitive and Developmental Systems* 8.4, pp. 237–255 (cit. on p. 91).

Mirza, Mehdi and Simon Osindero (2014). "Conditional generative adversarial nets". In: *arXiv preprint arXiv:1411.1784* (cit. on p. 94).

Mitchell, T., W. Cohen, E. Hruschka, et al. (Apr. 2018). "Never-ending Learning". In: *Commun. ACM* 61.5, pp. 103–115 (cit. on pp. 104, 115).

Mittal, Anurag and Larry S Davis (2008). "A general method for sensor planning in multi-sensor systems: Extension to random occlusion". In: *International Journal of Computer Vision* 76.1, pp. 31–52 (cit. on p. 79).

Mnih, Volodymyr, Nicolas Heess, Alex Graves, et al. (2014). "Recurrent models of visual attention". In: *Advances in neural information processing systems*, pp. 2204–2212 (cit. on p. 81).

Mogadala, Aditya, Marimuthu Kalimuthu, and Dietrich Klakow (2019). "Trends in Integration of Vision and Language Research: A Survey of Tasks, Datasets, and Methods". In: *arXiv preprint arXiv:1907.09358* (cit. on pp. 86, 136).

Mohseni, Sina, Niloofar Zarei, and Eric D. Ragan (2018). "A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems". In: *arXiv preprint arXiv:1811.11839* (cit. on pp. 130, 132).

Mole, Christopher (2017). "Attention". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Fall 2017. Metaphysics Research Lab, Stanford University (cit. on p. 42).

Montavon, Grégoire, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller (2017). "Explaining nonlinear classification decisions with deep taylor decomposition". In: *Pattern Recognition* 65, pp. 211–222 (cit. on p. 125).

Montavon, Grégoire, Wojciech Samek, and Klaus Robert Müller (2018). "Methods for interpreting and understanding deep neural networks". In: *Digital Signal Processing: A Review Journal* 73, pp. 1–15 (cit. on pp. 125, 131).

Moosavi-Dezfooli, Seyed-Mohsen, Alhussein Fawzi, and Pascal Frossard (2016). "Deepfool: a simple and accurate method to fool deep neural networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2574–2582 (cit. on p. 156).

Moosavi-Dezfooli, Seyed-Mohsen, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard (2017). "Universal Adversarial Perturbations". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1765–1773 (cit. on p. 156).

Morcos, Ari S, David GT Barrett, Neil C Rabinowitz, and Matthew Botvinick (2018). "On the importance of single directions for generalization". In: *arXiv preprint arXiv:1803.06959* (cit. on p. 135).

Mueller, Matthias, Vincent Casser, Jean Lahoud, Neil Smith, and Bernard Ghanem (2017). "UE4Sim: A photo-realistic simulator for computer vision applications". In: *arXiv preprint arXiv:1708.05869* (cit. on p. 153).

Müller, Matthias, Vincent Casser, Jean Lahoud, Neil Smith, and Bernard Ghanem (2018). "Sim4CV: A photo-realistic simulator for computer vision applications". In: *International Journal of Computer Vision* 126.9, pp. 902–919 (cit. on p. 93).

Mundy, Joseph L (2006). "Object recognition in the geometric era: A retrospective". In: *Toward category-level object recognition*. Springer, pp. 3–28 (cit. on pp. 41, 96).

Murez, Zak, Soheil Kolouri, David Kriegman, Ravi Ramamoorthi, and Kyungnam Kim (2018). "Image to image translation for domain adaptation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4500–4509 (cit. on p. 94).

Murphy, Kevin P (2012). *Machine learning: a probabilistic perspective*. MIT press (cit. on p. 87).

Nachum, Ofir, Shixiang Shane Gu, Honglak Lee, and Sergey Levine (2018). "Data-efficient hierarchical reinforcement learning". In: *Advances in Neural Information Processing Systems*, pp. 3303–3313 (cit. on p. 109).

Naghshvar, Mohammad and Tara Javidi (Dec. 2013). "Active sequential hypothesis testing". In: *The Annals of Statistics* 41.6, pp. 2703–2738 (cit. on p. 72).

Nair, Suraj and Chelsea Finn (2019). "Hierarchical Foresight: Self-Supervised Learning of Long-Horizon Tasks via Visual Subgoal Generation". In: *ArXiv* abs/1909.05829 (cit. on p. 109).

Nan, Feng, Joseph Wang, and Venkatesh Saligrama (2015). "Feature-Budgeted Random Forest". In: *International Conference on Machine Learning*, pp. 1983–1991 (cit. on p. 71).

Nanay, Bence (2015). "Perceptual representation/perceptual content". In: *Oxford Handbook for the Philosophy of Perception*, pp. 153–167 (cit. on p. 46).

Natarajan, Prabhu, Pradeep K Atrey, and Mohan Kankanhalli (2015). "Multi-camera coordination and control in surveillance systems: A survey". In: *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 11.4, p. 57 (cit. on p. 79).

Nguyen, Anh, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune (2016). "Synthesizing the preferred inputs for neurons in neural networks via deep generator networks". In: *Advances in Neural Information Processing Systems*, pp. 3387–3395 (cit. on p. 122).

Nguyen, Ngoc Quynh (2019). "Optimization landscape of deep neural networks". PhD thesis. Saarländische Universitäts (cit. on p. 108).

N'Guyen, Steve, Charles Thurat, and Benoît Girard (2014). "Saccade learning with concurrent cortical and subcortical basal ganglia loops". In: *Frontiers in computational neuroscience* 8, p. 48 (cit. on p. 13).

Nguyen, Tam V, Qi Zhao, and Shuicheng Yan (2018). "Attentive systems: A survey". In: *International Journal of Computer Vision* 126.1, pp. 86–110 (cit. on p. 43).

Nicolae, Maria-Irina, Mathieu Sinn, Minh Ngoc Tran, et al. (2018). "Adversarial Robustness Toolbox v0.3.0". In: *arXiv preprint arXiv:1807.01069* (cit. on p. 160).

Niu, Yulei, Hanwang Zhang, Manli Zhang, et al. (2019). "Recursive visual attention in visual dialog". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6679–6688 (cit. on p. 86).

Nobre, Kia (2014). *The Oxford handbook of attention*. Oxford University Press (cit. on p. 42).

Noë, Alva (2002). "Is the visual world a grand illusion?" In: *Journal of consciousness studies* 9.5-6, pp. 1–12 (cit. on p. 48).

Norman, Don (2009). *The design of future things*. Basic books (cit. on p. 172).

– (2013). *The design of everyday things: Revised and expanded edition*. Constellation (cit. on pp. 104, 135).
Northoff, Georg (2013). *Unlocking the brain: volume 2: consciousness*. Oxford University Press (cit. on p. 36).
Nushi, Besmira, Ece Kamar, Eric Horvitz, and Donald Kossmann (2017). "On Human Intellect and Machine Failures: Troubleshooting Integrative Machine Learning Systems." In: *AAAI*, pp. 1017–1025 (cit. on p. 132).
Nushi, Besmira, Ece Kamar, and Eric Horvitz (2018). "Towards Accountable AI: Hybrid Human-Machine Analyses for Characterizing System Failure". In: *HCOMP-18*. AAAI Press, Palo Alto, California USA, pp. 126–135 (cit. on p. 132).
Odena, Augustus and Ian Goodfellow (2018). "TensorFuzz: Debugging Neural Networks with Coverage-Guided Fuzzing". In: *arXiv preprint arXiv:1807.10875* (cit. on p. 162).
Oh, Junhyuk, Valliappa Chockalingam, Honglak Lee, et al. (2016). "Control of Memory, Active Perception, and Action in Minecraft". In: *International Conference on Machine Learning*, pp. 2790–2799 (cit. on p. 110).
Olah, Chris and Shan Carter (2016). "Attention and Augmented Recurrent Neural Networks". In: *Distill* (cit. on p. 81).
Olah, Chris, Alexander Mordvintsev, and Ludwig Schubert (2017). "Feature Visualization". In: *Distill* 2.11, e7 (cit. on p. 123).
Olah, Chris, Arvind Satyanarayan, Ian Johnson, et al. (2018). "The building blocks of interpretability". In: *Distill* 3.3, e10 (cit. on p. 135).
Oquab, Maxime, Léon Bottou, Ivan Laptev, and Josef Sivic (2015). "Is object localization for free?-weakly-supervised learning with convolutional neural networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 685–694 (cit. on p. 90).
O'Regan, J Kevin and Alva Noë (2001). "A sensorimotor account of vision and visual consciousness". In: *Behavioral and brain sciences* 24.5, pp. 939–973 (cit. on p. 19).
Osa, Takayuki, Joni Pajarinen, Gerhard Neumann, et al. (2018). "An algorithmic perspective on imitation learning". In: *Foundations and Trends® in Robotics* 7.1-2, pp. 1–179 (cit. on p. 71).
Osoba, Osonde A and William Welser IV (2017). *An intelligence in our image: The risks of bias and errors in artificial intelligence*. Rand Corporation (cit. on p. 139).
Oudeyer, Pierre-Yves (2018). "Computational theories of curiosity-driven learning". In: *arXiv preprint arXiv:1802.10546* (cit. on pp. 91, 114).
Oudeyer, Pierre-Yves and Frederic Kaplan (2009). "What is intrinsic motivation? A typology of computational approaches". In: *Frontiers in Neurorobotics* 1, p. 6 (cit. on p. 91).
Ovadia, Yaniv, Emily Fertig, Jie Ren, et al. (2019). "Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift". In: *arXiv preprint arXiv:1906.02530* (cit. on p. 83).
Paletta, Lucas and Axel Pinz (2000). "Active object recognition by view integration and reinforcement learning". In: *Robotics and Autonomous Systems* 31.1-2, pp. 71–86 (cit. on pp. 68, 76).
Pan, Sinno Jialin and Qiang Yang (2009). "A survey on transfer learning". In: *IEEE Transactions on knowledge and data engineering* 22.10, pp. 1345–1359 (cit. on pp. 26, 89).
Pan, Zhaoqing, Weijie Yu, Xiaokai Yi, et al. (2019). "Recent progress on generative adversarial networks (GANs): A survey". In: *IEEE Access* 7, pp. 36322–36333 (cit. on p. 94).
Panzeri, Stefano, Nicolas Brunel, Nikos K Logothetis, and Christoph Kayser (2010). "Sensory neural codes using multiplexed temporal scales". In: *Trends in neurosciences* 33.3, pp. 111–120 (cit. on p. 35).
Papandreou, George, Liang-Chieh Chen, Kevin P Murphy, and Alan L Yuille (2015). "Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation". In: *Proceedings of the IEEE international conference on computer vision*, pp. 1742–1750 (cit. on p. 90).
Papernot, Nicolas and Patrick McDaniel (2017). "Extending defensive distillation". In: *arXiv preprint arXiv:1705.05264* (cit. on p. 159).

Papernot, Nicolas, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami (2016a). "Distillation as a defense to adversarial perturbations against deep neural networks". In: *Security and Privacy (SP), 2016 IEEE Symposium on*. IEEE, pp. 582–597 (cit. on p. 159).

Papernot, Nicolas, Patrick McDaniel, Arunesh Sinha, and Michael Wellman (2016b). "Towards the science of security and privacy in machine learning". In: *arXiv preprint arXiv:1611.03814* (cit. on p. 151).

Papernot, Nicolas, Patrick McDaniel, and Ian Goodfellow (2016c). "Transferability in machine learning: from phenomena to black-box attacks using adversarial samples". In: *arXiv preprint arXiv:1605.07277* (cit. on p. 157).

Papernot, Nicolas, Patrick McDaniel, Ian Goodfellow, et al. (2017). "Practical black-box attacks against machine learning". In: *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*. ACM, pp. 506–519 (cit. on p. 156).

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu (2002). "BLEU: a method for automatic evaluation of machine translation". In: *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, pp. 311–318 (cit. on p. 126).

Parisi, German I, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter (2018). "Continual Lifelong Learning with Neural Networks: A Review". In: *arXiv preprint arXiv:1802.07569* (cit. on p. 105).

Park, Dong Huk, Lisa Anne Hendricks, Zeynep Akata, et al. (2016). "Attentive explanations: Justifying decisions and pointing to the evidence". In: *arXiv preprint arXiv:1612.04757* (cit. on p. 126).

Park, Dong Huk, Lisa Anne Hendricks, Zeynep Akata, et al. (2018). "Multimodal Explanations: Justifying Decisions and Pointing to the Evidence". In: *31st IEEE Conference on Computer Vision and Pattern Recognition* (cit. on p. 126).

Parthasarathy, Nikhil, Eleanor Batty, William Falcon, et al. (2017). "Neural networks for efficient bayesian decoding of natural images from retinal neurons". In: *Advances in Neural Information Processing Systems*, pp. 6434–6445 (cit. on p. 33).

Parvaneh, Amin, Ehsan Abbasnejad, Qi Wu, and Javen Shi (2019). "Show, Price and Negotiate: A Hierarchical Attention Recurrent Visual Negotiator". In: *arXiv preprint arXiv:1905.03721* (cit. on p. 86).

Patel, Vishal M, Raghuraman Gopalan, Ruonan Li, and Rama Chellappa (2015). "Visual domain adaptation: A survey of recent advances". In: *IEEE signal processing magazine* 32.3, pp. 53–69 (cit. on p. 26).

Pathak, Deepak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros (2016). "Context encoders: Feature learning by inpainting". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2536–2544 (cit. on p. 91).

Pathak, Deepak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell (2017). "Curiosity-Driven Exploration by Self-Supervised Prediction". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops* (cit. on p. 91).

Patten, Timothy (2016). "Active object classification from 3D range data with mobile robots". PhD thesis. University of Sydney (cit. on p. 81).

Pearl, Judea (1988). "Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference". In: (cit. on p. 96).

Peebles, David and Richard P Cooper (2015). "Thirty years after Marr's vision: levels of analysis in cognitive science". In: *Topics in cognitive science* 7.2, pp. 187–190 (cit. on p. 143).

Pei, Kexin, Yinzhi Cao, Junfeng Yang, and Suman Jana (2017). "Deepxplore: Automated whitebox testing of deep learning systems". In: *Proceedings of the 26th Symposium on Operating Systems Principles*. ACM, pp. 1–18 (cit. on p. 162).

Perez, Ethan, Harm De Vries, Florian Strub, Vincent Dumoulin, and Aaron Courville (2017). "Learning Visual Reasoning Without Strong Priors". In: *ICML 2017's Machine Learning in Speech and Language Processing Workshop* (cit. on p. 128).

Petitot, Jean, Francisco J. Varela, Bernard Pachoud, and Jean-Michel Roy, eds. (1999). *Naturalizing phenomenology: Issues in contemporary phenomenology and cognitive science.* Stanford University Press (cit. on p. 49).

Pfülb, B and A Gepperth (2019). "A comprehensive, application-oriented study of catastrophic forgetting in DNNS". In: *arXiv preprint arXiv:1905.08101* (cit. on p. 106).

Piciarelli, Claudio, Lukas Esterle, Asif Khan, Bernhard Rinner, and Gian Luca Foresti (2015). "Dynamic reconfiguration in camera networks: A short survey". In: *IEEE Transactions on Circuits and Systems for Video Technology* 26.5, pp. 965–977 (cit. on p. 79).

Pillow, Jonathan W, Jonathon Shlens, Liam Paninski, et al. (2008). "Spatio-temporal correlations and visual signalling in a complete neuronal population". In: *Nature* 454.7207, p. 995 (cit. on p. 35).

Pimentel, Marco AF, David A Clifton, Lei Clifton, and Lionel Tarassenko (2014). "A review of novelty detection". In: *Signal Processing* 99, pp. 215–249 (cit. on p. 164).

Poggio, Tomaso, Hrushikesh Mhaskar, Lorenzo Rosasco, Brando Miranda, and Qianli Liao (2017). "Why and when can deep-but not shallow-networks avoid the curse of dimensionality: a review". In: *International Journal of Automation and Computing* 14.5, pp. 503–519 (cit. on p. 108).

Pomerantz, James R. (2003). "Perception: Overview". In: *Encyclopedia of Cognitive Science*. Ed. by L. Nadel. Nature Publishing Group (cit. on p. 40).

Ponce, Jean, Tamara L Berg, Mark Everingham, et al. (2006). "Dataset issues in object recognition". In: *Toward category-level object recognition*. Springer, pp. 29–48 (cit. on p. 152).

Poole, David L and Alan K Mackworth (2017). *Artificial Intelligence: foundations of computational agents*. second. Cambridge University Press (cit. on pp. 17, 18).

Potthast, Christian, Andreas Breitenmoser, Fei Sha, and Gaurav S Sukhatme (2016). "Active multi-view object recognition: A unifying view on online feature selection and view planning". In: *Robotics and Autonomous Systems* 84, pp. 31–47 (cit. on p. 68).

Preece, Alun, Dan Harborne, Dave Braines, Richard Tomsett, and Supriyo Chakraborty (2018). "Stakeholders in Explainable AI". In: *arXiv preprint arXiv:1810.00184* (cit. on p. 118).

Pu, Shi, Yibing Song, Chao Ma, Honggang Zhang, and Ming-Hsuan Yang (2018). "Deep attentive tracking via reciprocative learning". In: *Advances in Neural Information Processing Systems*, pp. 1931–1941 (cit. on p. 80).

Pylyshyn, Zenon (1999). "Is vision continuous with cognition?: The case for cognitive impenetrability of visual perception". In: *Behavioral and brain sciences* 22.3, pp. 341–365 (cit. on pp. 17, 41).

Péré, Alexandre, Sébastien Forestier, Olivier Sigaud, and Pierre-Yves Oudeyer (2018). "Unsupervised Learning of Goal Spaces for Intrinsically Motivated Goal Exploration". In: *International Conference on Learning Representations* (cit. on pp. 91, 114).

Quo, Yanlin, S. Hsu, H.S. Sawhney, R. Kumar, and Ying Shan (2007). "Robust Object Matching for Persistent Tracking with Heterogeneous Features". In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 29.5, pp. 824–839 (cit. on p. 79).

Rahwan, Iyad, Manuel Cebrian, Nick Obradovich, et al. (2019). "Machine behaviour". In: *Nature* 568.7753, pp. 477–486 (cit. on p. 151).

Raissi, Maziar, Paris Perdikaris, and George E Karniadakis (2019). "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations". In: *Journal of Computational Physics* 378, pp. 686–707 (cit. on p. 102).

Rajalingham, Rishi, Elias B Issa, Pouya Bashivan, et al. (2018). "Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks". In: *Journal of Neuroscience* 38.33, pp. 7255–7269 (cit. on p. 16).

Rajani, Nazneen Fatema and Raymond J. Mooney (2017). "Using Explanations to Improve Ensembling of Visual Question Answering Systems". In: *IJCAI* (cit. on p. 125).

Rajasegaran, Jathushan, Munawar Hayat, Salman Khan, Fahad Shahbaz Khan, and Ling Shao (2019). "Random Path Selection for Incremental Learning". In: *arXiv preprint arXiv:1906.01120* (cit. on p. 106).

Ramirez-Loaiza, Maria E, Manali Sharma, Geet Kumar, and Mustafa Bilgic (2017). “Active learning: an empirical study of common baselines”. In: *Data mining and knowledge discovery* 31.2, pp. 287–313 (cit. on p. 103).

Ranjan, Anurag, Joel Janai, Andreas Geiger, and Michael J Black (2019). “Attacking optical flow”. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2404–2413 (cit. on p. 156).

Rao, Rajesh PN and Dana H Ballard (1999). “Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects”. In: *Nature neuroscience* 2.1, p. 79 (cit. on p. 36).

Ras, Gabrielle, Pim Haselager, and Marcel van Gerven (2018). “Explanation Methods in Deep Learning: Users, Values, Concerns and Challenges”. In: *arXiv preprint arXiv:1803.07517* (cit. on p. 118).

Rebuffi, Sylvestre-Alvise, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert (2017). “icarl: Incremental classifier and representation learning”. In: *Proc. CVPR* (cit. on p. 106).

Rebuffi, Sylvestre-Alvise, Hakan Bilen, and Andrea Vedaldi (2018). “Efficient parametrization of multi-domain deep neural networks”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8119–8127 (cit. on p. 92).

Reed, Scott, Zeynep Akata, Xinchen Yan, et al. (2016). “Generative adversarial text to image synthesis”. In: *ICML* (cit. on p. 94).

Régis, Nicolas, Frédéric Dehais, Emmanuel Rachelson, et al. (2014). “Formal detection of attentional tunneling in human operator–automation interactions”. In: *IEEE Transactions on Human-Machine Systems* 44.3, pp. 326–336 (cit. on p. 24).

Reichle, Erik D, Keith Rayner, and Alexander Pollatsek (2003). “The EZ Reader model of eye-movement control in reading: Comparisons to other models”. In: *Behavioral and brain sciences* 26.4, pp. 445–476 (cit. on p. 16).

Reichle, Erik D, Alexander Pollatsek, and Keith Rayner (2012). “Using EZ Reader to simulate eye movements in nonreading tasks: A unified framework for understanding the eye–mind link.” In: *Psychological review* 119.1, p. 155 (cit. on p. 15).

Remazeilles, Anthony and François Chaumette (2007). “Image-based robot navigation from an image memory”. In: *Robotics and Autonomous Systems* 55.4, pp. 345–356 (cit. on p. 77).

Ren, Mengye and Richard S Zemel (2017). “End-to-end instance segmentation with recurrent attention”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6656–6664 (cit. on p. 80).

Ren, Mengye, Sachin Ravi, Eleni Triantafillou, et al. (2018). “Meta-Learning for Semi-Supervised Few-Shot Classification”. In: *International Conference on Learning Representations* (cit. on p. 92).

Ren, Shaoqing, Kaiming He, Ross Girshick, and Jian Sun (2015). “Faster r-cnn: Towards real-time object detection with region proposal networks”. In: *Advances in neural information processing systems*, pp. 91–99 (cit. on pp. 80, 81, 88).

Rensink, Ronald A., J. Kevin O’Regan, and James J. Clark (1997). “To See or Not to See: The Need for Attention to Perceive Changes in Scenes”. In: *Psychological Science* 8, pp. 368–373 (cit. on p. 44).

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016). “Why should i trust you?: Explaining the predictions of any classifier”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, pp. 1135–1144 (cit. on p. 124).

Richard D. Wright, Lawrence M. Ward (2008). *Orienting of Attention*. Oxford University Press, USA (cit. on p. 19).

Richardson, Matthew and Pedro Domingos (2006). “Markov logic networks”. In: *Machine learning* 62.1-2, pp. 107–136 (cit. on p. 101).

Richter, Charles and Nicholas Roy (2017). “Safe Visual Navigation via Deep Learning and Novelty Detection”. In: *Robotics: Science and Systems XIII, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA, July 12-16, 2017* (cit. on p. 169).

Richter, Michael M. and Rosina O. Weber (2016). *Case-based reasoning*. Springer (cit. on p. 130).

Richter, Stephan R, Vibhav Vineet, Stefan Roth, and Vladlen Koltun (2016). "Playing for data: Ground truth from computer games". In: *European conference on computer vision*. Springer, pp. 102–118 (cit. on p. 93).

Rick Chang, JH, Chun-Liang Li, Barnabas Poczos, BVK Vijaya Kumar, and Aswin C Sankaranarayanan (2017). "One Network to Solve Them All–Solving Linear Inverse Problems Using Deep Projection Models". In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5888–5897 (cit. on p. 101).

Robert, Thomas, Nicolas Thome, and Matthieu Cord (2018). "Hybridnet: Classification and reconstruction cooperation for semi-supervised learning". In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 153–169 (cit. on p. 90).

Roijers, Diederik M, Peter Vamplew, Shimon Whiteson, and Richard Dazeley (2013). "A survey of multi-objective sequential decision-making". In: *Journal of Artificial Intelligence Research* 48, pp. 67–113 (cit. on p. 71).

Rolls, Edmund T and Alessandro Treves (2011). "The neuronal encoding of information in the brain". In: *Progress in neurobiology* 95.3, pp. 448–490 (cit. on p. 35).

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox (2015). "U-net: Convolutional networks for biomedical image segmentation". In: *International Conference on Medical image computing and computer-assisted intervention*. Springer, pp. 234–241 (cit. on p. 33).

Rosch, Eleanor (1999). "Principles of categorization". In: *Concepts: core readings*. Ed. by E. Margolis and S. Laurence. MIT Press. Chap. 10, pp. 189–206 (cit. on p. 11).

Ross, Stéphane, Joelle Pineau, Sébastien Paquet, and Brahim Chaib-Draa (2008). "Online planning algorithms for POMDPs". In: *Journal of Artificial Intelligence Research* 32, pp. 663–704 (cit. on p. 71).

Roy, Sumantra Dutta, Santanu Chaudhury, and Subhashis Banerjee (2004). "Active recognition through next view planning: a survey". In: *Pattern Recognition* 37.3, pp. 429 –446 (cit. on p. 77).

Rozantsev, Artem, Mathieu Salzmann, and Pascal Fua (2018). "Beyond sharing weights for deep domain adaptation". In: *IEEE transactions on pattern analysis and machine intelligence* 41.4, pp. 801–814 (cit. on p. 90).

RTCA, Special Committee 205 of (2011). *DO-178C, Software Considerations in Airborne Systems and Equipment Certification*. Tech. rep. RTCA (cit. on p. 140).

Ruder, Sebastian (2017). "An overview of multi-task learning in deep neural networks". In: *arXiv preprint arXiv:1706.05098* (cit. on pp. 91, 109, 177).

Ruff, Lukas, Nico Goernitz, Lucas Deecke, et al. (2018). "Deep One-Class Classification". In: *International Conference on Machine Learning*, pp. 4390–4399 (cit. on p. 165).

Rullen, Rufin Van and Simon J Thorpe (2001). "Rate coding versus temporal order coding: what the retinal ganglion cells tell the visual cortex". In: *Neural computation* 13.6, pp. 1255–1283 (cit. on p. 35).

Russakovsky, Olga, Li-Jia Li, and Li Fei-Fei (2015). "Best of both worlds: human-machine collaboration for object annotation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2121–2131 (cit. on p. 171).

Russell, Stuart J and Peter Norvig (2016). *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited (cit. on p. 96).

Rusu, Andrei A, Neil C Rabinowitz, Guillaume Desjardins, et al. (2016). "Progressive neural networks". In: *arXiv preprint arXiv:1606.04671* (cit. on p. 105).

Rusu, Andrei A., Dushyant Rao, Jakub Sygnowski, et al. (2019). "Meta-Learning with Latent Embedding Optimization". In: *International Conference on Learning Representations* (cit. on p. 92).

Sabour, Sara, Nicholas Frosst, and Geoffrey E Hinton (2017). "Dynamic routing between capsules". In: *Advances in neural information processing systems*, pp. 3856–3866 (cit. on p. 52).

SAE (1996). *ARP4761 Guidelines and Methods for Conducting the Safety Assessment Process on Civil Airborne Systems and Equipment*. Tech. rep. SAE International (cit. on p. 141).

– (2010). *ARP4754A Guidelines for Development of Civil Aircraft and Systems*. Tech. rep. SAE International (cit. on p. 141).

Saito, Kuniaki, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada (2018). “Maximum classifier discrepancy for unsupervised domain adaptation”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3723–3732 (cit. on p. 90).

Salay, Rick and Krzysztof Czarnecki (2018). “Using Machine Learning Safely in Automotive Software: An Assessment and Adaption of Software Process Requirements in ISO 26262”. In: *CoRR* abs/1808.01614 (cit. on p. 140).

Salman, Hadi, Greg Yang, Huan Zhang, Cho-Jui Hsieh, and Pengchuan Zhang (2019). “A convex relaxation barrier to tight robustness verification of neural networks”. In: *Advances in Neural Information Processing Systems*, pp. 9832–9842 (cit. on p. 163).

Salvagnini, Pietro, Loris Bazzani, Marco Cristani, and Vittorio Murino (2013). “Person re-identification with a ptz camera: an introductory study”. In: *2013 IEEE International Conference on Image Processing*. IEEE, pp. 3552–3556 (cit. on p. 78).

Salvagnini, Pietro, Federico Pernici, Marco Cristani, et al. (2015). “Non-myopic information theoretic sensor management of a single pan–tilt–zoom camera for multiple object detection and tracking”. In: *Computer Vision and Image Understanding* 134, pp. 74–88 (cit. on p. 78).

Samek, Wojciech, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller (2017). “Evaluating the visualization of what a deep neural network has learned”. In: *IEEE transactions on neural networks and learning systems* 28.11, pp. 2660–2673 (cit. on p. 131).

Sanchez-Gonzalez, Alvaro, Nicolas Heess, Jost Tobias Springenberg, et al. (2018). “Graph Networks as Learnable Physics Engines for Inference and Control”. In: *International Conference on Machine Learning*, pp. 4467–4476 (cit. on p. 102).

Satsangi, Yash (2019). “Active perception for person tracking”. PhD thesis. University of Amsterdam (cit. on p. 79).

Savva, Manolis, Angel X. Chang, Alexey Dosovitskiy, Thomas Funkhouser, and Vladlen Koltun (2017). “MINOS: Multimodal Indoor Simulator for Navigation in Complex Environments”. In: *arXiv:1712.03931* (cit. on p. 85).

Savva, Manolis, Abhishek Kadian, Oleksandr Maksymets, et al. (2019). “Habitat: A platform for embodied ai research”. In: *arXiv preprint arXiv:1904.01201* (cit. on p. 85).

Saxe, Andrew Michael, Yamini Bansal, Joel Dapello, et al. (2018). “On the Information Bottleneck Theory of Deep Learning”. In: *International Conference on Learning Representations* (cit. on p. 121).

Saxena, Dhruv Mauria, Vince Kurtz, and Martial Hebert (2017). “Learning robust failure response for autonomous vision based flight”. In: *Robotics and Automation (ICRA), 2017 IEEE International Conference on*. IEEE, pp. 5824–5829 (cit. on p. 168).

Scaman, Kevin and Aladin Virmaux (2018). “Lipschitz regularity of deep neural networks: analysis and efficient estimation”. In: *arXiv preprint arXiv:1805.10965* (cit. on p. 163).

Schiele, Bernt and James L Crowley (1998). “Transinformation for active object recognition”. In: *Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271)*. IEEE, pp. 249–254 (cit. on p. 68).

Schmicking, Daniel (2010). “A toolbox of phenomenological methods”. In: *Handbook of phenomenology and cognitive science*. Springer, pp. 35–55 (cit. on p. 48).

Schmidhuber, Jürgen (2015). “Deep learning in neural networks: An overview”. In: *Neural networks* 61, pp. 85–117 (cit. on p. 87).

Schwartzstein, Joshua (2014). “Selective attention and learning”. In: *Journal of the European Economic Association* 12.6, pp. 1423–1452 (cit. on p. 111).

Scott, William R., Gerhard Roth, and Jean-François Rivest (2003). “View planning for automated three-dimensional object reconstruction and inspection”. In: *ACM Comput. Surv.* 35.1, pp. 64–96 (cit. on p. 77).

Seifert, Christin, Aisha Aamir, Aparna Balagopalan, et al. (2017). “Visualizations of Deep Neural Networks in Computer Vision: A Survey”. In: *Transparent Data Mining for Big and Small Data*. Springer, pp. 123–144 (cit. on p. 123).

Sejnowski, Terrence J and Gerald Tesauro (1989). “The Hebb rule for synaptic plasticity: algorithms and implementations”. In: *Neural models of plasticity*. Elsevier, pp. 94–103 (cit. on p. 113).

Selvaraju, Ramprasaath R, Michael Cogswell, Abhishek Das, et al. (2017). “Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 618–626 (cit. on pp. 125, 126, 131).

Sener, Ozan and Silvio Savarese (2018). “Active Learning for Convolutional Neural Networks: A Core-Set Approach”. In: *International Conference on Learning Representations* (cit. on p. 104).

Sermanet, Pierre, Andrea Frome, and Esteban Real (2015). “Attention for fine-grained categorization”. In: *International Conference on Learning Representations (ICLR 2015) workshop* (cit. on p. 80).

Seshia, Sanjit A, Dorsa Sadigh, and S Shankar Sastry (2016). “Towards verified artificial intelligence”. In: *arXiv preprint arXiv:1606.08514* (cit. on pp. 151, 163).

Seth, Anil K. (2015). “The Cybernetic Bayesian Brain”. In: *Open MIND*. Ed. by Thomas K. Metzinger and Jennifer M. Windt. Frankfurt am Main: MIND Group. Chap. 35(T) (cit. on p. 36).

Settles, Burr (2009). *Active learning literature survey*. Tech. rep. University of Wisconsin-Madison Department of Computer Sciences (cit. on p. 103).

Shafahi, Ali, Mahyar Najibi, Mohammad Amin Ghiasi, et al. (2019a). “Adversarial training for free!” In: *Advances in Neural Information Processing Systems*, pp. 3353–3364 (cit. on p. 158).

Shafahi, Ali, W. Ronny Huang, Christoph Studer, Soheil Feizi, and Tom Goldstein (2019b). “Are adversarial examples inevitable?” In: *International Conference on Learning Representations* (cit. on p. 161).

Shah, Shital, Debadeepta Dey, Chris Lovett, and Ashish Kapoor (2018). “Airsim: High-fidelity visual and physical simulation for autonomous vehicles”. In: *Field and service robotics*. Springer, pp. 621–635 (cit. on p. 153).

Shams, Ladan and Robyn Kim (2010). “Crossmodal influences on visual perception”. In: *Physics of life reviews* 7.3, pp. 269–284 (cit. on p. 44).

Shapiro, Lawrence (2014). *The Routledge Handbook of Embodied Cognition*. Routledge (cit. on p. 47).

Shi, Weiwei, Yihong Gong, Chris Ding, Zhiheng MaXiaoyu Tao, and Nanning Zheng (2018). “Transductive semi-supervised deep learning using min-max features”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 299–315 (cit. on p. 90).

Shim, Hajin, Sung Ju Hwang, and Eunho Yang (2018). “Joint active feature acquisition and classification with variable-size set encoding”. In: *Advances in Neural Information Processing Systems*, pp. 1368–1378 (cit. on p. 71).

Shin, Hanul, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim (2017). “Continual learning with deep generative replay”. In: *Advances in Neural Information Processing Systems*, pp. 2990–2999 (cit. on p. 105).

Shlens, Jonathon, Fred Rieke, and EJ Chichilnisky (2008). “Synchronized firing in the retina”. In: *Current opinion in neurobiology* 18.4, pp. 396–402 (cit. on p. 32).

Shoham, Yoav, Rob Powers, and Trond Grenager (2007). “If multi-agent learning is the answer, what is the question?” In: *Artificial Intelligence* 171.7, pp. 365–377 (cit. on p. 112).

Shrikumar, Avanti, Peyton Greenside, and Anshul Kundaje (2017). “Learning Important Features Through Propagating Activation Differences”. In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. International Convention Centre, Sydney, Australia: PMLR, pp. 3145–3153 (cit. on p. 125).

Shrivastava, Abhinav, Abhinav Gupta, and Ross Girshick (2016). “Training region-based object detectors with online hard example mining”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 761–769 (cit. on p. 111).

Shrivastava, Ashish, Tomas Pfister, Oncel Tuzel, et al. (2017). “Learning from Simulated and Unsupervised Images through Adversarial Training”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 2242–2251 (cit. on pp. 94, 153).

Shu, Jun, Zongben Xu, and Deyu Meng (2018). “Small Sample Learning in Big Data Era”. In: *arXiv preprint arXiv:1808.04572* (cit. on p. 89).
Shwartz-Ziv, Ravid and Naftali Tishby (2017). “Opening the black box of deep neural networks via information”. In: *arXiv preprint arXiv:1703.00810* (cit. on pp. 108, 118, 121).
Siegel, Susanna (2016). “The Contents of Perception”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2016. Metaphysics Research Lab, Stanford University (cit. on p. 46).
Siewert, Charles (2015). “Phenomenological Approaches”. In: *The Oxford Handbook of Philosophy of Perception* (cit. on p. 48).
Sigaud, Olivier and Alain Droniou (2015). “Towards deep developmental learning”. In: *IEEE Transactions on Cognitive and Developmental Systems* 8.2, pp. 99–114 (cit. on p. 91).
Silver, D, T Hubert, J Schrittwieser, et al. (2018). “A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play.” In: *Science (New York, NY)* 362.6419, pp. 1140–1144 (cit. on p. 71).
Simons, Daniel J and Ronald A Rensink (2005). “Change blindness: Past, present, and future”. In: *Trends in cognitive sciences* 9.1, pp. 16–20 (cit. on p. 44).
Simonyan, Karen and Andrew Zisserman (2014). “Very deep convolutional networks for large-scale image recognition”. In: *arXiv preprint arXiv:1409.1556* (cit. on p. 88).
Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman (2013). “Deep inside convolutional networks: Visualising image classification models and saliency maps”. In: *arXiv preprint arXiv:1312.6034* (cit. on pp. 122, 125).
Singh, Gagandeep, Timon Gehr, Matthew Mirman, Markus Püschel, and Martin Vechev (2018). “Fast and effective robustness certification”. In: *Advances in Neural Information Processing Systems*, pp. 10802–10813 (cit. on p. 167).
Singh, Gagandeep, Rupanshu Ganvir, Markus Püschel, and Martin Vechev (2019a). “Beyond the Single Neuron Convex Barrier for Neural Network Certification”. In: *Advances in Neural Information Processing Systems*, pp. 15072–15083 (cit. on p. 163).
Singh, Gagandeep, Timon Gehr, Markus Püschel, and Martin Vechev (2019b). “Boosting Robustness Certification of Neural Networks”. In: *International Conference on Learning Representations* (cit. on p. 163).
Sinha, Aman, Hongseok Namkoong, and John Duchi (2018). “Certifiable Distributional Robustness with Principled Adversarial Training”. In: *International Conference on Learning Representations* (cit. on p. 158).
Sitawarin, Chawin and David Wagner (2019a). “Defending Against Adversarial Examples with K-Nearest Neighbor”. In: *arXiv preprint arXiv:1906.09525* (cit. on p. 161).
– (2019b). “On the Robustness of Deep K-Nearest Neighbors”. In: *arXiv preprint arXiv:1903.08333* (cit. on p. 161).
Sitawarin, Chawin, Arjun Nitin Bhagoji, Arsalan Mosenia, Mung Chiang, and Prateek Mittal (2018). “Darts: Deceiving autonomous cars with toxic signs”. In: *arXiv preprint arXiv:1802.06430* (cit. on p. 159).
Smirnov, Evgeny, Aleksandr Melnikov, Andrei Oleinik, et al. (2018). “Hard example mining with auxiliary embeddings”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 37–46 (cit. on p. 112).
Snoek, Jasper, Yaniv Ovadia, Emily Fertig, et al. (2019). “Can you trust your model’s uncertainty? Evaluating predictive uncertainty under dataset shift”. In: *Advances in Neural Information Processing Systems*, pp. 13969–13980 (cit. on p. 166).
Soatto, Stefano (2013). “Actionable information in vision”. In: *Machine learning for computer vision*. Springer, pp. 17–48 (cit. on p. 82).
Sommerlade, E. and I. Reid (2008a). “Influence of zoom selection on a Kalman filter”. In: *Intelligent Robots and Systems, 2008. IROS 2008. IEEE/RSJ International Conference on*, pp. 2565–2571 (cit. on p. 78).

– (2008b). “Information-theoretic active scene exploration”. In: *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–7 (cit. on pp. 77, 78).

Sommerlade, Eric and Ian Reid (2008c). “Information-theoretic Decision Making for Exploration of Dynamic Scenes”. In: *Proceedings of the 5th International Workshop on Attention in Cognitive Systems (WAPCV)* (cit. on p. 78).

Song, Bi, Chong Ding, A.T. Kamal, J.A. Farrell, and A.K. Roy-chowdhury (2011). “Distributed Camera Networks”. In: *Signal Processing Magazine, IEEE* 28.3, pp. 20 –31 (cit. on p. 79).

Souza, Cesar Roberto de, Adrien Gaidon, Yohann Cabon, and Antonio Manuel Lopez (2017). “Procedural generation of videos to train deep action recognition networks”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4757–4767 (cit. on p. 93).

Spratling, Michael W (2012). “Unsupervised learning of generative and discriminative weights encoding elementary image components in a predictive coding model of cortical function”. In: *Neural computation* 24.1, pp. 60–103 (cit. on p. 56).

– (2016). “A neural implementation of Bayesian inference based on predictive coding”. In: *Connection Science* 28.4, pp. 346–383 (cit. on p. 56).

– (2017). “A hierarchical predictive coding model of object recognition in natural images”. In: *Cognitive computation* 9.2, pp. 151–167 (cit. on p. 56).

Steinhardt, Jacob, Pang Wei W Koh, and Percy S Liang (2017). “Certified defenses for data poisoning attacks”. In: *Advances in neural information processing systems*, pp. 3517–3529 (cit. on p. 114).

Stutz, David, Matthias Hein, and Bernt Schiele (2019). “Disentangling adversarial robustness and generalization”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6976–6987 (cit. on p. 158).

Su, Dong, Huan Zhang, Hongge Chen, et al. (2018). “Is Robustness the Cost of Accuracy?–A Comprehensive Study on the Robustness of 18 Deep Image Classification Models”. In: pp. 631–648 (cit. on pp. 158, 160).

Sukhbaatar, Sainbayar, Jason Weston, Rob Fergus, et al. (2015). “End-to-end memory networks”. In: *Advances in neural information processing systems*, pp. 2440–2448 (cit. on p. 81).

Sun, Chen, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta (2017). “Revisiting unreasonable effectiveness of data in deep learning era”. In: *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, pp. 843–852 (cit. on p. 91).

Sundararajan, Mukund, Ankur Taly, and Qiqi Yan (2017). “Axiomatic Attribution for Deep Networks”. In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. International Convention Centre, Sydney, Australia: PMLR, pp. 3319–3328 (cit. on p. 125).

Sünderhauf, Niko, Oliver Brock, Walter Scheirer, et al. (2018). “The limits and potentials of deep learning for robotics”. In: *The International Journal of Robotics Research* 37.4-5, pp. 405–420 (cit. on pp. 84, 88).

Sutton, Charles, Andrew McCallum, et al. (2012). “An introduction to conditional random fields”. In: *Foundations and Trends® in Machine Learning* 4.4, pp. 267–373 (cit. on p. 101).

Sutton, Richard S and Andrew G Barto (1998). *Introduction to reinforcement learning*. Vol. 2. 4. MIT press Cambridge (cit. on p. 71).

Swanson, Link R (2016). “The predictive processing paradigm has roots in Kant”. In: *Frontiers in systems neuroscience* 10, p. 79 (cit. on p. 38).

Swersky, Lorne, Henrique O Marques, Jöerg Sander, Ricardo JGB Campello, and Arthur Zimek (2016). “On the evaluation of outlier detection and one-class classification methods”. In: *Data Science and Advanced Analytics (DSAA), 2016 IEEE International Conference on*. IEEE, pp. 1–10 (cit. on p. 166).

Szegedy, Christian, Wojciech Zaremba, Ilya Sutskever, et al. (2014). “Intriguing properties of neural networks”. In: *ICLR* (cit. on p. 156).

Szegedy, Christian, Wei Liu, Yangqing Jia, et al. (2015). “Going deeper with convolutions”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9 (cit. on p. 88).

Szeliski, Richard (2010). *Computer vision: algorithms and applications*. Springer Science & Business Media (cit. on p. 97).

Sztipanovits, Janos, Xenofon Koutsoukos, Gabor Karsai, et al. (2011). "Toward a science of cyber–physical system integration". In: *Proceedings of the IEEE* 100.1, pp. 29–44 (cit. on p. 84).

Tan, Chuanqi, Fuchun Sun, Tao Kong, et al. (2018). "A survey on deep transfer learning". In: *International Conference on Artificial Neural Networks*. Springer, pp. 270–279 (cit. on p. 26).

Tang, Yong, Jens R Nyengaard, Didima MG De Groot, and Hans Jørgen G Gundersen (2001). "Total regional and global number of synapses in the human brain neocortex". In: *Synapse* 41.3, pp. 258–273 (cit. on p. 35).

Tartakovsky, Alexander, Igor Nikiforov, and Michele Basseville (2014). *Sequential analysis: Hypothesis testing and changepoint detection*. Chapman and Hall/CRC (cit. on p. 72).

Tatler, Benjamin W, Nicholas J Wade, Hoi Kwan, John M Findlay, and Boris M Velichkovsky (2010). "Yarbus, eye movements, and vision". In: *i-Perception* 1.1, pp. 7–27 (cit. on pp. 42, 80).

Thacker, Neil A, Adrian F Clark, John L Barron, et al. (2008). "Performance characterization in computer vision: A guide to best practices". In: *Computer vision and image understanding* 109.3, pp. 305–334 (cit. on p. 152).

Thinès, Georges (2016). "Perception". French. In: *Dictionnaire de la philosophie*. Encyclopaedia Universalis (cit. on p. 8).

Thomson, Alex M. and Christophe Lamy (2007). "Functional maps of neocortical local circuitry". In: *Frontiers in neuroscience* 1, p. 2 (cit. on p. 33).

Thrun, Sebastian and Tom M. Mitchell (1995). "Lifelong robot learning". In: *Robotics and Autonomous Systems* 15, pp. 25–46 (cit. on p. 104).

Tian, Yuchi, Kexin Pei, Suman Jana, and Baishakhi Ray (2018). "Deeptest: Automated testing of deep-neural-network-driven autonomous cars". In: *Proceedings of the 40th International Conference on Software Engineering*. ACM, pp. 303–314 (cit. on p. 162).

Tishby, Naftali and Noga Zaslavsky (2015). "Deep learning and the information bottleneck principle". In: *Information Theory Workshop (ITW), 2015 IEEE*. IEEE, pp. 1–5 (cit. on p. 121).

Tishby, Naftali, Fernando C Pereira, and William Bialek (2000). "The information bottleneck method". In: *arXiv preprint physics/0004057* (cit. on p. 121).

Tjeng, Vincent and Russ Tedrake (2017). "Verifying neural networks with mixed integer programming". In: *arXiv preprint arXiv:1711.07356* (cit. on p. 163).

Tomsett, Richard, Dave Braines, Dan Harborne, Alun Preece, and Supriyo Chakraborty (2018). "Interpretable to whom? A role-based model for analyzing interpretable machine learning systems". In: *ICML Workshop on Human Interpretability in Machine Learning* (cit. on p. 118).

Tordoff, B. and D. Murray (Jan. 2004). "Reactive control of zoom while fixating using perspective and affine cameras". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 26.1, pp. 98–112 (cit. on p. 78).

Tordoff, B.J. and D.W. Murray (2007). "A method of reactive zoom control from uncertainty in tracking". In: *Computer Vision and Image Understanding* 105, pp. 131–144 (cit. on p. 78).

Tousch, Anne-Marie (2010). "Hiérarchies sémantiques pour l'annotation multifacette d'images". PhD thesis. Université Paris-Est (cit. on pp. 130, 233).

Tousch, Anne-Marie, Stéphane Herbin, and Jean-Yves Audibert (2008). "Semantic lattices for multiple annotation of images". In: *Proceedings of the 1st ACM SIGMM International Conference on Multimedia Information Retrieval, MIR 2008, Vancouver, British Columbia, Canada, October 30-31, 2008*, pp. 342–349 (cit. on pp. 130, 233).

– (2012). "Semantic hierarchies for image annotation: A survey". In: *Pattern Recognition* 45.1, pp. 333–345 (cit. on p. 11).

Town, Christopher (2006). "Ontological inference for image and video analysis". In: *Machine Vision and Applications* 17.2, p. 94 (cit. on p. 109).

Tramèr, Florian and Dan Boneh (2019). "Adversarial training and robustness for multiple perturbations". In: *Advances in Neural Information Processing Systems*, pp. 5858–5868 (cit. on p. 158).

Tramèr, Florian, Alexey Kurakin, Nicolas Papernot, et al. (2018). “Ensemble Adversarial Training: Attacks and Defenses”. In: *International Conference on Learning Representations* (cit. on p. 158).

Trott, Alexander, Caiming Xiong, and Richard Socher (2018). “Interpretable Counting for Visual Question Answering”. In: *International Conference on Learning Representations* (cit. on p. 127).

Tsin, Yanghai, Y. Genc, and V. Ramesh (2009). “Explicit 3D Modeling for Vehicle Monitoring in Non-overlapping Cameras”. In: *Advanced Video and Signal Based Surveillance, 2009. AVSS ’09. Sixth IEEE International Conference on*, pp. 110–115 (cit. on p. 79).

Tsipras, Dimitris, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry (2019). “Robustness May Be at Odds with Accuracy”. In: *International Conference on Learning Representations* (cit. on pp. 158, 160).

Tsotsos, John K. (1992). “On the relative complexity of active vs. passive visual search”. In: *International Journal of Computer Vision* 7.2, pp. 127–141 (cit. on p. 82).

Tsotsos, John K (2011). *A computational perspective on visual attention* (cit. on p. 80).

Tuia, Devis, Frédéric Ratle, Fabio Pacifici, Mikhail F Kanevski, and William J Emery (2009). “Active learning methods for remote sensing image classification”. In: *IEEE Transactions on Geoscience and Remote Sensing* 47.7, pp. 2218–2232 (cit. on p. 103).

Uijlings, Jasper RR, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders (2013). “Selective search for object recognition”. In: *International journal of computer vision* 104.2, pp. 154–171 (cit. on p. 81).

Valiant, Leslie G (1984). “A theory of the learnable”. In: *Proceedings of the sixteenth annual ACM symposium on Theory of computing*. ACM, pp. 436–445 (cit. on pp. 79, 87).

Vanschoren, Joaquin (2018). “Meta-learning: A survey”. In: *arXiv preprint arXiv:1810.03548* (cit. on p. 92).

Vapnik, Vladimir (2013). *The nature of statistical learning theory*. Springer science & business media (cit. on p. 87).

Varela, F., E. Thompson, and E. Rosch (1991). *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge, MA: MIT Press (cit. on pp. 19, 47).

Varshney, Kush R (2016). “Engineering safety in machine learning”. In: *Information Theory and Applications Workshop (ITA), 2016*. IEEE, pp. 1–5 (cit. on p. 163).

Varshney, Kush R and Homa Alemzadeh (2017). “On the safety of machine learning: Cyber-physical systems, decision sciences, and data products”. In: *Big data* 5.3, pp. 246–255 (cit. on p. 163).

Vaswani, Ashish, Noam Shazeer, Niki Parmar, et al. (2017). “Attention is all you need”. In: *Advances in Neural Information Processing Systems*, pp. 5998–6008 (cit. on pp. 19, 43, 81).

Vedantam, Ramakrishna, C Lawrence Zitnick, and Devi Parikh (2015). “Cider: Consensus-based image description evaluation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575 (cit. on p. 126).

Veit, Andreas, Michael Wilber, Rajan Vaish, et al. (2015). “On optimizing human-machine task assignments”. In: *HCOMP-15* (cit. on p. 171).

Velez, Javier, Garrett Hemann, Albert S Huang, Ingmar Posner, and Nicholas Roy (2011). “Planning to perceive: Exploiting mobility for robust object detection”. In: *Twenty-First International Conference on Automated Planning and Scheduling* (cit. on p. 77).

Velez, Javier, Garrett Hemann, Albert S. Huang, Ingmar Posner, and Nicholas Roy (2012). “Modelling Observation Correlations for Active Exploration and Robust Object Detection”. In: *Journal of Artificial Intelligence Research* 44, pp. 423–453 (cit. on p. 77).

Ven, Gido M van de and Andreas S Tolias (2018). “Generative replay with feedback connections as a general strategy for continual learning”. In: *arXiv preprint arXiv:1809.10635* (cit. on p. 105).

– (2019). “Three scenarios for continual learning”. In: *arXiv preprint arXiv:1904.07734* (cit. on p. 105).

Venkateswara, Hemanth, Shayok Chakraborty, and Sethuraman Panchanathan (2017). “Deep-learning systems for domain adaptation in computer vision: Learning transferable feature representations”. In: *IEEE Signal Processing Magazine* 34.6, pp. 117–129 (cit. on p. 89).

Vezhnevets, Alexander, Joachim M Buhmann, and Vittorio Ferrari (2012). “Active learning for semantic segmentation with expected change”. In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 3162–3169 (cit. on p. 103).

Villani, C. (2018). *Donner un sens à l’intelligence artificielle. Pour une stratégie nationale et européenne*. en français (cit. on pp. 117, 139).

Vinyals, Oriol, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. (2016). “Matching networks for one shot learning”. In: *Advances in neural information processing systems*, pp. 3630–3638 (cit. on p. 92).

Vinyals, Oriol, Igor Babuschkin, Junyoung Chung, et al. (2019). *AlphaStar: Mastering the Real-Time Strategy Game StarCraft II*. https://deepmind.com/blog/alphastar-mastering-real-time-strategy-game-starcraft-ii/ (cit. on p. 71).

Viola, Paul and Michael Jones (2001). “Rapid object detection using a boosted cascade of simple features”. In: *CVPR 2001*. Vol. 1. IEEE, I:511–518 (cit. on p. 72).

Vries, Harm de, Kurt Shuster, Dhruv Batra, et al. (2018). “Talk the walk: Navigating new york city through grounded dialogue”. In: *arXiv preprint arXiv:1807.03367* (cit. on p. 86).

Vu, Tuan-Hung, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez (2019). “Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2517–2526 (cit. on p. 94).

Wagemans, Johan (2015). *The Oxford handbook of perceptual organization*. 1. ed. Oxford library of psychology. Oxford Univ. Press (cit. on p. 41).

Wald, Abraham and Jacob Wolfowitz (1948). “Optimum character of the sequential probability ratio test”. In: *The Annals of Mathematical Statistics* 19.3, pp. 326–339 (cit. on p. 72).

Wang, Fei, Mengqing Jiang, Chen Qian, et al. (2017). “Residual attention network for image classification”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3156–3164 (cit. on p. 81).

Wang, Feng and David MJ Tax (2016). “Survey on the attention based RNN model and its applications in computer vision”. In: *arXiv preprint arXiv:1601.06823* (cit. on p. 80).

Wang, Fulton and Cynthia Rudin (2015). “Falling rule lists”. In: *Artificial Intelligence and Statistics*, pp. 1013–1022 (cit. on pp. 124, 130).

Wang, K., D. Zhang, Y. Li, R. Zhang, and L. Lin (2017). “Cost-Effective Active Learning for Deep Image Classification”. In: *IEEE Transactions on Circuits and Systems for Video Technology* 27.12, pp. 2591–2600 (cit. on p. 104).

Wang, Mei and Weihong Deng (2018). “Deep visual domain adaptation: A survey”. In: *Neurocomputing* 312, pp. 135–153 (cit. on p. 26).

Wang, Meng and Xian-Sheng Hua (2011). “Active learning in multimedia annotation and retrieval: A survey”. In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 2.2, p. 10 (cit. on p. 103).

Wang, Qiang, Zhu Teng, Junliang Xing, et al. (2018a). “Learning attentions: residual attentional siamese network for high performance online visual tracking”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4854–4863 (cit. on p. 80).

Wang, Shiqi, Kexin Pei, Justin Whitehouse, Junfeng Yang, and Suman Jana (2018b). “Efficient formal safety analysis of neural networks”. In: *Advances in Neural Information Processing Systems*, pp. 6367–6377 (cit. on p. 163).

Wang, Ting-Chun, Ming-Yu Liu, Jun-Yan Zhu, et al. (2018c). “High-Resolution Image Synthesis and Semantic Manipulation With Conditional GANs”. In: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 8798–8807 (cit. on pp. 94, 153).

Wang, Ting-Chun, Ming-Yu Liu, Jun-Yan Zhu, et al. (2018d). “Video-to-Video Synthesis”. In: *Advances in Neural Information Processing Systems*, pp. 1144–1156 (cit. on p. 94).

Wang, Tong (2018). “Hybrid Decision Making: When Interpretable Models Collaborate With Black-Box Models”. In: *arXiv preprint arXiv:1802.04346* (cit. on p. 130).

Wang, Tong, Cynthia Rudin, Finale Doshi-Velez, et al. (2017). “A bayesian framework for learning rule sets for interpretable classification”. In: *The Journal of Machine Learning Research* 18.1, pp. 2357–2393 (cit. on p. 130).

Wang, Wei, Vincent W Zheng, Han Yu, and Chunyan Miao (2019a). “A survey of zero-shot learning: Settings, methods, and applications”. In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 10.2, p. 13 (cit. on p. 92).

Wang, Xiaogang (2013). “Intelligent multi-camera video surveillance: A review”. In: *Pattern recognition letters* 34.1, pp. 3–19 (cit. on p. 79).

Wang, Yaqing and Quanming Yao (2019). “Few-shot learning: A survey”. In: *arXiv preprint arXiv:1904.05046* (cit. on p. 92).

Wang, Yizhen and Kamalika Chaudhuri (2018). “Data poisoning attacks against online learning”. In: *arXiv preprint arXiv:1808.08994* (cit. on p. 114).

Wang, Yu-Xiong and Martial Hebert (2016). “Learning to learn: Model regression networks for easy small sample learning”. In: *European Conference on Computer Vision*. Springer, pp. 616–634 (cit. on p. 92).

Wang, Zhengwei, Qi She, and Tomas E Ward (2019b). “Generative Adversarial Networks: A Survey and Taxonomy”. In: *arXiv preprint arXiv:1906.01529* (cit. on p. 94).

Wang, Zhou and Alan C Bovik (2001). “Embedded foveation image coding”. In: *IEEE Transactions on image processing* 10.10, pp. 1397–1410 (cit. on p. 80).

Wässle, Heinz (2004). “Parallel processing in the mammalian retina”. In: *Nature Reviews Neuroscience* 5.10, p. 747 (cit. on p. 31).

Watzl, Sebastian (2017). *Structuring mind: The nature of attention and how it shapes consciousness*. Oxford University Press (cit. on p. 43).

Weber, Cornelius and Jochen Triesch (2009). “Implementations and implications of foveated vision”. In: *Recent Patents on Computer Science* 2.1, pp. 75–85 (cit. on p. 80).

Webster, Ryan, Julien Rabin, Loic Simon, and Frederic Jurie (2019). “Detecting overfitting of deep generative networks via latent recovery”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11273–11282 (cit. on p. 95).

Wei, Donglai, Bolei Zhou, Antonio Torrabla, and William Freeman (2015). “Understanding intra-class knowledge inside CNN”. In: *arXiv preprint arXiv:1507.02379* (cit. on p. 123).

Weinshall, Daphna and Dan Amir (2018). “Theory of curriculum learning, with convex loss functions”. In: *arXiv preprint arXiv:1812.03472* (cit. on p. 114).

Weiss, David J and Ben Taskar (2013). “Learning adaptive value of information for structured prediction”. In: *Advances in neural information processing systems*, pp. 953–961 (cit. on p. 71).

Weiss, Gerhard, ed. (2013). *Multiagent systems* (cit. on pp. 18, 19).

Weiss, Karl, Taghi M Khoshgoftaar, and DingDing Wang (2016). “A survey of transfer learning”. In: *Journal of Big data* 3.1, p. 9 (cit. on pp. 26, 89).

Wen, Haiguang, Kuan Han, Junxing Shi, et al. (2018). “Deep Predictive Coding Network for Object Recognition”. In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. Stockholmsmässan, Stockholm Sweden: PMLR, pp. 5266–5275 (cit. on p. 56).

Weng, Tsui-Wei, Huan Zhang, Pin-Yu Chen, et al. (2018). “Evaluating the Robustness of Neural Networks: An Extreme Value Theory Approach”. In: *ICLR* (cit. on p. 163).

Wenhardt, S., B. Deutsch, J. Hornegger, H. Niemann, and J. Denzler (2006). “An Information Theoretic Approach for Next Best View Planning in 3-D Reconstruction”. In: *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*. Vol. 1, pp. 103–106 (cit. on p. 77).

Wenhardt, S., B. Deutsch, E. Angelopoulou, and H. Niemann (2007). “Active Visual Object Reconstruction using D-, E-, and T-Optimal Next Best Views”. In: *Computer Vision and Pattern Recognition, 2007. CVPR ’07. IEEE Conference on*, pp. 1–7 (cit. on p. 77).

Wermter, Stefan and Ron Sun, eds. (2000). *Hybrid neural systems*. Lecture Notes on Computer Science 1778. Springer Science & Business Media (cit. on p. 96).

Weston, Jason, Sumit Chopra, and Antoine Bordes (2014). “Memory networks”. In: *arXiv preprint arXiv:1410.3916* (cit. on p. 52).

Wickens, Christopher D and Amy L Alexander (2009). “Attentional tunneling and task management in synthetic vision displays”. In: *The International Journal of Aviation Psychology* 19.2, pp. 182–199 (cit. on p. 24).

Wicker, Matthew, Xiaowei Huang, and Marta Kwiatkowska (2018). “Feature-guided black-box safety testing of deep neural networks”. In: *International Conference on Tools and Algorithms for the Construction and Analysis of Systems*. Springer, pp. 408–426 (cit. on p. 162).

Wienbar, Sophia and GregoryW Schwartz (2018). “The dynamic receptive fields of retinal ganglion cells”. In: *Progress in retinal and eye research* (cit. on p. 32).

Wiese, Wanja and Thomas K. Metzinger (2017). “Vanilla PP for Philosophers: A Primer on Predictive Processing”. In: *Philosophy and Predictive Processing*. Ed. by Thomas K. Metzinger and Wanja Wiese. Frankfurt am Main: MIND Group. Chap. 1 (cit. on p. 37).

Wiesel, Torsten N and David H Hubel (1963). “Single-cell responses in striate cortex of kittens deprived of vision in one eye”. In: *Journal of neurophysiology* 26.6, pp. 1003–1017 (cit. on p. 113).

Wikipedia (2019). *Neural coding — Wikipedia, The Free Encyclopedia*. [Online; accessed 26-March-2019] (cit. on p. 35).

Wilson, Robert A. and Lucia Foglia (2017). “Embodied Cognition”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Spring 2017. Metaphysics Research Lab, Stanford University (cit. on p. 47).

Wirth, Christian, Riad Akrour, Gerhard Neumann, and Johannes Fürnkranz (2017). “A survey of preference-based reinforcement learning methods”. In: *The Journal of Machine Learning Research* 18.1, pp. 4945–4990 (cit. on p. 71).

Wloka, Calden, Iuliia Kotseruba, and John K Tsotsos (2018). “Active Fixation Control to Predict Saccade Sequences”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3184–3193 (cit. on p. 80).

Wong, Eric, Frank Schmidt, Jan Hendrik Metzen, and J Zico Kolter (2018). “Scaling provable adversarial defenses”. In: *Advances in Neural Information Processing Systems*, pp. 8400–8409 (cit. on p. 167).

Wong, Eric, Leslie Rice, and J Zico Kolter (2020). “Fast is better than free: Revisiting adversarial training”. In: *arXiv preprint arXiv:2001.03994* (cit. on p. 158).

Woo, Sanghyun, Jongchan Park, Joon-Young Lee, and In So Kweon (2018). “Cbam: Convolutional block attention module”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 3–19 (cit. on p. 81).

Wu, Jiajun, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum (2016). “Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling”. In: *Advances in neural information processing systems*, pp. 82–90 (cit. on p. 94).

Wu, Weibin, Hui Xu, Sanqiang Zhong, Michael R Lyu, and Irwin King (2019). “Deep validation: Toward detecting real-world corner cases for deep neural networks”. In: *2019 49th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*. IEEE, pp. 125–137 (cit. on p. 162).

Xia, Fei, Amir R Zamir, Zhiyang He, et al. (2018). “Gibson env: Real-world perception for embodied agents”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9068–9079 (cit. on p. 85).

Xian, Yongqin, Christoph H Lampert, Bernt Schiele, and Zeynep Akata (2018). “Zero-shot learning-A comprehensive evaluation of the good, the bad and the ugly”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (cit. on p. 92).

Xiao, Hu, Rongxin Cui, and Demin Xu (2017). “A sampling-based bayesian approach for cooperative multiagent online search with resource constraints”. In: *IEEE transactions on cybernetics* 48.6, pp. 1773–1785 (cit. on p. 79).

Xie, Cihang, Jianyu Wang, Zhishuai Zhang, et al. (2017). “Adversarial examples for semantic segmentation and object detection”. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1369–1378 (cit. on p. 156).

Xu, Han, Yao Ma, Haochen Liu, et al. (2019). “Adversarial attacks and defenses in images, graphs and text: A review”. In: *arXiv preprint arXiv:1909.08072* (cit. on p. 156).

Xu, Huijuan and Kate Saenko (2016). “Ask, attend and answer: Exploring question-guided spatial attention for visual question answering”. In: *European Conference on Computer Vision*. Springer, pp. 451–466 (cit. on pp. 80, 127).

Xu, Kelvin, Jimmy Ba, Ryan Kiros, et al. (2015). “Show, attend and tell: Neural image caption generation with visual attention”. In: *International conference on machine learning*, pp. 2048–2057 (cit. on p. 80).

Xu, Weilin, David Evans, and Yanjun Qi (2017). “Feature squeezing: Detecting adversarial examples in deep neural networks”. In: *arXiv preprint arXiv:1704.01155* (cit. on p. 167).

Xu, Xiaojun, Xinyun Chen, Chang Liu, et al. (2018). “Fooling Vision and Language Models Despite Localization and Attention Mechanism”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4951–4961 (cit. on pp. 126, 127).

Yampolskiy, Roman V and MS Spellchecker (2016). “artificial intelligence safety and cybersecurity: a timeline of AI failures”. In: *arXiv preprint arXiv:1610.07997* (cit. on p. 151).

Yan, Claudia, Dipendra Misra, Andrew Bennnett, et al. (2018). “CHALET: Cornell house agent learning environment”. In: *arXiv preprint arXiv:1801.07357* (cit. on p. 85).

Yang, Hongyu, Cynthia Rudin, and Margo Seltzer (2017). “Scalable Bayesian Rule Lists”. In: *International Conference on Machine Learning*, pp. 3921–3930 (cit. on p. 130).

Yang, Jianwei, Zhile Ren, Mingze Xu, et al. (2019). “Embodied Visual Recognition”. In: *arXiv preprint arXiv:1904.04404* (cit. on p. 84).

Yang, Tianyu and Antoni B Chan (2018). “Learning dynamic memory networks for object tracking”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 152–167 (cit. on p. 80).

Yang, Yongxin and Timothy Hospedales (2016). “Deep multi-task representation learning: A tensor factorisation approach”. In: *ICLR* (cit. on p. 91).

Yarbus, A. (1967). *Eye Movements and Vision.* New York: Plenum Press. (cit. on p. 42).

Ye, Yiming and John K. Tsotsos (2001). “A Complexity-Level Analysis of the Sensor Planning Task for Object Search”. In: *Computational Intelligence* 17.3, pp. 605–620 (cit. on pp. 77, 82).

Yin, Dong, Ramchandran Kannan, and Peter Bartlett (2019). “Rademacher Complexity for Adversarially Robust Generalization”. In: *International Conference on Machine Learning*, pp. 7085–7094 (cit. on p. 158).

Yin, Zhichao and Jianping Shi (2018). “Geonet: Unsupervised learning of dense depth, optical flow and camera pose”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1983–1992 (cit. on p. 101).

Yoo, Donggeun and In So Kweon (2019). “Learning Loss for Active Learning”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 93–102 (cit. on p. 104).

Yoo, Donggeun, Sunggyun Park, Joon-Young Lee, Anthony S Paek, and In So Kweon (2015). “Attentionnet: Aggregating weak directions for accurate object detection”. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2659–2667 (cit. on p. 80).

Yosinski, Jason, Jeff Clune, Yoshua Bengio, and Hod Lipson (2014). “How transferable are features in deep neural networks?” In: *Advances in neural information processing systems*, pp. 3320–3328 (cit. on pp. 89, 92).

You, Quanzeng, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo (2016). “Image captioning with semantic attention”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4651–4659 (cit. on p. 80).

Young, Tom, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria (2018). “Recent trends in deep learning based natural language processing”. In: *ieee Computational intelligenCe magazine* 13.3, pp. 55–75 (cit. on p. 80).

Yuan, Yuhui, Kuiyuan Yang, and Chao Zhang (2017). “Hard-aware deeply cascaded embedding”. In: *Proceedings of the IEEE international conference on computer vision*, pp. 814–823 (cit. on p. 112).

Yuille, Alan L. and Chenxi Liu (2019). *Limitations of Deep Learning for Vision, and How We Might Fix Them*. https://thegradient.pub/ (cit. on p. 108).

Yun, Sangdoo, Jongwon Choi, Youngjoon Yoo, Kimin Yun, and Jin Young Choi (2017). “Action-decision networks for visual tracking with deep reinforcement learning”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2711–2720 (cit. on p. 80).

Yun, Sangdoo, Jongwon Choi, Youngjoon Yoo, Kimin Yun, and Jin Young Choi (2018). “Action-driven visual object tracking with deep reinforcement learning”. In: *IEEE transactions on neural networks and learning systems* 29.6, pp. 2239–2252 (cit. on p. 80).

Zadra, Jonathan R and Gerald L Clore (2011). “Emotion and perception: The role of affective information”. In: *Wiley interdisciplinary reviews: cognitive science* 2.6, pp. 676–685 (cit. on p. 44).

Zahavi, Dan (2010). “Naturalized phenomenology”. In: *Handbook of phenomenology and cognitive science*. Springer, pp. 2–19 (cit. on p. 49).

– (2012). *The Oxford handbook of contemporary phenomenology*. Oxford University Press Oxford, UK (cit. on p. 47).

– (2018). “Brain, Mind, World: Predictive coding, neo-Kantianism, and transcendental idealism”. In: *Husserl Studies* 34.1, pp. 47–61 (cit. on p. 38).

Zajc, Luka Cehovin, Alan Lukežic, Aleš Leonardis, and Matej Kristan (2017). “Beyond standard benchmarks: Parameterizing performance evaluation in visual object tracking”. In: *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, pp. 3343–3351 (cit. on p. 153).

Zamir, Amir R, Alexander Sax, and William Shen (2018). “Taskonomy: Disentangling Task Transfer Learning”. In: *CVPR* (cit. on pp. 21, 109).

Zarka, John, Louis Thiry, Tomás Angles, and Stéphane Mallat (2019). “Deep Network classification by Scattering and Homotopy dictionary learning”. In: *arXiv preprint arXiv:1910.03561* (cit. on p. 108).

Zeiler, Matthew D and Rob Fergus (2014). “Visualizing and understanding convolutional networks”. In: *European conference on computer vision*. Springer, pp. 818–833 (cit. on pp. 122, 123).

Zeimbekis, John and Athanassios Raftopoulos (2015). *The cognitive penetrability of perception: New philosophical perspectives*. Oxford university Press (cit. on p. 39).

Zendel, Oliver, Markus Murschitz, Martin Humenberger, and Wolfgang Herzner (2015). “Cv-hazop: Introducing test data validation for computer vision”. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2066–2074 (cit. on p. 147).

Zendel, Oliver, Katrin Honauer, Markus Murschitz, Martin Humenberger, and Gustavo Fernández Domínguez (2017a). “Analyzing Computer Vision Data—The Good, the Bad and the Ugly”. In: *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. IEEE, pp. 6670–6680 (cit. on p. 148).

Zendel, Oliver, Markus Murschitz, Martin Humenberger, and Wolfgang Herzner (2017b). “How Good Is My Test Data? Introducing Safety Analysis for Computer Vision”. In: *International Journal of Computer Vision* 125.1-3, pp. 95–109 (cit. on pp. 147, 148, 153, 162, 174).

Zeng, Jiaming, Berk Ustun, and Cynthia Rudin (2017). “Interpretable classification models for recidivism prediction”. In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 180.3, pp. 689–722 (cit. on p. 130).

Zenke, Friedemann, Ben Poole, and Surya Ganguli (2017a). “Continual Learning Through Synaptic Intelligence”. In: *ICML* (cit. on pp. 104, 105).

Zenke, Friedemann, Wulfram Gerstner, and Surya Ganguli (2017b). “The temporal paradox of Hebbian learning and homeostatic plasticity”. In: *Current opinion in neurobiology* 43, pp. 166–176 (cit. on p. 105).

Zhai, Shuangfei, Yu Cheng, Weining Lu, and Zhongfei Zhang (2016). “Deep Structured Energy Based Models for Anomaly Detection”. In: *International Conference on Machine Learning*, pp. 1100–1109 (cit. on p. 165).

Zhang, Chiyuan, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals (2016). “Understanding deep learning requires rethinking generalization”. In: *arXiv preprint arXiv:1611.03530* (cit. on p. 108).

Zhang, Han, Tao Xu, Hongsheng Li, et al. (2017a). “Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks”. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5907–5915 (cit. on p. 94).

Zhang, Hongyang, Yaodong Yu, Jiantao Jiao, et al. (2019a). “Theoretically Principled Trade-off between Robustness and Accuracy”. In: *International Conference on Machine Learning*, pp. 7472–7482 (cit. on p. 158).

Zhang, Hongyi, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz (2018a). “mixup: Beyond empirical risk minimization”. In: (cit. on p. 158).

Zhang, Huan, Tsui-Wei Weng, Pin-Yu Chen, Cho-Jui Hsieh, and Luca Daniel (2018b). “Efficient neural network robustness certification with general activation functions”. In: *Advances in neural information processing systems*, pp. 4939–4948 (cit. on p. 167).

Zhang, Jianming, You Wu, Wenjun Feng, and Jin Wang (2019b). “Spatially Attentive Visual Tracking Using Multi-Model Adaptive Response Fusion”. In: *IEEE Access* 7, pp. 83873–83887 (cit. on p. 80).

Zhang, Mengshi, Yuqun Zhang, Lingming Zhang, Cong Liu, and Sarfraz Khurshid (2018c). “DeepRoad: GAN-based Metamorphic Autonomous Driving System Testing”. In: *arXiv preprint arXiv:1802.02295* (cit. on p. 162).

Zhang, Peng, Jiuling Wang, Ali Farhadi, Martial Hebert, and Devi Parikh (2014). “Predicting Failures of Vision Systems”. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 3566–3573 (cit. on p. 132).

Zhang, Quanshi, Ruiming Cao, Feng Shi, Ying Nian Wu, and Song-Chun Zhu (2017b). “Interpreting cnn knowledge via an explanatory graph”. In: *arXiv:1708.01785* (cit. on p. 131).

Zhang, Quanshi, Yu Yang, Ying Nian Wu, and Song-Chun Zhu (2018d). “Interpreting CNNs via decision trees”. In: *arXiv:1802.00121* (cit. on p. 124).

Zhang, Quanshi, Yu Yang, Yuchen Liu, Ying Nian Wu, and Song-Chun Zhu (2018e). “Unsupervised Learning of Neural Networks to Explain Neural Networks”. In: *arXiv preprint arXiv:1805.07468* (cit. on p. 124).

Zhang, Wei Emma, Quan Z Sheng, AHOUD Alhazmi, and CHENLIANG LI (2019c). “Adversarial attacks on deep learning models in natural language processing: A survey”. In: *arXiv preprint arXiv:1901.06796* (cit. on p. 156).

Zhao, Feng and Xianghua Xie (2013). “An overview of interactive medical image segmentation”. In: *Annals of the BMVA* 2013.7, pp. 1–22 (cit. on p. 171).

Zheng, Heliang, Jianlong Fu, Tao Mei, and Jiebo Luo (2017). “Learning multi-attention convolutional neural network for fine-grained image recognition”. In: *Proceedings of the IEEE international conference on computer vision*, pp. 5209–5217 (cit. on p. 80).

Zheng, Liang, Yi Yang, and Alexander G Hauptmann (2016). “Person re-identification: Past, present and future”. In: *arXiv preprint arXiv:1610.02984* (cit. on p. 79).

Zhou, Bolei, Yiyou Sun, David Bau, and Antonio Torralba (2018a). “Revisiting the importance of individual units in cnns via ablation”. In: *arXiv preprint arXiv:1806.02891* (cit. on p. 135).

Zhou, Yanzhao, Yi Zhu, Qixiang Ye, Qiang Qiu, and Jianbin Jiao (2018b). “Weakly supervised instance segmentation using class peak response”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3791–3800 (cit. on p. 90).

Zhou, Zhi-Hua (2017). “A brief introduction to weakly supervised learning”. In: *National Science Review* 5.1, pp. 44–53 (cit. on p. 90).

Zhu, Jun-Yan, Taesung Park, Phillip Isola, and Alexei A Efros (2017). “Unpaired image-to-image translation using cycle-consistent adversarial networks”. In: *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232 (cit. on p. 94).

Zhu, Xiaojin and Andrew B Goldberg (2009). “Introduction to semi-supervised learning”. In: *Synthesis lectures on artificial intelligence and machine learning* 3.1, pp. 1–130 (cit. on p. 90).

Zhu, Xiaojin Jerry (2005). *Semi-supervised learning literature survey*. Tech. rep. University of Wisconsin-Madison Department of Computer Sciences (cit. on p. 90).

Zhu, Yuke, Oliver Groth, Michael Bernstein, and Li Fei-Fei (2016). "Visual7w: Grounded question answering in images". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4995–5004 (cit. on p. 127).

Zhuang, Bohan, Qi Wu, Chunhua Shen, Ian Reid, and Anton van den Hengel (2018). "Parallel attention: A unified framework for visual object discovery through dialogs and queries". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4252–4261 (cit. on p. 86).

Zimek, Arthur, Erich Schubert, and Hans-Peter Kriegel (2012). "A survey on unsupervised outlier detection in high-dimensional numerical data". In: *Statistical Analysis and Data Mining: The ASA Data Science Journal* 5.5, pp. 363–387 (cit. on pp. 164, 165).

Zio, Enrico and Nicola Pedroni (2013). "Methods for representing uncertainty". In: *Cahiers de la Sécurité Industrielle, Fondation pour une Culture de Sécurité Industrielle* 2013-03 (cit. on p. 129).

Zitnick, C Lawrence and Piotr Dollár (2014). "Edge boxes: Locating object proposals from edges". In: *European conference on computer vision*. Springer, pp. 391–405 (cit. on p. 81).

Zou, Difan, Yuan Cao, Dongruo Zhou, and Quanquan Gu (2018). "Stochastic gradient descent optimizes over-parameterized deep relu networks". In: *arXiv preprint arXiv:1811.08888* (cit. on p. 108).

Previous work summary

Towards Autonomous PErceptual Systems

The objective of this document is to introduce the principle of Autonomous PErceptual System (APES) as an object of study.

The functionalities of artificial perception, in particular vision, have become both easier to design and more efficient through the use of a set of techniques and development environments grouped under the term "Deep Learning". They have reached a certain level of maturity making it possible to envisage their use for real or even critical applications.

The research direction proposed here is to provide perception with a certain degree of autonomy envisaged as a means of guaranteeing its reliability.

The introduction of such a property implies to reconsider the status of perception no longer as passive functionality but as an activity involving as explicit stakeholders the environment to be perceived but also the recipient of the perceptual products with which the system maintains a contractual relationship determining the nature of the expected service and the means to guarantee it.

The study of autonomous perceptual systems thus leads to a research program organized along three axes: the design of a perceptual activity articulating functional dynamics and learning processes, the development of an inherent intelligibility of the mechanisms of perception for monitoring, specifying or justifying their behavior, and the implementation of a general approach to guarantee their safe and controlled use.

Keywords :
COMPUTER VISION ; ARTIFICIAL INTELLIGENCE ; ACTIVE VISION ; MACHINE LEARNING ; EXPLAINABIIITY ; CERTIFICATION ; AUTONOMY

Vers des systèmes perceptifs autonomes

L'objectif de ce mémoire est d'introduire le principe de système perceptif autonome comme objet d'étude.

Les fonctionnalités de perception artificielle, en particulier de vision, sont devenues à la fois plus faciles à concevoir et plus performantes par l'utilisation d'un ensemble de techniques et d'environnements de développement regroupés sous l'expression apprentissage profond "Deep Learning". Elles ont atteint un certain niveau de maturité permettant d'envisager leur utilisation pour des application réelles voire critiques.

La direction de recherche proposée ici est de munir la perception d'un certain degré d'autonomie considéré comme moyen de garantir sa fiabilité.

L'introduction d'une telle propriété implique de reconsidérer le statut de la perception non plus comme fonctionnalité passive mais comme une activité impliquant comme parties prenantes explicites l'environnement à percevoir mais également le destinataire des produits perceptifs avec lequel le système entretient une relation contractuelle déterminant la nature du service attendu et les moyens de le garantir.

L'étude des systèmes perceptifs autonomes conduit ainsi à un programme de recherche organisé selon trois axes: la conception d'une activité perceptive articulant dynamique fonctionnelle et processus d'apprentissage, le développement d'une intelligibilité propre des mécanismes de perception pour surveiller, spécifier ou justifier leur comportement, et la mise en œuvre d'une démarche générale permettant de garantir leur utilisation sûre et maîtrisée.

Mots-clés :
VISION ARTIFICIELLE ; INTELLIGENCE ARTIFICIELLE ; VISION ACTIVE ; APPRENTISSAGE AUTOMATIQUE ; EXPLICABILITE ; CERTIFICATION ; AUTONOMIE

ONERA
THE FRENCH AEROSPACE LAB

BP 80100 - Chemin de la Hunière FR-91123 PALAISEAU CEDEX - Tél. : +33 1 80 38 60 60 - Fax : +33 1 80 38 65 10

www.onera.fr