Contributions et perspectives pour l'amélioration de l'acceptabilité de l'apprentissage profond

Adrien Chan-Hon-Tong

Mémoire pour l'obtention de l'Habilitation à Diriger des Recherches

SORBONNE UNIVERSITÉ
Spécialité
Sciences pour l'Ingénieur

Présentée par

Adrien CHAN-HON-TONG

soutenu le 09/12/2022

Titre :

Contributions et perspectives pour l'amélioration de l'acceptabilité de l'apprentissage profond

Devant le jury composé de :

Catherine ACHARD	Professeure, Sorbonne Université	Examinatrice
Clément MALLET	Directeur de recherche, IGN	Examinateur
Teddy FURON	Directeur de recherche, INRIA	Rapporteur
Fabien MOUTARDE	Professeur, Mines ParisTech	Rapporteur
Sébastien LEFEVRE	Professeur, Université Bretagne Sud	Rapporteur

Remerciements

J'ai beaucoup de personnes à remercier pour cette HDR.

D'abord je remercie mes doctorants Matthieu Nugue, Guillaume Vaudaux-Ruth, Gaston Lenczner, Magdeleine Airiau et Pol Labarbarie. Je remercie aussi leurs directeurs/directrices de thèse : Guy Le Besnerais, Bertrand Le Saux, Stéphane Herbin et Catherine Achard pour m'avoir mis le pied à l'étrier de l'encadrement de doctorants.

Je remercie aussi l'équipe IVA de l'ONERA (notamment Elise Colin, Stéphane Herbin, Aurélien Plyer, Pierre Fournier, Maxime Chareyre, Guy Le Besnerais, Martial Sanfourche, mes doctorants et mes stagiaires) pour la très bonne ambiance propice à un travail scientifique de qualité.

Je remercie ensuite à nouveau ma directrice de thèse Catherine Achard de m'avoir poussé à rédiger cette HDR.

Je remercie l'ONERA (notamment Philippe Bidaud) de m'avoir accordé l'équivalent d'un mois de travail pour m'aider à la rédaction de cette HDR (cela fut peu mais indispensable). Je remercie Hélène Piet Lahanier et surtout Benjamin Pannetier pour la relecture d'une première version.

Je remercie enfin le jury pour avoir accepté de sanctionner ce travail.

Mais surtout je tiens à remercier ma femme qui m'a aidé à chacune de ses étapes. Merci pour les nombreuses relectures d'articles passés. Merci de m'avoir poussé à rédiger et soutenir cette HDR dès cette année. Notamment, merci pour les quelques vacances sacrifiées sur l'autel de la rédaction de ce manuscrit. Merci aussi pour les nombreuses relectures du manuscrit et pour les répétitions de la soutenance. Merci pour tout.

Introduction

En 2012, apparaissait ce que nous connaissons aujourd'hui sous le terme générique d'apprentissage profond ou de *deep learning* en anglais. Cette technologie, qui devient visible avec la victoire du réseau de neurones convolutif *Alexnet* [162] au challenge ImageNet [174] de reconnaissance par ordinateur, est rapidement vécue comme une véritable révolution : l'explosion phénoménale du nombre de publications comportant les mots clés *apprentissage profond* qui passe (selon [69]) de 4000 en 2011 à plus de 20000 dès 2015 en atteste. Cet emballement de la communauté scientifique est accompagné d'attentes très élevées en termes de changements industriels et sociétaux tant le champ des applications possibles est large et l'envol des performances impressionnant.

En 2022, dix ans après, le rythme des publications ne semble pas faiblir. De nombreux domaines d'application cherchent à s'approprier cette technologie : voitures et véhicules autonomes, diagnostics médicaux, analyses d'images satellites à des fins militaires ou pour des enjeux économiques... Cependant, bien que de nombreuses preuves de concept aient démontré le possible transfert de l'apprentissage profond à ces multiples domaines d'application, l'apprentissage profond n'a pas encore atteint aujourd'hui notre vie de tous les jours que ce soit pour notre santé, nos loisirs ou dans les laboratoires de recherche.

Au sein de l'ONERA, établissement fortement multi-disciplinaire, ma mission est d'appliquer ces méthodes d'apprentissage profond sur des données issues d'autres domaines scientifiques (ou des données de défense militaire). J'anime ainsi des projets autour de ce mot clé *apprentissage profond* au sein de mon établissement depuis 2015. J'ai ainsi pris part à des réalisations variées dans les domaines d'études de matériaux, de traitement d'électroencéphalogrammes, d'imageries médicales, de mécanique des fluides, d'observations de la Terre et de défense. Mais, cette diversité de champs d'application (et de questions posées dans chacune de ces collaborations) m'a aussi permis d'observer que l'acceptabilité de cette technologie n'est pas acquise. En effet, quand ces algorithmes ne sont pas indispensables, leurs inconvénients sont souvent rédhibitoires. Si en vision par ordinateur, les gains de performances sont tels que personne ne peut faire l'impasse sur l'apprentissage profond, dans d'autres domaines, ces gains peuvent être plus modestes et donc ne pas compenser leurs inconvénients.

Aujourd'hui, au delà de toujours plus de performance, deux obstacles me semblent majeurs pour permettre à l'apprentissage profond d'être plus acceptable. L'enjeu est, d'une part, de donner confiance en un outil qui cumule de nombreux défauts : la non reproductibilité des apprentissages, les problématiques de stabilité, le manque de robustesse, le manque d'équité, le manque d'explicabilité, le manque de maîtrise globale, autant d'éléments qui donnent à ces algorithmes un aspect *boite noire*. Il s'agit, d'autre part, de rendre le coût d'annotations des bases de données, pré-requis essentiel, plus accessible compte tenu du bénéfice attendu.

La cohérence de mes recherches et des thèses que j'ai co-encadrées est ainsi d'essayer d'adresser l'ensemble de ces inconvénients pour apporter un soutien dans la mise en œuvre de ces algorithmes d'apprentissage profond notamment

dans l'écosystème ONERA. Dans ce manuscrit, je présenterai une synthèse [1] de ces travaux.

Tout d'abord, je présenterai un bref historique de l'apprentissage profond, illustré par la diffusion de ces méthodes à différents domaines d'application. Le cœur du manuscrit s'attache ensuite à ces deux limites majeures qui sont devenues mes axes de recherches principaux. En deuxième partie, je considérerai les problématiques dites d'IA de confiance et en troisième partie j'aborderai la problématique de coût d'annotation des bases de données.

Enfin, le manuscrit présente un ensemble de perspectives pour l'avenir de l'apprentissage profond ainsi que mon projet de recherche à 5 ans qui porte sur l'intérêt d'utiliser les caractéristiques spécifiques des problèmes associés à des lois physiques connues et observables dans le cadre de l'utilisation de l'apprentissage profond. Une des portes de sortie permettant d'obtenir des méthodes d'IA de confiance pourrait en effet être de proposer en entrée des problèmes mieux posés.

1. **Dans l'ensemble du manuscrit, les publications dont je suis co-auteur sont indiquées en *bleu* pour les distinguer de la littérature en *noir*. Ces publications sont d'ailleurs détaillées en annexe et non accolées à la bibliographie.**

Table des matières

Chapitre 1

La diffusion de l'apprentissage profond

1.1 La montée en performance sur Imagenet

En 2009, [174] introduit *Imagenet* une base de données annotées de plus de 1 million d'images de type *réseaux sociaux* et l'associe à une compétition mondiale visant à accélérer la recherche en vision par ordinateur. En 2010 et 2011, les meilleures performances mondiales sont [169] et [172] avec 72% et 74% de taux de classification correcte (5 prédictions autorisées). Ces deux méthodes sont basées sur des sacs de mots, avec, [169] travaillant sur l'encodage local de primitives images classique comme des histogrammes de gradient [186] et [172] utilisant des vecteurs de Fisher.

En 2012, [162] obtient 84% de taux de classification correcte avec un réseau de neurones convolutif à 8 couches *Alexnet*. C'est 10% de plus qu'en 2011 et c'est le point de départ de la popularisation de l'apprentissage profond. Pourtant, Alexnet est assez similaire à Lenet [201], un réseau de neurones convolutif des années 90. Les principales différences sont :

- L'utilisation d'une activation *relu* là où [201] utilisait une activation *scaled hyperbolic tangent*.
- La présence d'une régularisation basée sur la déconnexion de neurones - *dropout* en anglais.
- La présence d'une décimation spatiale - *pooling* en anglais - à la place d'un lissage dans Lenet.
- Une plus grande profondeur (plus généralement une plus grande taille à la fois en terme de profondeur et d'épaisseur).

Cependant, à part la taille, les autres *ingrédients* n'ont pas été introduits par Alexnet : le *relu* est souvent associé à [171], le *dropout* à [183] et le *pooling* à [176]. C'est donc la concomitance d'améliorations techniques et de la puissance de calcul permettant des réseaux plus profonds (et plus gros) qui conduit à cette augmentation de performance sur Imagenet.

algorithmes	taux de bonnes classifications (top5)
sac de vecteurs de Fisher (2011) [172]	74%
Alexnet (2012) [162]	84%
VGG (2013) [152]	90%
ResNet-101 (2016) [119]	94%
EfficientNet (2019) [67]	97%
Transformer (2020) [34]	97%

TABLE 1.1 – Quelques articles clés illustrant l'augmentation des performances de classification d'images sur Imagenet.

Cette augmentation des performances et de la profondeur des réseaux va se poursuivre notamment grâce à la technique de normalisation de paquet à la volée [119] - *batch norm* en anglais - qui permet de limiter le problème du *vanishing gradient*. En effet, si chaque couche atténue le gradient entrant, alors plus le réseau est grand, moins les premières couches reçoivent du gradient [139]. Inversement, la couche dite de normalisation de paquets est dynamique. Elle prend un paquet de données en entrée et le normalise pour qu'il ait une moyenne nulle et une variance de 1. Au final, [119] parvient à créer des réseaux ResNet à plus de 100 couches grâce à la *batch norm*.

L'optimisation des architectures devient alors un problème en soi. De nombreuses approches d'optimisation d'architectures - *Neural architecture search* en anglais - comme par exemple [132] ont alors tenté de produire des réseaux plus performants. Par exemple, [56] introduit des réseaux géants (cependant plutôt en terme de paramètres que de nombre de couches). De même, [67] introduit EfficientNet une famille d'architectures offrant des performances élevées même avec des réseaux relativement petits. Cette famille se définit par un seul paramètre qui couple champ récepteur, largeur et profondeur.

Récemment, l'état de l'art semble basculer vers des architectures dites *transformer* [34] (bien que des perspectives existent aussi sur la modernisation des architectures convolutives comme [3]). Ces architectures plongent les bouts d'images dans un espace latent puis traitent ces ensembles de mots comme une séquence. Cette approche est donc intellectuellement proche d'une méthode *sac de mots* des années 2010, sauf que le vocabulaire n'est pas créé avec un simple regroupement basé sur une distance mais il est optimisé par l'algorithme. L'absence de convolution et de *pooling* dans cette architecture (le voisinage spatial n'étant présent que dans l'utilisation de bouts d'images) est très surprenante. Cependant, les performances semblent être au rendez-vous : elles sont du même niveau que celles d'EfficientNet alors que ce type d'architecture n'en est qu'à son début là où les réseaux convolutifs sont très aboutis. Ce bref historique est résumé dans la table 1.1.

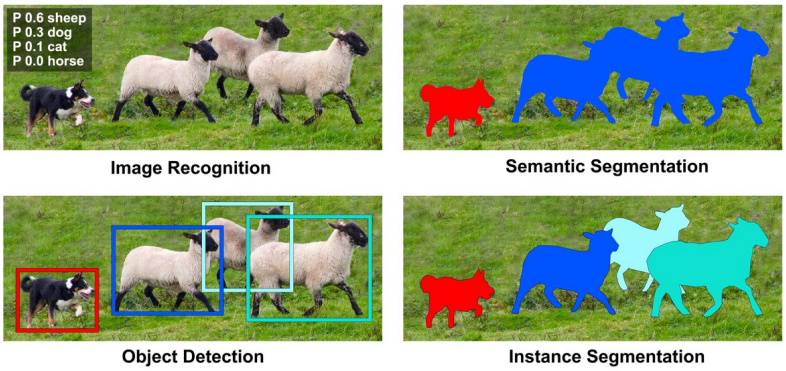


figure extraite de ai-pool.com/d/could-you-explain-me-how-instance-segmentation-works

FIGURE 1.1 – Illustrations de problèmes dérivés de la classification d'images (en haut à gauche) : la segmentation sémantique en haut à droite (chaque pixel a une classe), la détection d'objets en bas à gauche (les objets ont une emprise spatiale) ou la segmentation d'instances qui combine les 2.

1.2 Des performances transférables

En réalité, les améliorations de performance rappelées en table 1.1 n'auraient sans doute pas provoqué un aussi gros changement en vision par ordinateur, si ces améliorations n'étaient pas transférables à d'autres problèmes et/ou d'autres types d'images que celles d'ImageNet c'est-à-dire globalement des images de réseaux sociaux.

Mais, l'augmentation des performances en classification d'images conduit, d'une part, presque mécaniquement à une augmentation de performance dans des problèmes dérivés comme la détection d'objets ou la segmentation sémantique (illustrés dans la figure 1.1). En effet, la plupart des architectures traitant ces problèmes sont construites autour d'un réseau de neurones de classification jouant le rôle d'encodeur (l'archétype étant Unet [141] construit avec deux VGG [152] mis en regard).

D'autre part, ces techniques d'apprentissage profond ont démontré leur efficacité sur d'autres types d'images. Les modèles ImageNet eux-même sont souvent transférables. Enfin, il est possibles de poursuivre un apprentissage à partir de poids préexistants. Ce processus de poursuite d'apprentissage - *finetuning* en anglais - est une idée qui est apparue très tôt après Alexnet (exemple [158]).

On peut ainsi traiter des images médicales [118] via un apprentissage complet mais aussi (notamment dans un contexte où peu de données sont disponibles) directement à partir d'un réseau préappris, ou mieux, en adaptant ce dernier par *finetuning*. De même, on peut ainsi traiter des images de télédétection [114], des images de conduite autonome [113], des images sous marines [52]... Cette transférabilité (directe ou après poursuite de l'apprentissage) n'a pas été fructueuse avec les méthodes antérieures à l'apprentissage profond comme [172].

Ainsi, très rapidement l'apprentissage profond s'est diffusé dans tout le traitement d'images puis globalement dans de nombreux autres domaines industriels ou académiques.

1.3 La diffusion dans l'ONERA

Au sein de l'ONERA, j'ai participé activement à cette diffusion. J'ai notamment coordonné le montage et les travaux du projet DELTA (projet interne ONERA) dont l'objectif était d'aider des chercheurs ONERA (d'autres domaines que celui de l'apprentissage par ordinateur) à prendre en main ces nouveaux outils d'apprentissage profond que ce soit d'un point de vue méthodologique mais aussi d'un point de vue logiciel et hardware. Ce projet a donné lieu à un ensemble de réalisations présenté ci-dessous. Ces réalisations doivent évidemment se voir comme des exemples de cette vague de travaux plus générale qui a bouleversé l'état de l'art.

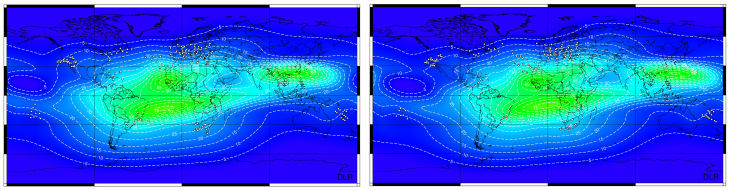
1.3.1 Apprentissage profond et télédétection

La première application de l'apprentissage profond de Imagenet vers des problématiques ONERA est, sans surprise pour cette introduction, une application à la télédétection. Étonnamment, ce type d'application n'a pas été immédiat dans la littérature, et le travail ONERA [235] de 2015 soit 3 ans après Alexnet fait néanmoins partie de la vague des premiers travaux à appliquer des approches d'apprentissage profond pour des tâches de segmentation sémantique d'images de télédétection (on peut citer [164] dès 2012 mais la rupture en termes de nombre de publication associant apprentissage profond et télédétection démarre plutôt en 2015).

Une explication possible est que le domaine de la télédétection a connu une révolution concomitante avec l'augmentation des volumes de données ouvertes et l'augmentation de la résolution. Or, l'apprentissage profond ne devient réellement pertinent que dans ce contexte. En effet, comme le remarque [35] et la figure 1.2, il n'est pas forcément pertinent de prendre en compte le voisinage spatial pour faire de la segmentation sémantique à basse résolution : dans ce cas, le problème revient principalement à classer chaque pixel indépendamment. Or pour cette tâche, des méthodes d'apprentissage classiques (comme le *boosting* [134]) sont compétitives avec des réseaux profonds. Par ailleurs, l'utilisation *d'a priori* physiques comme de simples indices construits sur des différences de bandes spectrales est très appréciée pour la classification de pixels dans la communauté.

Indépendamment, le projet DELTA a aussi permis de travailler sur des problématiques d'observation de la Terre. Il a ainsi conduit à des travaux sur l'étude de l'ionosphère [92] illustrés en figure 1.3 ou sur la prévision du risque foudre [208]. Dans ces applications, il est pertinent d'utiliser non seulement des réseaux convolutifs mais aussi récurrents [191]. L'état de l'ionosphère a un impact significatif sur certaines liaisons satellite-sol (dont GPS et communication). Aussi,

FIGURE 1.2 – Une même image de [235] à une résolution de 1m (Pléiade), de 10m (Sentinel2b) et 100m (Landsat) : l'utilisation de la texture spatiale n'a pas vraiment de sens à basse résolution.



Images extraites de https ://tec.hb9vqq.ch/

FIGURE 1.3 – Évolution de l'ionosphère. L'utilisation de réseaux de neurones récurrents pour capturer cette évolution a été investiguée dans [92].

être en mesure de prévoir cet état, pour estimer la disponibilité de communication est important tant pour des besoins civils que militaires, ce qui justifie l'existence de ces travaux à l'ONERA. Pour le risque orageux, il existe un besoin spécifique pour l'aviation (civile ou militaire) qui n'est pas forcément couvert par les prévisions météorologiques classiques, ce qui explique également que ce sujet soit présent à l'ONERA (alors que l'on pourrait penser que cette thématique serait plus pertinente chez Météo France).

Par ailleurs, même si cette activité n'est clairement pas dans le coeur de métier de l'ONERA, j'ai participé à un projet de transfert de technologie avec la startup VitaDx [1] justifié par la ressemblance (en terme de format, de constance d'échelle...) entre les images de cytologies urinaires et les images de télédétection (voir figure 1.4). Le diagnostic par ordinateur du cancer de la vessie à l'aide d'une cytologie urinaire de VitaDx a d'ailleurs reçu son marquage CE et est en cours de mise sur le marché. Ce projet m'aura d'ailleurs énormément appris sur les processus réglementaires en médecine.

1.3.2 Apprentissage profond et matériaux

D'autres applications, moins immédiates, d'apprentissage profond d'*ImageNet* vers des problématiques ONERA ont été réalisées notamment à des fins d'opti-

1. dont une trace tangible est le brevet data.inpi.fr/brevets/WO2019186073

FIGURE 1.4 – Illustration d'une image de télédétection et d'une image de cytologie urinaire : dans les deux cas, le volume de données est conséquent et la taille des objets est fixe. Il reste cependant une différence importante : une cytologie correspond à beaucoup de cellules mais néanmoins à 1 seul échantillon.

misations de matériaux aéronautiques.

Afin d'embarquer des matériaux dans un avion, il est pertinent de mesurer leur comportement vis-à-vis de stress. Un point particulièrement important est de connaître l'intensité des fissures en fonction d'une contrainte mécanique. L'ONERA a une expertise reconnue pour réaliser ce type de caractérisation notamment pour mesurer cette courbe *fissures vs tension*. Des équipes des départements matériaux conçoivent et mettent en oeuvre des protocoles expérimentaux pour générer et imager des fissures. Cela génère de grandes quantités de données expérimentales (par exemple, des images) dans lesquelles il faut ensuite venir extraire l'information utile (par exemple, le nombre de fissures).

C'est ainsi le cas avec un protocole de coupe et d'imagerie par microscopie. La plaque de matériel considérée y est étirée avec des cycles *élongation puis repos*. Durant les phases de repos le bord de la plaque est coupé mettant à nu les fissure internes. Cette coupe est alors imagée. La figure 1.5 illustre ce dispositif et les images qu'on peut en extraire. Appliquer des méthodes de segmentation sémantique à ces images pour extraire les fissures s'est révélé pertinent et original (dans la communauté *matériau*) [242]. On peut noter que compte tenu de la texture du matériau, les images sont difficilement traitables avec des approches *plus simples*.

Une réalisation relativement similaire a été obtenue sur des images de propergol dans le cadre de la thèse de Matthieu Nugue [63] dirigée par Guy Le Besnerais et dont j'ai rejoint l'encadrement notamment pour les aspects apprentissage profond. Il se trouve que des billes d'aluminium sont ajoutées au propergol utilisé pour les fusées (dont les fusées *Ariane*) car cela permet d'augmenter significativement l'efficacité des propulseurs. Cependant, ces particules d'aluminium peuvent aussi créer des instabilités en fonction de la distribution des tailles de gouttes dans l'écoulement. Or, malheureusement, des phénomènes d'agglomération et de fractionnement durant la combustion font que la taille des particules en combustion n'est pas déductible de la taille des billes ajoutées au propergol. Aussi, pour optimiser la dose d'aluminium, il est nécessaire de

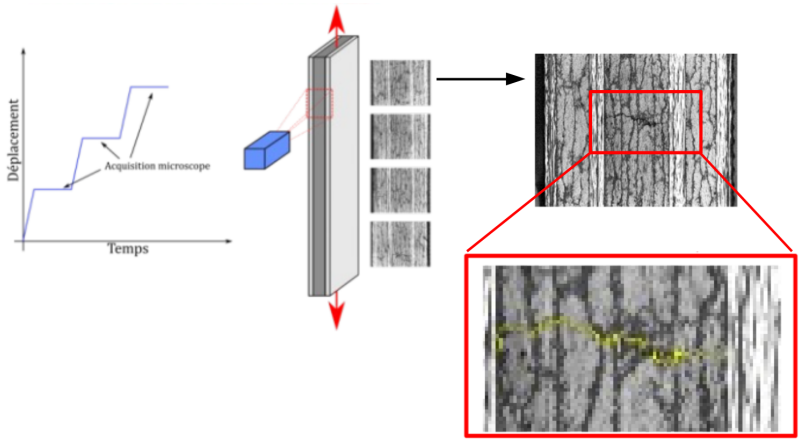


FIGURE 1.5 – Illustration d'un protocole mis en oeuvre à l'ONERA pour observer les fissures dans un matériau : la plaque est étirée pendant que la coupe est imagée (gauche), amenant à un ensemble d'images dans lesquels il faut trouver les fissures (en jaune en bas à droite).

coupler des mesures expérimentales des distributions de taille de gouttes et des simulations numériques. L'ONERA a ainsi développé un protocole d'imagerie dit de spectroscopie (qui consiste à imager la déformation de l'indice optique du milieu) illustré en figure 1.6 pour réaliser ces *mesures* dans et durant la combustion. Cela produit des *images* dont il faut extraire la distribution de taille des gouttes. Mais, là encore, l'utilisation de méthodes classiques comme MSER [185] pour extraire les particules se heurte à l'aspect des images qu'on obtient. Ainsi, utiliser des approches d'apprentissage profond de segmentation sémantique (voir figure 1.7) a apporté une plus value significative sur ce problème [242] (ce dont on reparlera au chapitre 2).

À noter que dans ces deux cas, l'apprentissage par ordinateur vient débloquer une situation où d'ingénieux protocoles expérimentaux produisent l'information

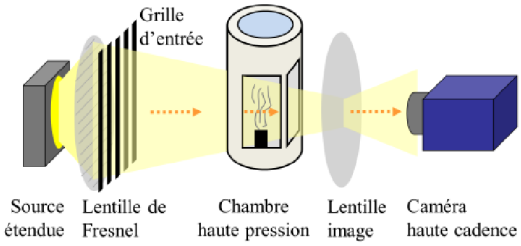


FIGURE 1.6 – Illustration d'un protocole mis en oeuvre à l'ONERA pour observer les particules d'aluminium dans un propergol en combustion.

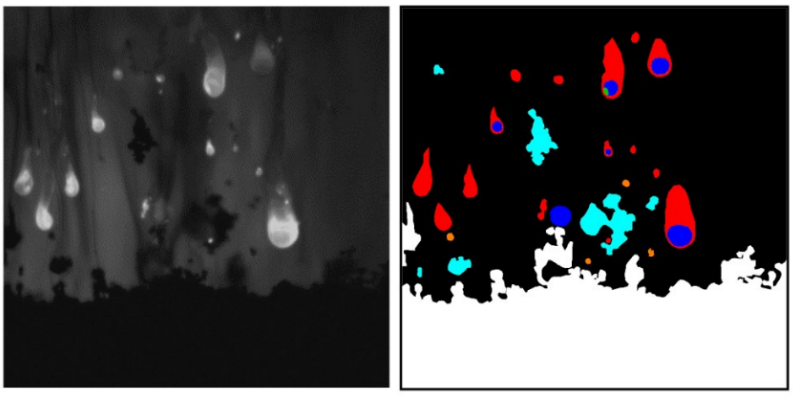


FIGURE 1.7 – Illustration des images issues de la captation du propergol en combustion (à gauche) et une vérité terrain (à droite) de ce problème vu comme un problème de segmentation sémantique (les gouttes en bleu - l'alumine en vert - les flammes en orange - le propergol en blanc ou bleu clair).

voulue mais où cette information est noyée dans un important volume de données rendant l'extraction de l'information voulue impossible *à la main (ou plutôt à l'oeil)*. Ces méthodes d'apprentissage profond ont justement le potentiel pour extraire ces informations à large échelle.

1.3.3 Apprentissage profond et données non images

D'autres méthodes d'apprentissage profond notamment des réseaux récurrents ont aussi été utilisées pour traiter des données d'EEG [210]. En effet, l'ONERA s'intéresse à l'étude des signaux physiologiques avec comme objectif d'être capable d'avoir des informations sur l'état cognitif d'un pilote. Dans [210], l'objectif était de déterminer si la personne avait regardé à gauche ou à droite.

Ensuite, ces approches d'apprentissage profond s'étendent à des problèmes de renforcement notamment en robotique [223, 222] (sur lequel je reviendrai en chapitre 4).

Enfin, d'autres travaux inclus dans DELTA ont concerné des signaux radars, le traitement de nuages de points [27], l'estimation de vorticité dans un fluide [22] ou encore des travaux de super résolution de champs en mécanique des fluides [47] (voir figure 1.8).

1.4 Des réussites mais aussi des limites

Ces applications de l'apprentissage profond sont *communes* : elles sont représentatives d'une vague de travaux de la communauté visant à appliquer ces nouvelles méthodes à toutes sortes de thématiques liées au traitement de l'image et/ou du signal.

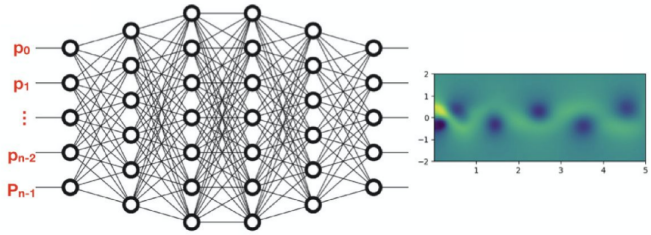


FIGURE 1.8 – Illustration d'un travail [47] porté par le projet DELTA en mécanique des fluides : l'objectif est de reconstruire un champ de pression 2D dense (à droite) en utilisant uniquement des mesures de pression 1D locales mais à hautes fréquences (à gauche).

Elles montrent l'enthousiasme de nombreuses communautés scientifiques vis-à-vis de ces approches : il parait rare de voir un socle méthodologique conduire à des publications en matériau, mécanique des fluides, télédétection et électroencéphalogramme en un temps court. Cependant, ces réalisations cachent en réalité des attentes partiellement voire peu comblées. En effet, dans ces domaines autres que la vision par ordinateur (et plus généralement le traitement du signal et du langage), ces algorithmes suscitent finalement de fortes réticences de partleur durée d'entraînement, leur manque de robustesse, leur aspect *boite noire*[2].

Personnellement, la contre-partie d'avoir pu participer à ces diverses réalisations est d'essayer d'augmenter l'acceptabilité de ces méthodes, et notamment en essayant de maintenir une expertise sur chacune de leurs limites (plutôt qu'une excellence sur quelques unes). Précisément, trois grandes limites s'illustrent : les temps de calcul (une discussion sur le sujet est renvoyée en annexe 5.4.1), la problématique de la constitution de base de données (qui est discutée en chapitre 3) et le côté *boite noire*.

Ce côté *boite noire* est intéressant car il limite non seulement l'acceptabilité de ces méthodes vis-à-vis d'autres domaines scientifiques mais surtout vers le grand public. En effet, bien que l'apprentissage profond ait été popularisé, il n'a pas conduit à des réalisations industrielles majeures en dehors du numérique : alors que les performances semblent au rendez-vous, le passage à la voiture réellement autonome reste lointain principalement pour des raisons d'acceptabilité. Cela justifie que le prochain chapitre soit orienté sur ces problématiques dites *d'IA de confiance*.

2. Cet aspect *boite noire* est double : il souligne le manque de maîtrise globale vis-à-vis de cette technologie, et, le fait qu'étant fixée une donnée, il n'est pas simple de décortiquer les mécanismes qui conduisent à la décision.

Chapitre 2

L'IA de confiance

La limite la plus saillante de l'acceptabilité des méthodes d'apprentissage par ordinateur (et en particulier des méthodes d'apprentissage profond) est la réticence à appliquer ce type d'algorithmes dans des contextes *critiques* c'est-à-dire dans des contextes où une erreur de l'algorithme provoque des morts. Pourtant, c'est aussi dans ces contextes critiques que ces algorithmes pourraient justement sauver des vies en prenant de bonnes décisions. Il est donc pertinent d'analyser les obstacles à ce type d'application et éventuellement d'essayer d'améliorer ces algorithmes pour qu'ils puissent y être appliqués.

Ainsi, ce chapitre propose de présenter les réticences à ces méthodes [1] et la réponse globale de la communauté pour essayer de les dépasser tout en se focalisant sur mes travaux et/ou encadrements.

2.1 Des réticences fondées

2.1.1 Quelques exemples saillants

— La haute autorité de santé (la HAS) considère que la performance d'un médicament peut être évalué par un essai clinique randomisé mais pas un dispositif médical contenant de l'IA. Elle a ainsi émis une grille d'évaluation [2] imposant d'autres contraintes à ces systèmes.

— Dans l'aviation, il y a une *tolérance* de 10^{-9} crash par heure de vol. Mais, cette tolérance ne concerne que les pannes matérielles. Inversement, aujourd'hui un algorithme critique (DAL-A) dans un avion ne doit pas produire de défaillance du tout (et doit être développé suivant la réglementation DO-178). Ainsi, un crash provoqué par un boulon qui casse (tolérance de 10^{-9}) est plus acceptable qu'un crash provoqué par

1. Ce mot n'est pas utilisé de façon péjorative : la suite insiste sur l'existence de raisons scientifiques qui invitent à la prudence.

2. dataanalyticspost.com/grille-evaluation-dispositifs-medicaux, j'ai personnellement participé à l'élaboration de cette grille via une invitation du LNE.

Cette figure utilise une image googlemap (Charleston, États Unis).

FIGURE 2.1 – Une sortie du détecteur [230] : Sur cette image, la précision est de 100% (toutes les détections sont exactes) et le rappel est de 90% (10 avions détectés sur les 11) ce qui correspond à un excellent F-score de 95%. Cependant, aucune *raison apparente* n'explique que l'avion raté le soit, ce qui peut être inacceptable pour les utilisateurs indépendamment du niveau de performances.

un algorithme (tolérance 0). Cette position était pertinente car les algorithmes qui étaient intégrés jusqu'alors pouvaient être *sans erreur* (et notamment formellement vérifiés). Mais, elle ne l'est plus forcément avec les algorithmes d'apprentissage (je participe à un groupe de travail sur la question [239]).

- Dans le domaine de la défense, il existe des missiles dont le guidage terminal est basé sur des algorithmes de vision par ordinateur [3]. Pour autant, l'utilisation d'algorithmes d'apprentissage profond sur ces mêmes systèmes ne pourra pas être validée de la même façon.
- L'Union Européenne, à la traîne dans le développement de ces méthodes d'apprentissage profond, tente de prendre les devant sur leur réglementation [4] et laisse transparaître que ces méthodes d'apprentissage ne seront pas traitées comme d'autres systèmes.
- Enfin, les réticences sont aussi présentes chez le grand public [53].
- De façon plus anecdotique, j'ai pris la mesure de ces réticences suite à des retours assez négatifs d'interprètes images de la défense sur le produit de détection présenté en figure 2.1 : alors que détecter 10 avions sur 11 était sur le papier largement au dessus de l'attendu, le fait d'avoir des erreurs inexplicables était quasiment rédhibitoire.

Vu de loin, l'idée qu'un médicament puisse être évalué via un essai clinique randomisé mais pas un algorithme peut sembler irrationnelle. Mais ce n'est pas le cas : les algorithmes d'apprentissage profond ne vérifient pas les hypothèses

3. Cela n'a évidemment rien de confidentiel : même wikipédia le dit fr.wikipedia.org/wiki/SCALP-EG

4. eur-lex.europa.eu/legal-content/EN/TXT/ ?uri=CELEX%3A52021PC0206

tacites sur lesquelles repose la procédure réglementaire de validation d'un médicament. Cette affirmation est précisée après une formalisation de ces notions.

2.1.2 La classification supervisée

Le problème de la classification d'images sur ImageNet entre dans le paradigme de la classification supervisée. On y accepte l'hypothèse qu'il existe une fonction y (inconnue mais accessible à l'aide d'un humain) de X un espace de données vers Y un ensemble discret de classes typiquement $Y = \{-1, 1\}$ s'il s'agit de séparer 2 sous-populations de X. Par exemple, X peut être l'ensemble des images, et y la fonction qui indique si une image est une image de chat. Un humain est tout à fait capable d'évaluer cette fonction, alors, que toute définition formelle semble vaine puisque *être une image de chat* n'est pas une propriété de l'image seule : le concept de *chat* n'a de réalité qu'en fonction *de la culture de celui qui regarde l'image.*

Justement, l'idée de la classification supervisée est d'être capable d'approximer y, sans chercher à trouver une règle exhaustive mais uniquement à l'aide d'exemples annotés (c'est-à-dire qu'on a montré à un humain pour avoir la valeur de y) $(x_1, y_1 = y(x_1)), ..., (x_N, y_N = y(x_N)) \in X \times Y$. Pour cela, on utilise un algorithme d'apprentissage $\mathcal{A}$ qui utilise la base d'apprentissage pour produire un modèle $f \sim \mathcal{A}((x_1, y_1), ..., (x_N, y_N))$ dont on voudrait que son signe soit une bonne approximation de y.

Plus formellement, une instance de classification supervisée est donnée [5] par une fonction y et une distribution P sur X selon laquelle sont tirés les exemples $x_1, ..., x_N$ de façon identiquement distribués. L'erreur réelle se définit alors par la proportion (pondérée par la densité) des points x où $y(x) \neq \text{sign}(f(x))$ c'est-à-dire

$$\text{erreur}_{\text{reelle}} = \int_{x, y(x) \neq \text{sign}(f(x))} P(x)\mathrm{d}x = \int \mathbf{1}_-(f(x)y(x))P(x)\mathrm{d}x$$

où $\mathbf{1}_-$ est la fonction qui vaut 1 pour les nombres négatifs incluant 0 et 0 sur les nombres strictement positifs i.e. $\forall x \in \mathbf{R},\ \mathbf{1}_-(x) = \begin{cases} 1 & \text{si } x \in]-\infty, 0] \\ 0 & \text{si } x \in]0, \infty[\end{cases}$.

2.1.3 Des garanties de performance

Ce paradigme étant posé, il est possible de dériver un certain nombre de bornes statistiques sur cette erreur réelle (qu'on ne peut pas calculer car on ne connaît ni y ni P). Par exemple, puisque la probabilité d'avoir $y(\chi) = \text{sign}(f(\chi))$ sur un échantillon χ tiré selon P est par définition 1-l'erreur réelle, la probabilité de n'observer aucune erreur sur une base de test de K échantillons $\chi_1, ..., \chi_K$ tirée selon P est $P(\forall k, y(\chi_k) = \text{sign}(f(\chi_k))) = (1 - \text{erreur}_{\text{reelle}})^K$.

5. Il est aussi possible de ne pas supposer l'existence d'une fonction y mais seulement d'une distribution jointe sur $X \times Y$. L'erreur est alors $\int_x P(x, -\text{sign}(f(x)))\mathrm{d}x$.

Plus généralement, on introduit l'erreur empirique comme

$$\text{erreur}_{\text{empirique}} = \frac{|\{k \in \{1,...,K\}, y(\chi_k)f(\chi_k) \leq 0|}{K} = \sum_{k\in\{1,...,K\}} \mathbf{1}_{-}(f(x_k)y(x_k))$$

, et on peut établir [187] que : pour toute distribution P et $\forall \delta \in]0,1]$ et si la base de test de taille K est tirée de façon identiquement distribuée selon P alors

$$P\left(\text{erreur}_{\text{reelle}} \leq \text{InvBin}\left(K, \text{erreur}_{\text{empirique}}, \delta\right)\right) \geq 1 - \delta$$

où $\text{InvBin}(K, e, \delta) = \max_p \{p,\ \text{Bin}(K, e, p) \geq \delta\}$ est l'inverse du binomial donné par $\text{Bin}(K, e, p) = \sum_{j=0}^{e} \frac{K!}{j!(K-j)!} p^j (1-p)^{K-j}$. Alternativement, [187] donne aussi la formulation :

$$P\left(\text{erreur}_{\text{reelle}} \leq \text{erreur}_{\text{empirique}} + \sqrt{\frac{-\log(\delta)}{2K}}\right) \geq 1 - \delta$$

Ces travaux connaissent évidemment de très nombreuses extensions qui ne seront pas plus détaillées ici (borne de Vapnik [193], bornes PAC spécifiques aux classificateur à vaste marge [184], [1] pour des réseaux de neurones, l'erreur réelle sous attaques adverses bornées en norme [21]...).

2.1.4 Les limites de ces garanties

Ces bornes sont évidemment intéressantes car elles donnent des garanties sur l'erreur réelle incalculable à partir de quantités observables comme l'erreur empirique. *Pourquoi alors ne pourrait-on pas évaluer un algorithme comme un médicament ?*

Le problème est que ces bornes supposent que le tirage de la base de test est identiquement distribué selon P (sinon, on calcule l'erreur du problème y, Q où Q est la distribution du tirage et pas celle de y, P). Mais, dans la plupart des applications industrielles, on ne pourra pas tirer des échantillons identiquement distribués notamment parce que les campagnes d'essais mettent en oeuvre du matériel dont la disponibilité est limitée, ou, parce que les distributions de données sont changeantes (*distribution drift* en anglais). Même en médical, la réglementation acte que *les patients arrivant dans un hôpital forment un tirage identiquement distribué* mais cette décision est plus réglementaire que scientifique. Dans un contexte de véhicule autonome, imaginons un détecteur de panneaux *stop*. L'option pratique sera de l'évaluer en imageant un certain nombre de panneaux *stop* au sein d'une ville ou d'une région lors qu'une campagne d'acquisition bornée en temps. Mais dans ce cas, la borne PAC ne s'applique pas. Il n'y a donc aucune garantie formelle que l'algorithme a la même performance ailleurs ailleurs (même si c'est probable, ce point est discuté en annexe). Inversement, il faudrait tirer les panneaux selon la fréquence de passage de voitures (qui est changeante) ou alternativement de façon uniforme (quitte à pondérer par la

fréquence) mais il faudrait aussi penser à tirer de façon uniforme en fonction de l'heure, de la saison, des conditions météo... Il en résulterait une campagne d'acquisition incompatible avec la pratique.

Ce problème qu'on ne peut **pas** tirer selon P n'est pas arrivé avec l'apprentissage profond. Cependant, et c'est pour cela que je parle *d'accord tacite*, la HAS pouvait *tolérer* l'approximation *hôpital=i.i.d.* pour évaluer un médicament. Mais, elle ne peut plus l'accepter pour des réseaux de neurones car l'impact de tels biais peut alors être significatif. Ainsi, les réticences présentées précédemment sont bien rationnelles : il est pertinent pour la HAS d'être prudente devant l'observation que les approximations habituelles ne sont plus valables pour ces algorithmes.

Je me permets une métaphore : les avions sont aujourd'hui mis sur le marché suite à une procédure qui ne tient pas compte de la relativité générale alors qu'on sait que la mécanique classique est *fausse*. Mais cela ne signifie pas que ces procédures sont insuffisantes : elles sont acceptables tant qu'on n'a pas affaire à un avion très très rapide. La montée en performance des algorithmes d'apprentissage par ordinateur (via l'apprentissage profond) c'est l'apparition d'avions très très rapides qui force à reposer la question des approximations qu'on fait dans les processus de validation.

On peut penser à cet autre exemple dans la défense : l'algorithme de ciblage du missile SCALP a été validé comme si on avait couvert tous les cas. Mais en réalité, on avait seulement couvert *assez* de cas pour avoir *confiance* pour un domaine d'emploi figé (mais sans être réellement exhaustif : on ne le peut jamais en vision par ordinateur). D'ailleurs, cette confiance est surtout apportée par le fait que cet algorithme est explicable par conception car il s'agit d'un *simple* appariement de segments. Mais, aujourd'hui, l'idée de mettre des algorithmes d'apprentissage profond dans ces mêmes systèmes remet en question l'ancienne définition de *assez*, dans un contexte d'algorithmes, qui sont aujourd'hui, *boite noire*. Pourtant, ces algorithmes modernes auraient une performance moyenne plus haute sur un domaine d'emploi plus large : ils sont plus performants que leurs prédécesseurs presque partout dans l'état de l'art de la vision par ordinateur.

Cette question de choisir les bonnes approximations qui garantissent à la fois une certaine sécurité des utilisateurs et une certaine faisabilité pour l'industriel est justement au coeur d'un effort considérable porté par la communauté. Cet effort résumé sous le concept de *l'IA de confiance* vise à reconstruire un consensus tacite sur la bonne façon de développer, évaluer et réglementer ces algorithmes.

2.2 Augmenter la maîtrise

Précisément, *l'IA de confiance* porte globalement sur le fait de mieux maîtriser et/ou comprendre ces algorithmes d'apprentissage (on a donc un objectif qualitatif). Cette maîtrise est supposée permettre une meilleur acceptabilité et la possibilité de définir - en confiance - les *bonnes* modalité de développement

et d'évaluation (sachant qu'en l'absence d'un tirage i.i.d. on n'a pas de garantie qualitative).

Ainsi, l'IA de confiance porte sur un objectif flou d'augmentation de la maîtrise de ces algorithmes d'apprentissage profond qui englobe énormément de travaux dont l'explicabilité, l'apprentissage frugal (avec peu de données), l'apprentissage hybride (couplant des approches d'apprentissage et des approches issues de modélisation), la détection d'anomalie, la confidentialité des données (*privacy* en anglais), les représentations internes, l'équité, la robustesse... Cependant, ces thèmes sont tous importants, et par ailleurs, tous plus ou moins inter-connectés entre eux. C'est pourquoi j'essaye d'avoir a minima une compétence (et si possible des contributions) sur la plupart. Bien entendu, cela se fait avec plus ou moins de succès, et le manuscrit s'étendra un peu plus sur certains des thèmes. Mais, la cohérence d'ensemble de mes travaux et/ou encadrements tient au fait d'essayer de traiter un peu de chacun de ces thèmes.

Notamment, je me permets de proposer la figure 2.2 pour illustrer la façon dont je vois les différentes interactions entre les problèmes de robustesse, d'équité et de détection d'anomalie qui forment des piliers des thèmes de l'IA de confiance (ce qui n'enlève rien au problème d'explicabilité, de confidentialité, d'hybridation ...). Ainsi, les sections suivantes cherchent d'une part à justifier cette illustration et d'autre part à montrer que j'ai cherché à avoir des co-contributions sur chacun de ces thèmes.

2.2.1 La robustesse

Dans le langage courant la *robustesse* est associée à la solidité, la résistance. Transposé à des algorithmes d'apprentissage par ordinateur, cela pourrait recouvrir de nombreux aspects : la capacité à traiter des données légèrement différentes, à être reproductible (un problème réel pour ces algorithmes profondément dépendant de l'aspect aléatoire de la descente de gradient stochastique et de l'initialisation des poids). Cependant, la notion que la communauté d'apprentissage par ordinateur met derrière le terme de robustesse correspond plutôt à la stabilité du modèle en fonction d'une perturbation locale de l'entrée.

Il se trouve que les réseaux de neurones profonds sont extrêmement vulnérables à certaines perturbations locales. Précisément, les réseaux de neurones *naïfs*, c'est-à-dire non entraînés spécifiquement contre ces perturbations y sont très sensibles. Ces perturbations spécifiques sont dites adversaires - *adversarial attacks* ou *evasion attacks* en anglais - car elles profitent spécifiquement de failles du modèle (elles sont généralement optimisées pour attaquer un modèle cible).

Formellement, étant fixée une norme $||.||$ et une borne ε, un réseau f (binaire pour simplifier) admet un exemple adversaire en x s'il existe δ tel que $||\delta|| \leq \varepsilon$ et $f(x+\delta)f(x) < 0$ ([5] remarque qu'il faut normalement aussi rajouter des contraintes de format, typiquement si $x \in \mathbb{Z}^D$, alors cette contrainte s'étend à $x+\delta$). Or, cela démontre un problème puisqu'on peut supposer que $y(x+\delta) = y(x)$ (sauf sur un ensemble de mesure nulle).

En réalité, ces attaques sont connues depuis a minima 2004 [188] et peuvent

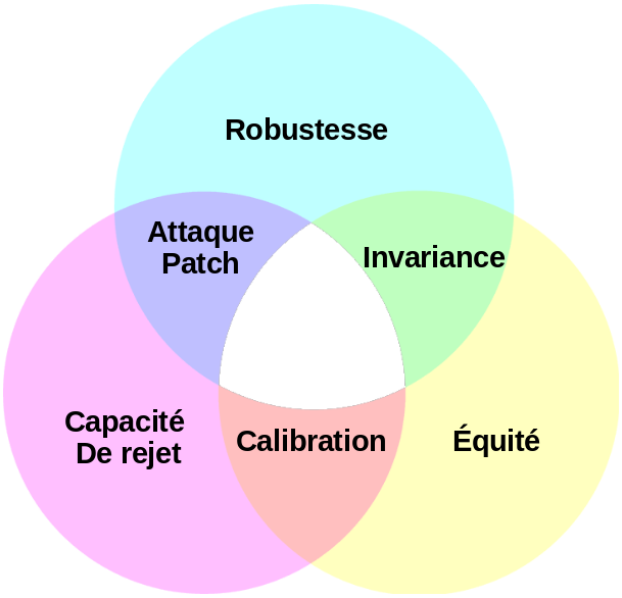


FIGURE 2.2 – Illustrations d'intersections de différents problèmes d'IA de confiance. Exemple : les problèmes de stabilité locale des prédictions dont l'archétype est la robustesse aux exemples adversaires intersecte avec les problèmes d'équité qui visent à une forme d'invariance c'est-à-dire une stabilité globale des prédictions que l'on retrouve dans la robustesse à des perturbations agnostiques ou des transformations géométriques. Cette cartographie n'est pas exhaustive car d'autres points émargent à l'IA de confiance comme l'explicabilité, la frugalité, l'hybridation, les bornes statistiques, les méthodes formelles ...

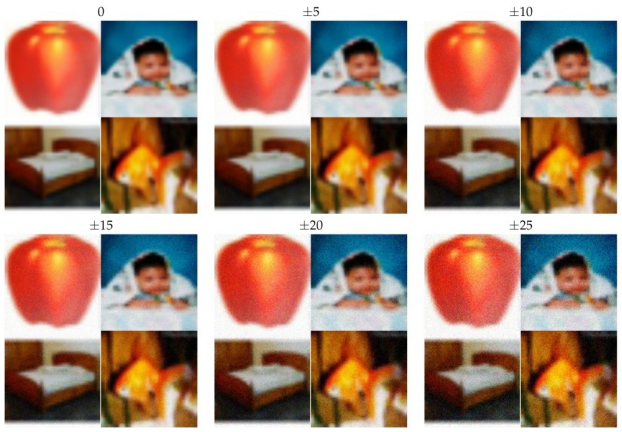


FIGURE 2.3 – L'illustration extraite de [211] de la sensibilité de l'oeil humain : une perturbation de moins de $10/255$ d'amplitude est déjà difficilement discernable, pourtant des perturbations adversaires de $3/255$ sont déjà capables de diminuer significativement la performance d'un réseau naïf.

toucher n'importe quel type d'algorithme d'apprentissage. Cependant, la facilité à attaquer un modèle de réseau profond et l'impact que peuvent avoir ces attaques n'étaient pas soupçonnés avant [153]. On sait aujourd'hui qu'il est possible de diminuer énormément la performance des réseaux (*naïfs*) à l'aide d'une perturbation **invisible pour l'oeil humain** [153] comme l'illustre la figure 2.3. Dit autrement, un problème des réseaux de neurones (naïf) est qu'ils admettent des exemples adversaires même pour une faible valeur de ε en norme infinie sur une grande proportion des points. Typiquement, un ResNet classique a 94% de prédictions correctes (top 5) sur Imagenet. Mais, [153] crée une perturbation invisible spécifique au réseau et à chaque image telle que le taux de prédictions correctes (top 5) sur l'ensemble des images perturbées ne soit plus que de 37%.

Évidemment, les points perturbés n'appartiennent pas à la distribution c'est-à-dire qu'on peut tout à fait avoir $f(x+\delta)f(x) < 0$ mais $P(x+\delta) = 0$. Donc, il est possible d'être sensible à ces attaques tout en ayant une erreur réelle très faible. Encore une fois, cela illustre la discussion autour de *l'IA de confiance*. D'une part, cela souligne les limites des bornes type PAC (bien que des extensions au cas adversaire existent) puisque l'erreur réelle pour la distribution actuelle peut être faible mais devenir importante pour une distribution *proche*. Au-delà, cela ouvre la question des changements de distribution qu'on pouvait tacitement penser négligeables.

D'autre part, cela souligne l'importance de la confiance *sociologique* indépendamment de la performance formelle. En réalité, ces attaques ne sont pas si problématiques : on commence à savoir s'en protéger (voir annexe 5.4.5) et elles sont généralement non réalisables dans le monde physique. Par contre, elles

démontrent que les réseaux de neurones n'apprennent **pas** comme on aurait pu espérer : ils n'apprennent pas comme nous puisque nous ne sommes même pas capables de percevoir ce qui va pour eux être significatif ! Or, ce point a eu un impact négatif significatif sur la confiance qu'on peut avoir en ces réseaux. Ainsi, construire des approches robustes n'est pas forcément nécessaire en terme de performance mais ça l'est en terme d'acceptabilité. Ce qui explique que ce thème pourtant *a priori* marginal vis-à-vis de l'apprentissage soit devenu très visible (nicholas.carlini.com/writing/2019/all-adversarial-example-papers.html répertorie plus de 5000 papiers sur le sujet).

À noter, depuis [153], de nombreux travaux s'intéressent au lien entre la sensibilité aux attaques adversaires et l'utilisation des *hautes fréquences* des images [57, 24, 2] plutôt que les formes des objets normalement portées par les basses fréquences (bien que cette assertion dépende de la résolution et bien que des exemples adversaires lisses soient constructibles [44]). J'ai travaillé sur ce thème [219] mais plus spécifiquement en lien avec l'empoisonnement de données décrit après.

2.2.2 L'équité

Si les attaques adversaires sont surtout liées aux réseaux de neurones profonds, les problèmes d'équité touchent l'apprentissage par ordinateur (et plus généralement les systèmes d'information). Ainsi [123] pointe le fait que les algorithmes (naïfs) tendent à produire des jugements biaisés (raciste, sexiste ...) et à pénaliser les minorités. Ces problèmes touchent aussi l'apprentissage profond [6], d'autant plus que l'aspect *boite noire* empêche d'avoir une maîtrise sur l'impact de telle ou telle caractéristique.

Il existe de multiples sous-problèmes associés au concept d'équité (biais encodés dans les données [105], biais de la population majoritaire [76], ...). Un de ces problèmes est celui des variables protégées [40] : on voudrait par exemple que le modèle résultant de l'apprentissage soit invariant à certaines variables (couleur de peau, sexe), sachant que supprimer ces variables n'est généralement pas suffisant (le modèle pouvant les reconstruire à l'aide du reste de la donnée : par exemple, le prénom permet de reconstruire approximativement le sexe).

Formellement, ce dernier problème peut se voir comme la volonté d'être invariant à des changements sur ces variables. Par exemple, si on considère que ces variables sont simplement un sous-ensemble I des composantes de la donnée $x \in \mathbb{R}^D$, alors, l'équité par rapport à ces variables consiste à (essayer de) garantir que : $\forall x, x',\ x_{\{1,\dots,D\}-I} = x'_{\{1,\dots,D\}-I} \Rightarrow f(x)f(x') \geq 0$. Cela dit, cette modélisation formelle est restreinte par rapport à l'esprit initial qui recouvre de nombreux autres problèmes d'équité.

Je co-encadre la thèse de Magdeleine Airiau qui porte notamment sur l'équité d'un détecteur vis-à-vis de la taille des objets. Cette thèse s'inscrit dans le même contexte applicatif de celle de Matthieu Nugue, c'est-à-dire dans la perception

6. www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai

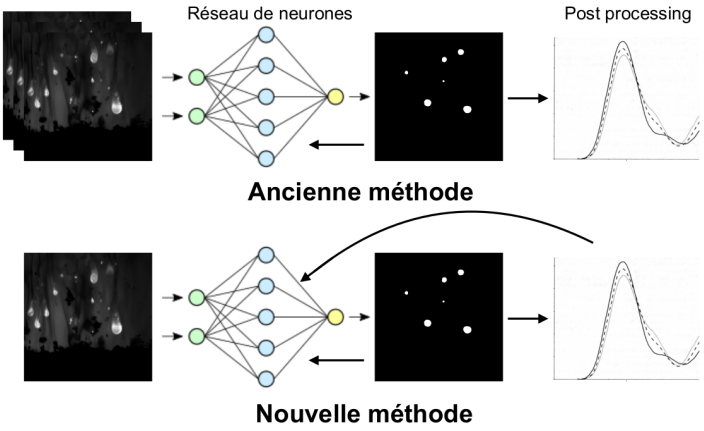


FIGURE 2.4 – Illustration d'une méthode visant à forcer le détecteur à avoir des performances similaires indépendamment de la taille des cibles : ici une fonction de perte qui compare la distribution de tailles prédites et la distribution de tailles réelles est ajoutée à la fonction de perte classique de détection.

appliquée à des vidéos de propergol en combustion. Comme indiqué au chapitre 1, il est critique pour améliorer les performances des propergols d'être capable de connaître la distribution des tailles de gouttes après agglomération.

Cependant, ce qui importe ce n'est pas d'avoir un détecteur de gouttes qui fasse peu d'erreurs mais bien un régresseur de distribution de tailles. Ainsi, on préférera un détecteur qui manque 1 goutte sur 2 indépendamment de la taille plutôt qu'un détecteur qui ne manque aucune grosse goutte mais 1 petite sur 2. En effet, la distribution est parfaitement restituée dans le premier cas et biaisée dans le second (alors même que l'erreur est plus basse). Ceci souligne l'importance d'avoir un détecteur qui soit équitablement performant entre les différentes sous-populations de gouttes. Or, les détecteurs de l'état de l'art tendent à être biaisés [37] en particulier par la taille. D'ailleurs dans le jeu de données de détection de MS COCO [150], les petits objets sont considérés différemment des autres.

Ainsi, les travaux de cette thèse s'intéressent (notamment) à augmenter l'équité de la détection relative à la taille des gouttes. Comme la performance de détection est ici moins importante que l'équité (ce qui ne sera pas forcément le cas dans d'autres contextes), l'option choisie pour renforcer l'équité est directement d'apprendre à l'algorithme à régresser la distribution de tailles de gouttes comme illustré par la figure 2.4. Plus précisément, une fonction de perte qui compare la distribution de tailles prédites et la distribution de tailles réelles est ajoutée à la fonction de perte classique de détection après une première phase de stabilisation.

Cette structure d'apprentissage permet de diminuer fortement la divergence entre les distributions estimées et prédites en test, comme illustré en figure 2.5.

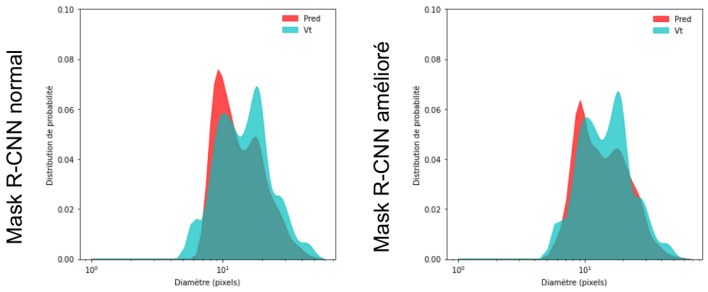


FIGURE 2.5 – Illustration de deux distributions de tailles de gouttes prédites par un détecteur appris classiquement à gauche et un détecteur appris avec une fonction de perte sur la divergence des deux distributions.

Cela correspond à une équité plus forte vis-à-vis de la taille des cibles.

Ces travaux ont donnée lieu à et devraient conduire à d'articles dans la communauté *propergol*.

Ces travaux sont assez spécifiques au cadre *propergol* mais ces problèmes d'équités ne sont pas que des problèmes éthiques abstraits. Dans CITYSCAPE [113], sur les 500 images de validation, on dénombre 3419 polygones *personne* [7] mais seulement 37 enfants (comptage manuel). Ainsi, alors que les enfants représentent 13% de la population allemande [8]), ils ne représentent que 1% des personnes dans CITYSCAPE. On voit donc qu'une campagne d'acquisition naïve peut être fortement biaisée (ici probablement car la présence d'enfants dans l'espace public est structurée autour de certaines heures). Certes dans le cas d'un véhicule autonome, on veut parfaitement distinguer petits et grands et pas juste détecter également les 2 sous-populations. Cependant, un grand piéton lointain apparaît petit mais sa distance fait qu'une détection ratée n'est pas critique alors qu'un petit piéton apparaît petit même s'il est près. Cette impact de la taille n'est donc pas tout à fait anodin.

2.2.3 L'invariance à des transformations images

La résistance aux attaques adversaires et d'équité par rapport à des variables protégées ne sont pas si différentes. Dans les deux cas, on cherche une invariance : locale dans le cas des attaques adversaires et spécifique à certaines variables protégées dans l'autre.

Il existe ainsi un continuum entre ces deux thématiques. Si une perturbation adversaire de forte amplitude est appliquée à une donnée, alors il n'y a aucune chance de pouvoir la traiter puisque cette perturbation peut détruire l'information. C'est pour cela qu'une attaque adversaire n'a de sens que si elle est de faible amplitude. Inversement, des bruits agnostiques de plus forte amplitude

7. Précisément, je ne compte que les polygones suffisamment au premier plan pour bien discerner s'il s'agit d'un adulte ou d'un enfant - mais il y en a au moins 1754.

8. https ://www.insee.fr/fr/statistiques/2867604

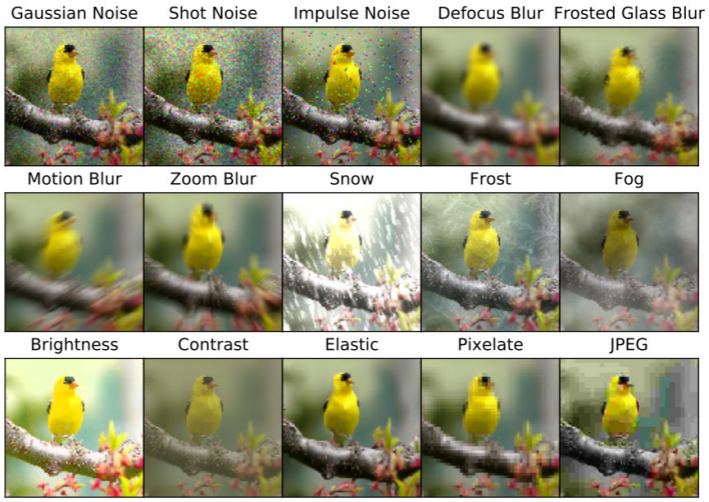


Figure 1: Our IMAGENET-C dataset consists of 15 types of algorithmically generated corruptions from noise, blur, weather, and digital categories. Each type of corruption has five levels of severity, resulting in 75 distinct corruptions. See different severity levels in Appendix B.

FIGURE 2.6 – Illustration extraite de [55] : la robustesse aux bruits communs n'entre pas dans le cadre des exemples adversaires car leur amplitude est bien supérieure. On est à mi-chemin entre la robustesse et la recherche d'invariance.

peuvent noyer l'information sans pour autant la détruire. Typiquement, dans la figure 2.6, l'humain est tout à fait capable de reconnaître l'objet même sous des bruits de cette amplitude. Inversement, les réseaux *naïfs* sont sensibles à de tels bruits : [55] montre par exemple que des bruits *communs* de forte amplitude diminuent fortement la performance des réseaux [9]. Mais ce résultat ne rentre pas dans le cadre de la robustesse au sens adversaire car la perturbation est de grande amplitude mais contrainte. Je me suis intéressé à ces différents problèmes de bruits agnostiques dans différents rapports non publiés comme hal.archives-ouvertes.fr/hal-01773170v3 (2019).

Dans le même esprit, l'invariance à la rotation sur une image de télédétection [112] ou l'invariance par translation [77] sont presque plus des problèmes d'équité que de robustesse.

2.2.4 Les mécanismes de rejet

Indépendamment de ces problèmes de robustesse et d'équité, [138] démontre que les réseaux naïfs pensent reconnaître des objets sur des images clairement

9. Précisémement, les réseaux sont plutôt robustes à des petits bruits blancs, ou des petites compressions jpg. Mais cette robustesse ne tient pas pour de forts bruits blancs, de fortes compressions etc. Inversement, l'humain reste très robuste même dans ces cas comme le montre la figure 2.6.

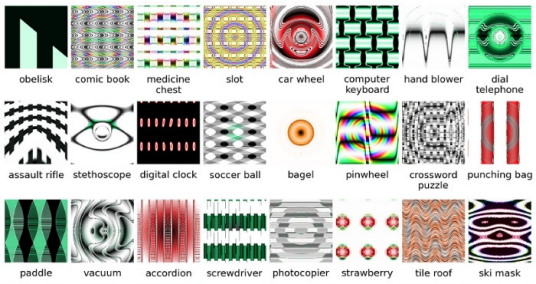


FIGURE 2.7 – Illustration extraite de [138] : les réseaux naïfs ne détectent pas comme étant hors distribution des images qui sont pourtant clairement non naturelles.

hors de la distribution comme illustré en figure 2.7.

Encore une fois, cela n'est nullement un problème dans le cadre du paradigme de classification supervisée : si les images sont hors de la distribution, on ne calcule pas l'erreur dessus, il n'est donc pas problématique de mal les classer.

Néanmoins, cela pose un problème de confiance. En effet, un opérateur aura plus confiance envers un système moins performant mais capable de détecter les cas qu'il ne sait pas traiter que devant un système sûr de lui même dans une situation totalement hors du domaine de fonctionnement [200, 161, 168]. Ainsi, il est indispensable d'être capable de doter les algorithmes d'une capacité à répondre qu'ils ne savent pas répondre. Par ailleurs, c'est aussi un problème de confiance, au sens où ces images illustrent que les représentations internes que le modèle se fait des classes ne correspond pas forcément aux représentations attendues.

Je n'ai actuellement pas de contribution sur ce thème bien que des discussions en ce sens soit en cours notamment dans un contexte d'interaction homme machine.

2.2.5 Les attaques adversaires par patch

L'illustration 2.7 qui montre que les représentations internes du réseau ne sont pas forcément adéquates ouvre la porte à des attaques adversaires non plus bornées en norme mais spatialement. En effet, [89] démontre qu'il est possible de créer physiquement un motif qui attire l'attention du réseau. Ce faisant, il introduit un nouveau type d'attaque illustré par la figure 2.8 : les attaques par patch. Dans cette figure, un patch vient exiter la représentation interne de la classe *grille-pain*. Ajouter ce patch dans une image (ici de banane) fait basculer la décision de la classe d'origine (*banane*) vers la classe *grille pain*.

Ici encore, on a affaire à des images hors distribution (mais malheureusement réalisables physiquement ce qui peut éventuellement poser de réels problèmes pour des applications critiques), et par ailleurs, le problème est philosophiquement moins grave qu'avec les attaques invisibles : en soit si on met un grille pain

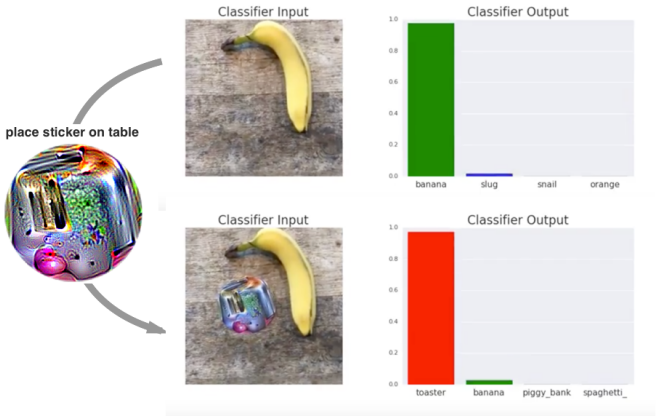


FIGURE 2.8 – Illustration d'une attaque par patch physiquement réalisable (image extraite de [89]).

et une banane dans une image, le classificateur ne fait pas vraiment d'erreur en choisissant 1 des 2. En ce sens, comme on *voit bien* qu'il y a un objet (même si c'est un objet bizarre), il n'est pas si étonnant que la classe puisse changer.

Sauf que ces patchs peuvent modifier si fortement la représentation globale de l'image qu'ils peuvent empêcher l'algorithme d'en avoir une vue d'ensemble. Par exemple, ils peuvent créer de vraies erreurs pour des taches de détection comme présenté en figure 2.9. Or, un tel comportement est problématique pour une application comme la voiture autonome puisqu'ils sont physiquement réalisables.

À noter, les méthodes formelle qui testent de façon exhaustive l'existence d'adversaire fonctionnent théoriquement aussi bien contre des adversaires invisibles que par patches (même si ces méthodes passent difficilement à l'échelle dans les 2 cas). Par contre il n'existe pas vraiment de méthode de défense efficace à ce jour. D'ailleurs, je co-encadre la thèse de Pol Labarbarie dirigée par Stéphane Herbin sur ce sujet.

2.3 Empoisonnement de données

Cette section est associée à 2.2.1 mais approfondie ici.

2.3.1 Empoisonnement vs Attaques adversaires

Si le thème des exemples adversaires a reçu un écho très fort dans la communauté, j'ai de mon côté plutôt étudié un autre aspect de la robustesse : l'empoisonnement des données - *data poisoning* en anglais.

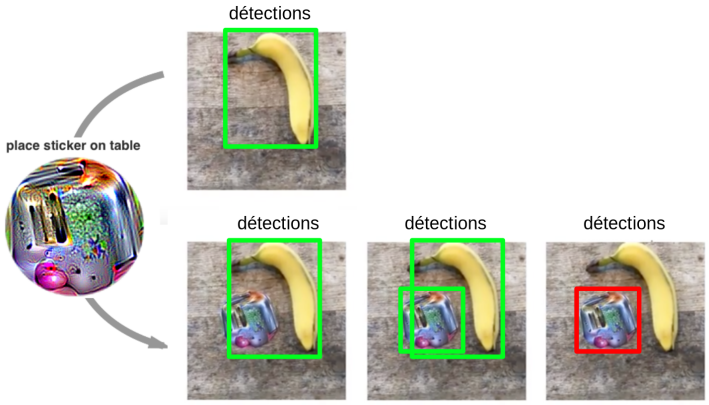


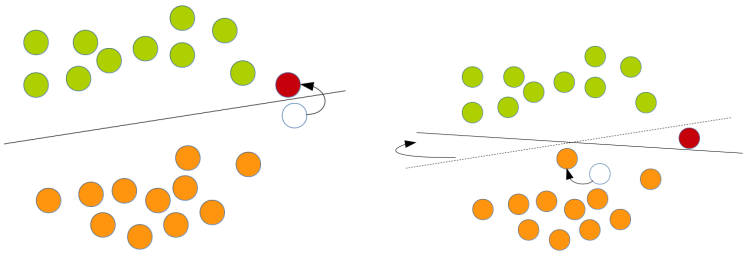
FIGURE 2.9 – Les attaques par patch sont philosophiquement moins problématiques que les attaques invisibles. Ainsi, il est acceptable que le système réagisse à l'objet introduit (en bas au milieu) même si la sortie en bas à gauche est préférable. Cependant, ces attaques peuvent empêcher le système de correctement réagir au reste de la scène (en bas à droite) ce qui est évidemment un problème pour une application comme le véhicule autonome.

Là où les exemples adversaires sont des attaques en production (le réseau est appris, et la modification concerne une image *de test*), l'empoisonnement de données est une attaque sur les données *d'apprentissage*. L'idée est de modifier les données d'apprentissage pour faire dévier le modèle induit vers un comportement spécifique. La figure 2.10 illustre cette différence.

En réalité, l'empoisonnement de données n'est pas apparu avec l'apprentissage profond [82]. Au contraire, les méthodes d'apprentissage classiques plus prédictibles sont peut être même plus fragiles face à ce type d'empoisonnement. En effet, il est alors éventuellement possible de savoir quel va être l'impact d'une modification de l'apprentissage sur le modèle résultant. L'exemple typique est le cas d'un classificateur à vaste marge (ou SVM [198]) sur un nuage de point linéairement séparable : imaginons que les points à séparer soient dans les matrices X, Y (chaque colonne de X étant un point d'apprentissage et Y le vecteur des classes - on omet les biais) alors pour toute matrice de rotation R, on a simplement $SVM(R(X), Y) = R(SVM(X, Y))$ i.e. la rotation commute avec l'apprentissage. On voit donc que dans ce cas, on peut anticiper l'effet d'un bruit (produisant une rotation du nuage).

Cette prédictibilité n'existe pas avec les réseaux de neurones profonds, mais, ils restent cependant potentiellement vulnérables à ce type d'attaque.

On peut noter que le problème de l'empoisonnement est beaucoup plus compliqué que celui des exemples adversaires en théorie : le calcul d'un exemple adversaire est simplement une optimisation à travers un réseau i.e. si f est le réseau x un point la génération d'un exemple adversaire (contraint d'avoir une



attaque adversaire empoisonnement de données

FIGURE 2.10 – Illustrations de la différence entre attaque adversaire et empoisonnement : dans les 2 cas, l'objectif de l'attaquant est de faire en sorte que le modèle (représenté par la ligne noire) se trompe sur le point rouge soit en le modifiant (attaque adversaire) soit en modifiant le modèle (via les données d'apprentissage).

norme $||.||$ inférieure à ε) est *simplement*

$$\min_{\delta, ||\delta|| \leq \varepsilon} f(x) f(x+\delta)$$

Alors que l'empoisonnement de données (si on cherche à maximiser l'erreur) consiste à optimiser non pas à travers le réseau mais le processus d'apprentissage lui même qu'on pourrait décrire comme :

$$\begin{array}{l} \max\limits_{\delta_1, ... \delta_N \ / \ ||\delta_n|| \leq \varepsilon} |\{m, f(\chi_m) \neq y(\chi_m)\}| \\ f \sim SGD(x_1 + \delta_1, y_1, ..., x_N + \delta_N, y_N) \end{array}$$

où SGD représenterait la descente de gradient stochastique sur la base d'apprentissage $x_1, ..., x_N$, alors que $|\{m, f(\chi_m) \neq y(\chi_m)\}|$ représente l'erreur de test sur la base de test $\chi_1, ...\chi_M$. Cependant, en pratique les approches d'empoisonnement sont très grossières et ne cherchent pas directement à traiter cette formulation.

2.3.2 L'empoisonnement classique vs invisible

Précisément, dans le contexte de l'empoisonnement de données, le hackeur peut modifier à la fois les données et/ou directement des classes. Cependant, les modifications du hackeur doivent passer inaperçues pour éviter d'être détectées par les propriétaires de la base de données. Ceci invite soit à faire des modifications fortes mais en toute petite quantité soit à faire des modifications invisibles. Dans le premier cas, le hackeur espère que le propriétaire ne regardera pas chaque élément de la base (typiquement s'il y a 1000000 d'images) et qu'ainsi les changements de labels - *label flip* en anglais - ou les images fortement corrompues (comme dans la figure 2.11) passeront inaperçues. Ce type d'attaque [101] est le plus classique.

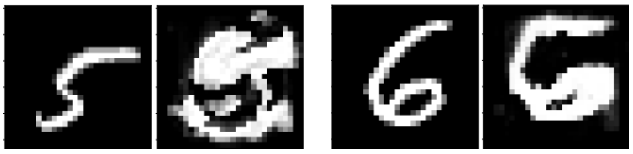


FIGURE 2.11 – Exemples de données empoisonnées dans [101] (contre un classificateur classique) : l'attaque est clairement visible mais seul 2% des échantillons sont modifiés

Dans le deuxième cas, l'idée est d'essayer de modifier un maximum de données mais de façon invisible pour une personne qui visionne ces données. On se rapproche alors d'une attaque adversaire invisible mais réalisée à l'apprentissage. On parle alors d'attaques *clean label* en anglais.

Globalement, les attaques dans [211, 85] sont relativement similaires et considèrent un encodeur pré-appris g sur lequel on vient apprendre un classificateur à vaste marge w (donc $f(x) = w^T g(x)$ où g correspond à un encodeur pré-appris). Dans [85], l'idée est d'attaquer une donnée cible x_{cible} (donnée) en trouvant une donnée x_{faille} avec un label différent telle qu'on puisse modifier x_{faille} pour que l'encodage de x_{faille} viennent correspondre à celui de x_{cible}. Ainsi une bonne classification de x_{faille} à l'apprentissage conduira à une mauvaise classification de x_{cible} en test.

Formellement, l'attaque consiste donc à calculer, pour un certain nombre de points x (avec $y(x) \neq y(x_{cible})$), le résultat de la minimisation $\min_{\delta, ||\delta|| \leq \varepsilon} ||g(x+\delta) - g(x_{cible})||$ puis à garder le x donnant le meilleur résultat. Si $||g(x+\delta) - g(x_{cible})||$ est très faible, alors il est probable que le w (classificateur à vaste marge appris au dessus de l'encodeur g) vérifie $w^T g(x+\delta) \approx w^T g(x_{cible})$ (alors que $y(x) \neq y(x_{cible})$).

Inversement, dans [211], l'objectif est de cibler un maximum des points du test (et pas un seul comme dans [85]). L'idée est donc de chercher un proxy ω tel que l'objectif du hackeur est de faire en sorte que le classificateur w appris par dessus g s'approche de ω. Il est alors possible de chercher une perturbation globale provoquant une rotation du modèle de w vers ω.

Dans les deux cas, l'empoisonnement est moins trivial dès qu'on considère l'apprentissage complet du réseau (et pas seulement celui d'un classificateur à partir d'un espace latent pré-appris). Cependant, cela est possible (déjà présent dans [85] avec une efficacité moindre et étendu de [211] à [224] dans mes travaux). À noter, dans [224], l'attaque est très efficace contre un VGG naïf qui passe de 80 à 30% de classification correcte sur CIFAR10 [177] (le jeu de données *bac à sable* de la vision par ordinateur). Cependant, cette attaque suppose que le hackeur peut à la fois modifier une grande partie des images de la base d'apprentissage et qu'il dispose d'informations sur la façon dont l'apprentissage est réalisé (ici, il a notamment besoin d'une paramétrisation du modèle cible donnant le comportement choisi).

D'ailleurs, la question de la dangerosité réelle de l'empoisonnement se pose : il

est peu probable qu'une véritable attaque d'empoisonnement de grande ampleur puisse être réalisée sur des modèles appris sur des bases de données consolidées. C'est donc plutôt un problème pour des algorithmes qui apprendraient de façon continue avec une supervision lâche de l'humain. De plus, j'ai récemment réalisé des expériences (non encore publiées [10]) qui montrent que les attaques [85, 224] ne fonctionnent pas quand on attaque une procédure d'apprentissage robuste (comme [124, 83]). Ainsi, ces méthodes d'empoisonnement basées sur des perturbations invisibles ont le même impactes que les attaques adversaires invisibles sur des réseaux naïfs mais aussi les mêmes limites sur des réseaux robustes. Par ailleurs, des approches de partitionnement des données d'apprentissage (*bootstrapping* en anglais) peuvent permettre de détecter (voire d'être insensible) à des empoisonnements concernant un faible nombre de données [38]. L'empoisonnement n'est donc probablement pas un risque majeur.

2.3.3 L'empoisonnement comme moyen de falsification

Bien que ce thème semble gagner en visibilité (exemple [23]), l'empoisonnement adversaire de données parait peu réaliste. Cependant, le point de départ de mes travaux porte sur l'empoisonnement **volontaire**. En effet, dans le cadre de la validation réglementaire d'un algorithme, il sera demandé de produire des preuves de performance. Et, il est probable que des tentatives de falsification puissent un jour avoir lieu [11].

Or, modifier les données de test ou d'apprentissage est un moyen très simple pour camoufler une falsification. Bien entendu, en théorie, les données de base de test doivent normalement être tirées de façon i.i.d. après la conception du modèle rendant une falsification impossible. Cependant, ce point n'est pas forcément compatible avec la pratique car les données sont souvent fournies par l'industriel lui même. Or, même non volontairement, le fait d'avoir des données de test visibles conduit à une surestimation des performances comme le rappelle [64] : en produisant une autre base de test supposée consistante avec Imagenet, les performances sont très différentes [12].

Ainsi, si les données de test sont modifiables, des falsifications triviales sont possibles comme le retrait des exemples gênants. De même, modifier les données de test de façon invisible pour les rendre plus faciles est aussi une falsification [13] invisible à l'oeil : ce serait donc des exemples complaisants dans ce contexte.

10. github.com/achanhon/AdversarialModel/blob/master/robustness_poisoning_final.pdf

11. Ceci ne constitue pas une accusation de l'industrie en général. Par contre, on peut se rappeler des exemples comme de Theranos ou du 737-Max qui montrent que des mauvaises pratiques peuvent subvenir de façon individuelle.

12. Cela peut au choix s'interpréter comme le fait que les données ne sont pas i.i.d. ou qu'il y a une surestimation due au fait que le dataset de test est plus ou moins utilisé comme de la validation.

13. À noter que des modifications naïves (par exemple ne respectant pas le mosaiquage spécifique d'une caméra [173]) pourraient être immédiatement détectées avec des outils basiques. Cependant, des modifications adaptées au senseur (qui irait donc plus loin que [5] qui considère des attaques entières) seraient difficilement détectables à la fois à l'oeil et à l'aide d'outils bas niveau.

Moyens difficilement détectables de falsification de résultats	
base de test modifiable	suppression de données et/ou modification *adversaire* des données
base de test visible	réglage des paramètres sur le test *empoisonnement* de la base d'apprentissage

TABLE 2.1 – Des falsifications difficilement détectables sont possibles dès que la base de données de test est modifiable. De façon moins évidente, cela est aussi possible dès qu'elle est simplement visible par exemple via des méthodes d'empoisonnement.

FIGURE 2.12 – Illustration de la stéganographie : l'image de droite est cachée dans celle de gauche de façon invisible à l'oeil. De la même façon, des motifs invisibles peuvent être ajoutés aux données d'apprentissage pour guider l'entraînement vers un modèle plus favorable en test ce qui constitue un moyen difficilement détectable de falsification.

Enfin, même si la base de test n'est pas modifiable, dès lors qu'elle est connue, il est possible de l'utiliser pour augmenter artificiellement ses performances comme le résume la table 2.1. On peut par exemple, régler des paramètres en regardant les performances en test (ce qui est une mauvaise pratique très courante consistant à confondre données de validation et données de test). Il est aussi possible de modifier la base d'apprentissage pour aiguiller l'entraînement vers un modèle meilleur en test. En effet, il est difficile d'exiger que la base d'apprentissage ne soit pas modifiable. Cependant, les modifications qu'on peut y faire peuvent être plus ou moins honnêtes. Ainsi, ajouter des exemples réels dans la base d'apprentissage est à la limite de l'apprentissage actif. Par contre, ajouter des hautes fréquences dans les images d'apprentissage (via une sorte d'empoisonnement) pour guider l'entraînement du modèle vers un modèle plus performant en test est une falsification difficilement détectable. La stéganographie est une illustration imparfaite mais cependant acceptable de cette idée d'empoisonnement volontaire comme présentée en figure 2.12.

Ce type de falsification est possible comme illustré dans mes travaux [224] (je n'ai trouvé aucun autre travail sur ce problème spécifique de la falsification).

Je continue actuellement de travailler sur cette question d'empoisonnement volontaire et sur les moyens de s'en prémunir (imposer un apprentissage robuste pourrait être une défense suffisante). Mais ces travaux ne sont pas encore assez matures pour les envisager pour un sujet de thèse.

2.4 Calibration

Les algorithmes d'apprentissage produisent des modèles f et l'objectif est que l'erreur $\int_{x,y(x)\neq \text{sign}(f(x))} P(x)\mathrm{d}x$ soit faible. Cependant, cette formulation laisse apparaître que $f(x)$ n'est utilisée que pour son signe $\text{sign}(f(x))$. Pourtant, si f est un modèle lisse, alors $|f(x)| \gg 1$ implique que le signe de $f(x)$ est dans une zone stable et $|f(x)| \approx 0$ implique que ce signe est dans une zone instable. Ce constat invite à s'intéresser à la valeur elle-même.

Ainsi, il existe un vaste sous-domaine de l'apprentissage associé à la notion de *calibration* qui porte notamment sur le fait de contraindre $f(x)$ à être interprétable comme une probabilité et/ou comme une incertitude et/ou s'intéresse à l'ordonnancement des valeurs de f indépendamment du signe.

Bien entendu, il est impossible de prédire si un modèle f est bien calibré [14]. Cependant, sur les *problèmes d'intérêts*, il n'est pas rare que la valeur f se généralise et/ou se transfère mieux que la décision (le signe). De même, en pratique quand $|f(x)|$ est grand (le modèle f prédit que x est fortement négatif ou fortement positif), le signe est souvent le bon. Ceci ouvre la voie à tout un pan de techniques pertinentes dans un contexte *d'IA de confiance* : l'utilisation de l'incertitude du modèle [199] et/ou la prédiction des erreurs [51] et/ou la construction de modèles calibrés, soit nativement [167], soit via une renormalisation des scores [62, 33].

J'ai co-encadré la thèse de Guillaume Vaudaux Ruth dirigée par Catherine Achard qui s'est intéressée à l'auto évaluation d'un détecteur et dont les travaux sont exposés ci dessous. Cependant, d'autres thèses à venir exploreront ce thème comme présenté en perspective.

2.4.1 Contexte général sur l'ordonnancement

2.4.1.1 Définitions

En classification d'images, les réseaux de neurones profonds sont classiquement appris pour prendre en entrée 1 image et pour produire 1 *nombre* par classe pour cette image. La classe prédite est alors celle qui a le *nombre* le plus haut - ce nombre peut donc s'interpréter comme un score - généralement appelé *likelihood* en anglais. Notons que cela se transpose en segmentation sémantique avec un score et/ou une probabilité par pixel. En détection, la situation est plus compliquée puisqu'il y a à la fois le fait d'accepter ou non une boite et le choix

14. Si on pouvait systématiquement trouver f tel que $|f(x)|$ corrèle avec le fait que $\text{sign}(f(x)) = y(x)$, on pourrait contredire le *No free lunch theorem* en énumérant tous les modèles pour en sélectionner un qui serait calibré pour un problème fixé (voir annexe 5.4.3).

d'une classe pour la boite mais cela reste comparable dans l'esprit. Quand il n'y a que 2 classes, on a alors un seul score (typiquement celui de la classe 1) qu'on compare à un seuil (naïvement 0).

On peut alors considérer l'ordonnancement des valeurs de f pour l'ensemble des x indépendamment du seuil : ainsi, on dira que l'ordonnancement est optimal si $\forall x, x', y(x) = 1 \wedge y(x') = -1 \Rightarrow f(x) > f(x')$. Cela peut avoir un intérêt si on considère non pas $\text{sign}(f(x))$ mais $\text{sign}(f(x) - \sigma)$ où σ est un seuil éventuellement réglable a posteriori (par exemple par un utilisateur).

Indépendamment, ces scores sont souvent normalisés pour être positifs et avoir une somme unitaire, on parle alors parfois abusivement de probabilité d'appartenance à la classe. Évidemment, si une donnée n'est pas ambiguë elle appartient à une et une seule classe. Mais cette probabilité doit se comprendre comme la probabilité qu'une donnée soit dans la classe sachant le nombre associé à cette classe en sortie du réseau. Précisément, un modèle est dit calibré si la probabilité qu'une donnée x tirée selon P soit effectivement dans la classe 1 est $f(x)$. C'est-à-dire, un modèle est dit calibré sur le problème y, P, si $\forall p \in [0, 1]$, $\frac{\int_{x, f(x)=y(x), f(x)\geq p} P(x)\mathrm{d}x}{\int_{x, f(x)\geq p} P(x)\mathrm{d}x} = p$.

2.4.1.2 Ordonnancement et détection

Cette idée de considérer l'ordonnancement des scores indépendamment du seuil est extrêmement majoritaire en détection d'objets dans l'état de l'art.

Cela me parait (toujours) étonnant puisqu'en terme de performance, la seule chose qui compte c'est si oui ou non le modèle fait des erreurs. Or, le fait d'être bien ordonnancé n'a pas d'intérêt sur l'erreur sauf avec un réglage a posteriori du seuil. Mais un tel réglage n'est pas possible puisque la vérité terrain de test n'est normalement pas connue. Donc, le seul contexte dans lequel un tel réglage a posteriori est possible est celui où l'algorithme est utilisé par un opérateur. Celui-ci peut alors le régler localement. Typiquement, si l'algorithme est utilisé sur une distribution légèrement différente ou sur un paquet de données qui représente une sous-distribution, il peut être pertinent de demander à l'opérateur d'avoir la possibilité de régler le seuil. C'est même positif pour l'opérateur de disposer d'un moyen d'action sur l'algorithme car cela augmente le sentiment de contrôle sur système. Mais cela n'a aucun intérêt pour un le système autonome.

Une raison expliquant (potentiellement) le choix d'une évaluation de l'ordonnancement en détection est de ne pas à avoir à pondérer les 2 types d'erreurs de détection. En effet, en détection, on ne peut pas considérer l'erreur de classification (définie en section 2.1.2) car le nombre de points avec $y(x) = -1$ est très largement dominant vis-à-vis du nombre de points avec $y(x) = 1$. On introduit alors deux types d'erreurs : les détections manquées (il y a un objet et on ne l'a pas vu $y(x) = 1$ mais $f(x) \leq \sigma$ où σ est le seuil de détection) et les fausses alarmes (il y a un objet prédit à un endroit où il n'y en a pas $y(x) = -1$ et $f(x) > \sigma$). Généralement, on introduit le rappel comme la pro-

portion du nombre d'objets détectés $\frac{\int_{x,y(x)=1,f(x)>\sigma} P(x)\mathrm{d}x}{\int_{x,y(x)=1} P(x)\mathrm{d}x}$ et la précision comme la proportion de fausses alarmes $\frac{\int_{x,y(x)=1,f(x)>\sigma} P(x)\mathrm{d}x}{\int_{x,f(x)>\sigma} P(x)\mathrm{d}x}$.

Mais on a alors une métrique faite de deux *nombres* là où on en voudrait qu'un seul pour pouvoir réaliser un classement. Mais, on ne peut pas juste prendre un seul de ces deux nombres, car il est facile de privilégier une erreur par rapport à l'autre : tout prédire ($f(x) = \sigma + 1$) permet d'être sûr de n'avoir aucune détection manquée (mais avec un trop grand nombre de fausses alarmes) et ne rien prédire ($f(x) = \sigma - 1$) permet d'être sûr de n'avoir aucune fausse alarme (mais aussi aucune détection). De plus, il est évident que les deux types d'erreurs ne sont pas nécessairement aussi coûteux en fonction des applications. Un module d'une voiture autonome qui doit déclencher un freinage d'urgence crée un danger moindre à freiner pour rien (fausse alarme) qu'à commettre un accident (détection manquée). Inversement, dans un moteur de recherche web, l'utilisateur ne saura pas qu'on ne lui a pas fourni une information pertinente (donc une détection manquée ne coûte rien individuellement) tant que toute l'information fournie l'est (donc seuls les fausses alarmes coûtent). Ainsi, une façon de ne pas à avoir à pondérer ces 2 erreurs est de considérer l'ordonnancement.

2.4.1.3 Courbe précision rappel

Précisément, la métrique majoritaire aujourd'hui en détection est l'aire sous la courbe précision rappel appelée *average precision* en anglais (abbreviée AP). L'idée est de considérer la courbe 2D tracée par les points *rappel en abscisse, précision en ordonnée* quand on fait décroître σ le seul de détection. Et, l'AP est l'aire sous cette courbe.

Cela dit, cette raison est discutable car les académiques aurait très bien pu considérer une *moyenne* que les industriels auraient ensuite pondéré pour tenir compte des spécificités de leur produit. Par exemple, on peut résumer la qualité d'un détecteur en 1 nombre via la moyenne harmonique (le F-score) ou l'intersection sur l'union IoU[15] qui offre un autre mélange de la précision et du rappel (mais qui au final est proche du F score).

Dit autrement, la métrique dominante en détection mesure l'ensemble des points de fonctionnement possibles pour un opérateur. Cela est pertinent mais ne correspond pas du tout au cadre classique de la littérature d'un système

15. **IoU vs tIoU :** Précision importante, pour l'intersection sur l'union, on parle ici bien de la métrique pour produire un score mesurant la qualité de l'ensemble des détections. Indépendamment, une détection (donc une emprise spatiale ou temporelle) n'est jamais exactement identique à l'emprise saisie par l'humain. Il y a donc toujours un critère de tolérance pour considérer qu'une détection correspond à l'emprise saisie par l'humain. Cette correspondance repose généralement aussi sur un critère d'intersection des emprises sur l'union des emprises. Mais, cette intersection sur l'union est locale (notée *tIoU* dans la suite) à une détection et ne doit pas être confondu avec l'intersection sur l'union comme métrique sur l'ensemble des détections.

autonome pour lequel il serait plus pertinent d'évaluer un et un seul point de fonctionnement avec par exemple un F-score.

Cela étant, il se trouve que dans un contexte d'emploi de détection d'objets d'intérêt dans des images satellites *défense*, il y a justement généralement un opérateur derrière le système. Ce qui explique (en partie) que l'ordonnancement fasse partie de mes thèmes d'intérêts.

2.4.2 SALAD

2.4.2.1 Motivation

Dans la littérature, l'ordonnancement est donc mesuré par la courbe précision rappel. Ce qui amène à des travaux qui maximisent directement cette aire [18, 28, 65, 121].

Cependant, ces travaux ne considèrent que le score de la sortie : *a-t-on confiance dans le fait que cette emprise/localisation (spatiale ou temporelle) est la bonne ?*. Or, on peut se poser une autre question : *a-t-on confiance dans le fait que cette emprise/localisation donne suffisamment d'information pour prendre la bonne décision ?*. Or, cette façon d'aborder le problème semble peu présente dans la littérature.

Cela a conduit à proposer SALAD une architecture qui optimise justement un score de confiance non pas sur la sortie mais sur l'entrée.

2.4.2.2 Contexte

Sans SALAD, l'architecture vise à être capable de détecter des actions à partir d'un seul *pivot* qui peut être un simple instant ou un groupe d'instant. Pour illustrer cette idée, on peut être sûr qu'un sportif fait du lancer de javelot avec une seule image représentative du lancer. Mais il est alors indispensable de bien choisir l'image utilisée. Ce qui motive de s'intéresser à la confiance envers la qualité de l'entrée.

Il n'y a pas vraiment d'équivalent image de ce type de structure. D'un côté, le détecteur YOLO [126] transforme l'image en une grille 13x13 dans laquelle 1 détection correspond à 1 pixel (précisément, une ancre parmi les k d'un pixel). Cependant, ces pixels encodent une grande partie de l'image. Inversement, une vidéo peut rapidement avoir plusieurs milliers d'instant et un pivot n'encodera qu'une petite partie de la vidéo (bien qu'un réseau récurrent soit utilisé pour fournir une information contextuelle autour de l'instant pivot considéré).

Plus généralement, l'état de l'art n'est pas tout à fait aussi avancé en détection d'actions dans des vidéos qu'en détection d'objets dans des images probablement probablement à cause de la très grande volumétrie qui caractérise les vidéos. Par exemple, BMN [61] qui était l'état de l'art en 2019 se contente d'estimer la probabilité pour chaque instant d'être le début et/ou la fin d'un segment d'action, puis, d'estimer un score d'appareillement début-fin. Cela explique pourquoi SALAD a pu être pertinent en vidéo alors qu'une architecture similaire aurait peut-être été insuffisante en image.

Cela dit, SALAD est donc fortement construit sur la possibilité d'estimer la *qualité du pivot* (bien plus que la *qualité du segment d'action prédit*). D'ailleurs ces segments sont juste directement régressés à partir de chaque pivot. Ainsi, cette structure voit la détection comme un problème de régression et non de classification : il faut régresser (i.e. déterminer) l'intervalle à partir d'un pivot. Mais, il n'est pas directement possible d'introduire un score d'acceptation dans un algorithme de régression. Un tel score peut être construit *a posteriori* de différentes manières. Une première façon de faire est de mesurer la variance de la réponse d'un ensemble de réseaux [117, 192]. Mais, l'approche choisie a été d'avoir une sortie supplémentaire qui produit ce score [75].

L'idée sous-jacente à [75] est donc de produire à la fois la régression et un score, étant donnée x associé à une valeur $z(x)$ (on est en régression ici donc z n'est pas dans un espace de label mais dans un espace continu), la régression produit $\hat{z}(x)$ et la sortie supplémentaire (d'auto-évaluation) produit $\hat{p}(x)$. On cherche alors à la fois à minimiser l'écart $||z(x) - \hat{z}(x)||_2^2$ mais aussi à avoir une sortie $\hat{p}(x)$ qui mesure la qualité de x pour effectuer la tâche de régression, par exemple, on peut se donner une valeur κ et vouloir que $||z(x) - \hat{z}(x)||_2^2 \leq \kappa \Rightarrow \hat{p}(x) = 1$ et $||z(x) - \hat{z}(x)||_2^2 > \kappa \Rightarrow \hat{p}(x) = 0$ (ce qu'on peut approximer via une fonction de perte type *cross entropy*, comme si $\hat{p}$ cherchait à prédire si $||z(x) - \hat{z}(x)||_2^2 \leq \kappa$ ou pas).

2.4.2.3 Score pivot vs score prédiction

L'algorithme proposé dans [221] (WACV 2021) est construit sur cette idée mais étendu car il s'agit de produire un score de pertinence du pivot plutôt qu'un score de pertinence de la prédiction. En pratique, cela signifie qu'on ne cherche pas un critère uniquement sur la boite produite (comme $||z(x) - \hat{z}(x)||_2^2 \leq \kappa \Rightarrow \hat{p}(x) = 1$ et $||z(x) - \hat{z}(x)||_2^2 > \kappa \Rightarrow \hat{p}(x) = 0$) mais un critère global qui compare les pivots entre eux.

Cette idée de donner un score à chaque pivot permet d'encoder un mécanisme de suppression des non maximaux directement dans ce score de confiance. Ce qu'on sait être un élément important de la performance de détection [88]. Cependant, le score de SALAD ne code pas uniquement la suppression des non maximaux car un maximum local n'est pas renforcé s'il ne recouvre pas un vrai positif.

SALAD introduit ainsi un algorithme dynamique où le modèle produit en chaque pivot potentiel un intervalle, puis compare cet intervalle avec les cibles potentielles, et enfin, compare chaque pivot avec ses voisins (voir table 2.4).

Grâce à ce processus, SALAD réussit à atteindre des performances supérieures à celles de l'état de l'art (début 2021) sur Thumoz [154] (table 2.2) et sur ActivityNet [133] (table 2.3 - avec un critère plus laxiste que dans l'état de l'art sur le recouvrement détection / vérité terrain).

Recouvrement	0.1	0.2	0.3	0.4	0.5
Oneata *et al.* [151]	36.6	33.6	27.0	20.8	14.4
Wang *et al.* [154]	18.2	17.0	14.0	11.7	8.3
Caba *et al.* [120]	-	-	-	-	13.5
Richard *et al.* [131]	39.7	35.7	30.0	23.2	15.2
Shou *et al.* [128]	47.7	43.5	36.3	28.7	19.0
Yeung *et al.* [144]	48.9	44.0	36.0	26.4	17.1
Yuan *et al.* [127]	51.4	42.6	33.6	26.1	18.8
DAPs [115]	-	-	-	-	13.9
SST [91]	-	-	37.8	-	23.0
CDC [103]	-	-	40.1	29.4	23.3
Yuan *et al.* [108]	51.0	45.2	36.5	27.8	17.8
SS-TAD [90]	-	-	45.7	-	29.2
CBR [95]	60.1	56.7	50.1	41.3	31.0
Hou *et al.* [97]	51.3	-	43.7	-	22.0
TCN [93]	-	-	-	33.3	25.6
TURN-TAP [94]	54.0	50.9	44.1	34.9	25.6
R-C3D [107]	54.5	51.5	44.8	35.6	28.9
SSN [109]	66.0	59.4	51.9	41.0	29.8
BSN [81]	-	-	53.5	45.0	36.9
BMN [61]	-	-	56.0	47.4	38.8
Chao *et al.* [74]	59.8	57.1	53.2	48.5	42.8
G-TAD [42]	-	-	54.5	47.6	40.2
SALAD	**73.3**	**70.7**	**65.7**	**57.0**	**44.6**
BSN + PGCN [70]	69.5	67.8	63.6	57.8	49.1
G-TAD + PGCN	-	-	66.4	60.4	**51.6**
SALAD + PGCN	**75.2**	**73.4**	**69.4**	**61.6**	49.8

TABLE 2.2 – Performances de la méthode SALAD sur THUMOS14 [154] mesurée par la AP moyennée par classe (en fonction du critère de recouvrement indiqué en colonne, un nombre plus haut correspond à un recouvrement plus strict).

Recouvrement	0.1	0.2	0.3	0.4	0.5
Singh *et al.* [129]					34.47
SCC [96]					40.00
CDC [103]					45.30
R-C3D [107]					26.80
SSN [109]					39.12
BSN [81]					46.45
Chao *et al.* [74]					38.23
P-GCN [70]					48.26
G-TAD [42]					50.36
BMN [61]	70.91	64.46	58.79	54.14	50.07
SALAD	**77.68**	**70.66**	**64.06**	**57.45**	**51.72**

TABLE 2.3 – Performances de la méthode SALAD sur ActivityNet [133] mesurée par la AP moyennée par classe (en fonction du critère de recouvrement indiqué en colonne, un nombre plus haut correspond à un recouvrement plus strict).

Algorithme :
Entrée : Soit $\{[\hat{d}_t, \hat{f}_t], \hat{p}_t\}_{t=1}^{T}$ l'ensemble des segments *(début, fin) et score* pour l'ensemble des pivots (i.e. instants). Et soit $\{[d_n, f_n]\}_{n=1}^{N}$ les vrais segments d'actions. μ la valeur de recouvrement minimal pour considérer qu'une détection recouvre une action.
Sortie : une fonction de perte *dérivable* combinant régression des segments à partir d'un petit ensemble de pivots sélectionnés.
Pseudo-code :

`Calculer` $\sigma(t)$ `un tri des pivots par valeur de` $\hat{p}_t$ i.e. $p_{\sigma(t+1)} \geq p_{\sigma(t)}$
$\alpha \leftarrow \mathbf{0}$, $\beta \leftarrow \mathbf{0}$, $y \leftarrow \mathbf{0}$, $\rho \leftarrow \mathbf{0}$
`pour` t `de` 1 `à` T
  `pour` n `de` 1 `à` N
    `si` $\beta_n = 0$
      `si` $d_n \leq \sigma(t) \leq f_n$
        $\alpha_{\sigma(t),n} \leftarrow 1$
        $\rho_{\sigma(t),n} \leftarrow \frac{\min(f_n, \hat{f}_{\sigma(t)}) - \max(d_n, \hat{d}_{\sigma(t)})}{\max(f_n, \hat{f}_{\sigma(t)}) - \min(d_n, \hat{d}_{\sigma(t)})}$
        `si` $\rho > \mu$
          $\beta_n \leftarrow 1$
          $y_{\sigma(t)} \leftarrow 1$

$$loss \leftarrow \sum_{t=1}^{T} (y_t \log(p_t) + (1 - y_t) \log(1 - p_t)) - \lambda_1 \sum_{t=1}^{T} \sum_{n=1}^{N} \alpha_{t,n} \rho_{t,n}$$

TABLE 2.4 – Processus dynamique introduit dans SALAD pour calculer la fonction de perte associée à un choix d'un score pour chaque pivot et à un choix d'un intervalle à partir de chaque pivot.

2.4.3 Perspectives

Bien entendu, dès fin 2021, SALAD n'était plus l'état de l'art de la détection d'actions. Mais le point fort de SALAD n'est pas tant la performance ayant été obtenue à date de publication, mais plutôt que le score d'auto-évaluation introduit est indispensable au processus.

En effet, la régression seule ne converge pas : la convergence semble freinée par le nombre de pivots puisqu'en l'occurrence chaque image est un pivot dans la version naïve. Inversement, ils sont élagués par la procédure dynamique présentée en table 2.4 (seul les pivots avec $\alpha_{n,t} = 1$ sont considérés pour entraîner la régression). De même, les deux composantes du score d'auto-évaluation sont très importantes car les performances baissent de façon importante quand on supprime l'une ou l'autre. Pour rappel, ces deux composantes sont le fait de tenir compte à la fois de la qualité du segment prédit (comme dans le cas d'un score classique d'ancre) et le fait d'estimer la pertinence de laisser la priorité à un pivot voisin (une sorte de suppression de non maximaux intégrée). Cela montre un intérêt à l'idée d'essayer de prédire la qualité de la décision prise à partir d'une zone, plutôt que de prédire la qualité de la décision elle-même.

Enfin, l'autre élément important est que les 100 premiers segments produits par SALAD ne sont pas (en moyenne) meilleurs que les 100 premiers produits par BMN. Par contre, leur ordonnancement est bien meilleur ce qui permet à la mAP d'être bien supérieure. Ainsi, même sans utiliser de méthodes du domaine

strict de la calibration, cet algorithme tend à être mieux calibré parce qu'il est structurellement construit sur l'optimisation du score d'auto-évaluation.

D'ailleurs, cette idée de s'intéresser à la calibration, pas uniquement comme une caractéristique souhaitable mais comme un élément structurant de l'algorithme est présente dans la littérature. Par exemple, dans [43], le fait de forcer la distribution des scores de chaque paquet à se rapprocher d'une distribution moyenne améliore de façon significative les performances de l'algorithme (en plus de produire une meilleure calibration).

Avoir un algorithme entièrement construit autour de la notion de probabilité d'occurrence est par ailleurs probablement indispensable pour essayer d'apprendre des phénomènes stochastiques. Par exemple, j'ai parlé dans le chapitre 1 d'estimation d'un risque foudre. Cependant, dans [208], il s'agit d'un risque à résolution spatio-temporelle basse relativement prévisible. Inversement, si on veut monter en résolution spatiale et temporelle, le risque foudre devient stochastique car un éclair peut éclater à tout moment et n'importe où au sein d'une cellule orageuse. D'ailleurs, la présence d'un éclair en un instant n'implique pas la présence d'un éclair à l'instant suivant (à la différence d'un nuage ou d'une dépression). Ainsi, pour étendre les travaux de [208], il est nécessaire de penser l'algorithme entièrement autour de la notion de calibration (car on peut estimer un risque foudre même sans être capable de prédire précisément où et quand l'éclair éclatera) de façon similaire à la construction de SALAD. Cette perspective devrait donner lieu à un sujet de thèse.

2.5 Synthèse

L'objectif de ce chapitre est de montrer que la cohérence de mes travaux est d'essayer de couvrir un large panel des questions relatives à l'IA de confiance résumé ci-dessous et dans la figure 2.13 :

- Sur la robustesse aux bruits adversaires à travers mes travaux sur l'empoisonnement.
- Sur la robustesse aux attaques par patch à travers la thèse de Pol Labarbarie.
- Sur l'équité des détecteurs à travers la thèse de Magdeleine Airiau.
- Sur la calibration à travers la thèse de Guillaume Vaudaux Ruth (en détection d'actions) et à travers une thèse à venir (sur la prédiction de risque orageux).

Par ailleurs, mes travaux futurs (chapitre 4) continueront à s'inscrire dans ce thème de l'IA de confiance qui conditionne l'acceptabilité de ces méthodes d'apprentissage profond. Cependant, cette acceptabilité (et ces travaux) pourrait être freinée indépendamment par le manque de données annotées. Ce qui amène d'autres travaux présentés dans le chapitre suivant.

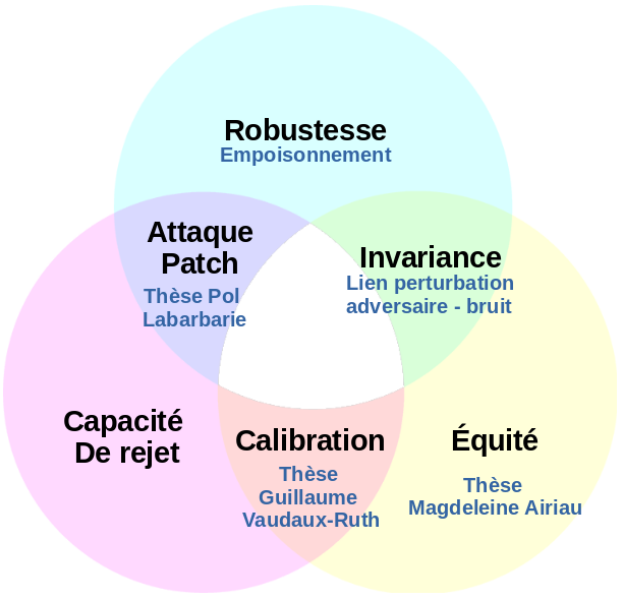


FIGURE 2.13 – Cette figure à mettre en regard de la figure 2.2 illustre mes travaux ou co-encadrements dans l'ensemble de thématique d'IA de confiance.

Chapitre 3

Le coût de l'annotation

L'IA de confiance est probablement la clé de l'acceptabilité de l'apprentissage profond dans des applications critiques (comme discuté en chapitre 2). Ainsi, une grande partie de la communauté travaille à *construire la confiance*. Cependant, mon travail m'amène aussi à appliquer l'apprentissage profond à d'autres sciences notamment pour traiter des données internes à l'ONERA. Or, l'obstacle principal est alors bien plus *pratique*. On peut le résumer comme suit : *personne ne veut annoter*. De nombreux articles du chapitre 1 (comme [239, 235]) n'auraient d'ailleurs pas été possibles sans un effort d'annotation des donnés brutes.

Malgré les apparences, ce problème de coût des données d'apprentissage est bien réel dès qu'on sort des grandes applications industrielles : des algorithmes même performants et de confiance ne trouveront que peu d'applications si des milliards d'exemples annotés sont nécessaires pour apprendre des modèles pertinents. Cela conduit à tout un écosystème allant de l'IA frugale à la mécanisation de l'annotation.

3.1 Contexte

3.1.1 Le coût des exemples

L'idéal de l'apprentissage par ordinateur est de réussir à comprendre un concept uniquement à partir de quelques exemples, ce que nous (êtres humains) arrivons à faire naturellement. Nous sommes ainsi tout à fait capables de nous faire une représentation mentale d'une licorne sans en avoir pourtant aucun exemple. Cela amène à une littérature sur l'apprentissage sans exemple [111] (*zero shot learning* en anglais) ou plus généralement à des approches qui visent à connecter des représentations symboliques (comme le langage) et des exemples [73]. Cependant, les performances atteignables aujourd'hui avec ce type de méthodes (notamment l'apprentissage sans exemple) sont très inférieures aux performances de l'apprentissage supervisé notamment en segmentation sémantique [48]. Ainsi, bien que ce type de méthodes soit l'avenir de l'apprentissage par

ordinateur, les applications à venir à court terme seront basées sur des méthodes supervisées (comme celles présentées en chapitre 1). Or, ces algorithmes cherchent plutôt à capturer des corrélations entre des motifs qu'à comprendre un concept sous-jacent. Mais, pour capturer ces corrélations entre des motifs, il faut beaucoup d'exemples annotés ce qui conduit à la question du coût d'obtention des exemples.

On peut noter que la donnée elle-même est souvent coûteuse. Bien entendu, ce problème se pose à des degrés divers en fonction des contextes. Le recueil de données ne sera probablement pas un problème pour la voiture autonome (où des budgets conséquents y sont dédiées) ou le numérique (où d'énorme quantité de données sont préexistantes et déjà partiellement annotées). Mais, il est dimensionnant dans le médical (pour donner un ordre de grandeur, dans le projet avec la startup VitaDX l'inclusion de chaque patient nécessite 1000€ en moyenne alors que le prélèvement d'urine n'est pas invasif).

Maintenant, dans le domaine des sciences physiques, les données brutes sont généralement préexistantes mais non formatées pour être utilisées par des algorithmes d'apprentissage profond. Ce qui pose principalement la question du coût de leur annotation qui sera l'objet de ce chapitre.

3.1.2 Le coût de l'annotation spatialisée

Précisément, de très nombreuses applications potentielles de l'apprentissage profond portent sur des problèmes de segmentation sémantique et/ou de détection d'objets dans des images (comme présenté en figure 1.1). Ces problèmes nécessitent une annotation spatialisée : il ne s'agit pas simplement de dire *si l'image est une image de chat* mais *de saisir les emprises spatiales correspondants à des chats* soit via des boites englobantes en détection soit au niveau de chaque pixel en segmentation sémantique. Cela conduit ainsi à une annotation bien plus complexe et très coûteuse en temps que ce soit en détection (surtout quand les images sont denses en objets) ou pire en segmentation car l'annotation doit a priori être précise au pixel près. Notamment, rien que pour la détection d'objets, [142] mesure que le temps de saisie d'un rectangle englobant un objet dans une image est 10 fois supérieur au temps nécessaire pour simplement classer la zone correspondante - valeur que l'on a confirmée à l'ONERA dans le cadre d'expérimentations technico-opérationnelles de la défense. Ce chiffre est encore plus élevé en segmentation sémantique, où l'annotation doit coller à l'emprise spatiale. D'autant que, dans ce cas, cela nécessite des outils de saisie complexes, et donc, des interfaces peu intuitives, comme le montre diverses publications sur le sujet comme LabelMe [182], ViLM [170], ViPER [145]... On pourra se rapporter à la page *github.com/heartexlabs/awesome-data-labeling* qui montre à quelle point la communauté a besoin de ce type d'outils.

Ce constat a conduit à envisager des approches mécanisées qui suppose néanmoins un modèle sous-jacent efficace. Ce dernier point explique que ces approches soient relativement récentes [1], notamment ViPER [145] insiste sur

1. Bien que [165] l'envisage avec des arbres de Hough qui sont des arbres de décision

cette dimension interactive mais plutôt propre à la vidéo avec utilisation de méthode de suivi - *tracking* en anglais (j'ai personnellement développé un prototype similaire a ViPER pour annoter [2] la base de données VIRAT [166]). Malheureusement, apprendre un détecteur *image* à partir de données vidéo n'est pas aussi intéressant qu'on pourrait le penser car les données obtenues sont très corrélées [212]. Cependant, aucune approche de mécanisation de l'annotation spatialisée ne s'était vraiment imposée en 2018, alors que le besoin de masques de segmentation sémantiques restait très important. Ce qui a motivé un ensemble de travaux décris ci après.

Ces travaux portent en particulier sur l'annotation **par** quelques clics et l'annotation **en** quelques clics.

3.2 Annoter par quelques clics

Être capable d'annoter des images pour des tâches de segmentation et/ou de détection par quelques clics correspond à concevoir des algorithmes capables d'apprendre à l'aide d'une vérité terrain partielle constituée de quelques points colorés seulement. Plus globalement, quand la vérité terrain est imparfaite, on parle d'apprentissage faiblement supervisé - *weakly supervised* en anglais.

3.2.1 Bruit d'annotation

En effet, il est commun que la vérité terrain soit imparfaite. Cette imperfection peut se manifester de façon diverse. L'annotation contient déjà toujours un bruit blanc résiduel lié au erreur de saisie. Mais ce bruit est parfois structurel. Par exemple, on peut avoir des contextes où seule une partie des positifs sont annotés. Cela ouvre vers une littérature dite *multiple instance learning* en anglais qui ne sera pas plus exploré dans ce manuscrit mais dont une porte d'entrée peut être trouvé dans [196, 155].

En segmentation sémantique, l'annotation est toujours bruitée à la jonction entre 2 emprises spatiales de 2 classes différentes comme illustrée par la figure 3.1 construite sur le jeu de données AIRS [49]. À noter, l'impact de ces bords n'est pas du tout négligeable ni pour l'apprentissage (il focalise l'attention du modèle dans des zones ambiguës [148, 20]), ni en test (ces zones bruitées peuvent représenter une fraction importante en terme de performance). Sur l'exemple de AIRS [49] (rééchantillonnée à 50cm), la vérité terrain rognée d'1 pixel conduit à seulement 90% d'IoU [215] ! D'ailleurs certains jeux de données comme ISPRS [163] ne fournissent qu'une vérité terrain déjà rognée. Cependant, ce bruit spatial peut être beaucoup plus important qu'un simple bruit aux frontières. Dans le cas extrême, l'annotation peut concerner l'ensemble de l'image sans aucune indication spatiale [58, 60, 125]. Parfois (notamment en télédétection), l'annotation est produite en recoupant d'autres sources de données ce qui donne parfois une spatialisation approximative : par exemple, si l'annotation suit le cadastre

produisant des votes spatialisées.

2. github.com/achanhon/VIRAT-AERIAL-ANNOTATION

FIGURE 3.1 – Illustration de l'ambiguïté du placement des bords des objets en segmentation sémantique même dans le cas d'une vérité terrain de très bonne qualité faite à la main (ici sur la base de données AIRS [49] rééchantillonnée à 50cm). À gauche, les bords externes (la ligne rouge est supposée être au bord mais en dehors des bâtiments), à droite les bords internes (la ligne rouge est supposée être le bord des bâtiments). Il est très difficile de faire la différence à l'oeil sans zoomer comme dans l'image de la 3ème ligne où on voit que le placement du bord est discutable. Ce problème n'est pourtant pas cosmétique car on perd 10% d'IoU en prédisant la ligne rouge de gauche au lieu de celle de droite.

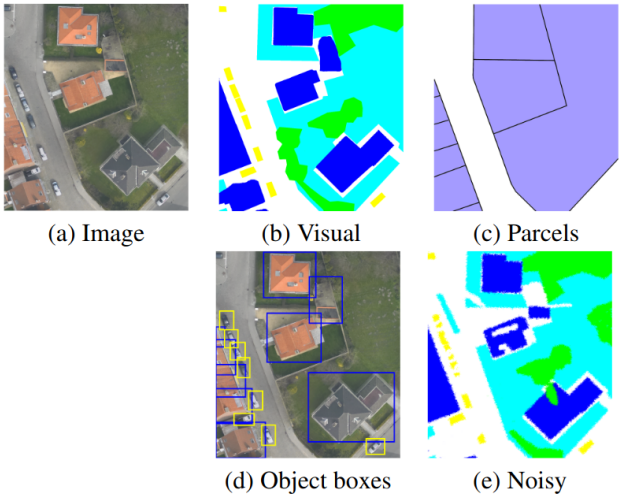


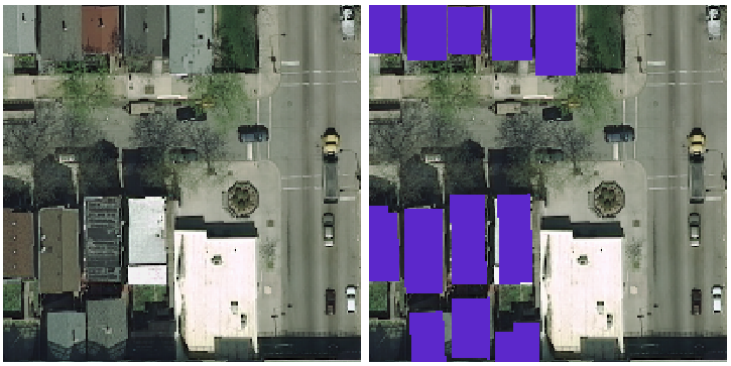
FIGURE 3.2 – Illustration de différents types de bruit d'annotation : un bruit *blanc* (en e) peut être simplement filtré mais pas un bruit systémique (comme en c). Dans notre cas, le bruit associé au clic se rapproche d'une emprise faible (comme en d). La figure utilise des données ISPRS [163].

[228] et/ou des bases ouvertes comme *open street map* [100]. En effet, dans le cadastre, une maison et son jardin seront englobés dans une unique parcelle alors que la classe sémantique de ces deux zones est différente comme illustré dans la figure 3.2. Dans le cas *d'open street map*, les emprises sont *supposées être correctes*, cependant, les données sont souvent non actualisées et avec une géométrie approximative comme illustrée en figure 3.3.

3.2.2 Débruitage

Les méthodes pour apprendre malgré le bruit d'annotation spatial vont dépendre du niveau de bruit. La simple correction des bords des objets [148, 20] n'est pas suffisante dans le cas de zones données via un cadastre. Une approche développée a l'ONERA [228] consiste à utiliser un réseau pour lisser l'annotation elle-même. Cette idée qu'un modèle peut difficilement apprendre le bruit est aussi au coeur du filtrage proposé par [72] qui rejoint les récentes approches de NERF [17] où les réseaux sont utilisés pour fusionner les données et non pour prédire une décision.

Dans le cadre d'une annotation faite de clic, l'information y est spatialisée contrairement au cas d'une information purement image [125] mais cette spatialisation est trop faible pour être simplement filtrée comme si c'était du bruit d'annotation. On se rapproche du cas d'une annotation de type cadastre sauf que le cadastre est englobant (aucun pixel n'est sûr mais on dispose d'emprise spatiales) alors que l'annotation faite de clic est interne (chaque point est sûr



Une image et l'emprise des bâtiments (d'après open streep map). Le gros bâtiment blanc n'est pas présent dans la base...

FIGURE 3.3 – Illustrations des erreurs de vérité terrain dans le jeu de données INRIA [100]

mais ne donne aucune information sur l'emprise). Actuellement, on peut noter que l'idée d'une supervision en un clic existe dès [110] (pour les mêmes raisons de coût et aussi parce que psychologiquement le clic est suffisant pour l'humain). Néanmoins, les performances obtenues y sont très inférieures à celle obtenue via une supervision normale.

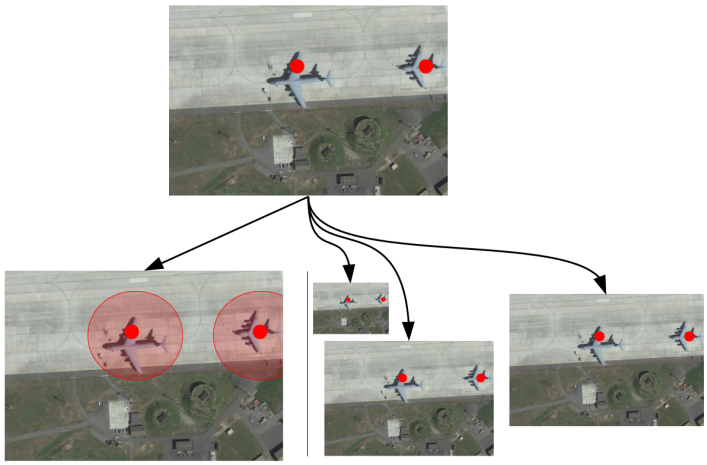
J'ai d'ailleurs essayé de reproduire ces résultats dans un contexte de télédétection [230] qui est plus favorable compte tenu de la constance d'échelle des objets. Précisément, deux approches illustrées en figure 3.4 sont comparées dans [230] :

— Une approche consistant à remplacer le clic par un disque puis à le filtrer avec une approche comparable à ce qui pourrait être fait avec des parcelles.
— Une approche (spécifique à un réseau de neurones) consistant à injecter le clic à différents étages du réseau de neurones.

Cependant, même en télédétection, et malgré une forte dynamique sur ce type de travaux dans la littérature [9, 14, 68, 39, 13, 45], les performances des techniques d'apprentissage faiblement supervisé pour réussir à tirer partie d'une annotation donnée par des points restent relativement basses par rapport à une approche supervisée.

3.2.3 Limites et alternatives

Pour relativiser cette observation que l'utilisation d'une annotation faite de clics conduit à des performances significativement plus faibles que l'utilisation d'une vérité terrain dense, il faut garder en tête que les performances mesurées en segmentation sémantique sont assez hautes dans l'absolu. Par exemple, sur AIRS [49], un UNet [141] construit sur un EfficientNet [67] atteint facilement plus de 90% d'IoU : c'est plus que la vérité terrain elle-même rognée (ou



La figure utilise une image googlemap (Ramstein, Allemagne).

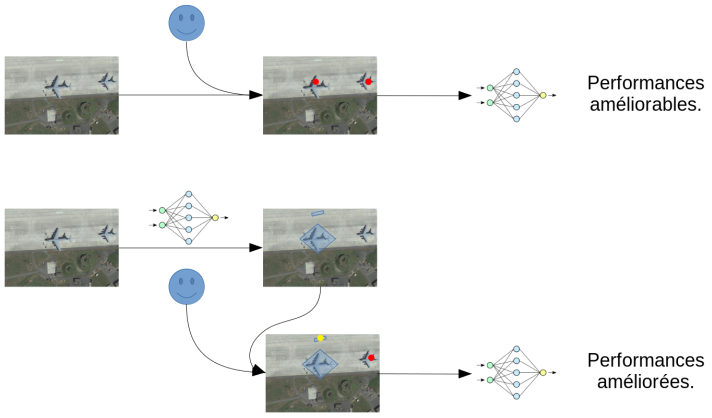
FIGURE 3.4 – Illustration de 2 approches pour traiter des annotations sous forme de clics [230] : raffiner une boule centrée sur chaque clic, ou injecter les clics à différents étages du réseau de neurones utilisés.

étendue) de 1 pixel (alors même que cette vérité terrain est justement bruitée sur ses bords). Bien entendu, en réalité les prédictions des réseaux contiennent de *vraies erreurs* (des bâtiments entiers sont oubliés ou hallucinés) mais cela est compensé par une prédiction qui arrive à copier la *convention d'annotation du jeu de données* mieux que ne ferait un humain. Ainsi, dans ce contexte ne pas disposer d'une annotation fine vient logiquement dégrader les performances qui sont compensées par la capacité des algorithmes à coller aux frontières des classes.

Dit autrement, l'utilisation d'une métrique rigide (par exemple, qui va pénaliser k pixels de bâtiment oubliés de la même façon qu'ils forment un seul bloc ou qu'ils soient répartis dans des zones ambiguës comme les bords) est un problème qui biaise l'évaluation des algorithmes. D'ailleurs, il est fréquent en télédétection que deux produits de même IoU soit en réalité de qualité très différents pour l'oeil humain.

Néanmoins, l'IoU est la métrique standard de la littérature et cela a un sens dans toutes les applications où les bords sont importants. Or pour cette métrique, il est clair qu'une supervision dense est indispensable pour atteindre des performances élevées. Ce qui conduit à considérer une mécanisation de l'annotation présentée en section suivante.

Cela dit, on peut quand même noter qu'indépendamment de la simple considération de temps d'annotation, l'utilisation d'un simple point pour représenter une emprise spatiale est quelque chose de commun pour les humains. En réalité, la quasi-totalité des informations géographiques disponibles sont représentées par des points : une adresse, une coordonnées GPS, ... Par exemple, la base offi-



La figure utilise une image googlemap (Ramstein, Allemagne).

FIGURE 3.5 – Illustration de l'intérêt d'une approche d'apprentissage interactif (en bas) vis-à-vis d'une approche d'apprentissage faiblement supervisée (en haut) : pour un même budget (ici de 2 clics), l'opérateur peut focaliser son effort dans les zones mal traitées par le modèle courant dans l'approche interactive.

cielle de la Poste *www.data.gouv.fr/fr/datasets/base-officielle-des-codes-postaux* donne un unique point GPS pour représenter une commune. Ainsi, être capable de tirer partie d'une annotation faite de clics est connecté avec une perspective importante pour la télédétection mais ce problème ne fait pas partie de mes principaux axes de recherche futurs.

3.3 Annoter en quelques clics

Le constat de la discussion précédente est qu'apprendre à partir de quelques clics figés ne conduit pas à des performances satisfaisantes. Mais, c'est pourtant parfois la seule chose qu'on peut attendre de l'utilisateur (par exemple, sur des thématiques scientifiques de niche [244]). Cependant, une solution serait que ces clics soient réalisés de façon dynamique car l'utilisateur peut être guidé par le résultat du modèle courant, et donc cliquer là où il pense pouvoir aider celui-ci. Dit autrement, on dispose d'un budget de clics. Utiliser une annotation figée revient ainsi à utiliser la totalité du budget sans considérer le modèle courant. Inversement optimiser le budget d'annotation de façon dynamique peut conduire à de meilleures performances en focalisant l'annotation là où elle semble le plus utile compte tenu du modèle courant comme l'illustre la figure 3.5.

Ceci invite à considérer l'apprentissage interactif [59, 80] dans lequel l'utilisateur interagit avec un algorithme - *interactive learning* en anglais. Il est alors possible d'envisager maintenir une annotation dense tout en ne demandant que des points : ce sujet est au coeur de la thèse de Gaston Lenczner dirigée par Guy Le Besneray, et coencadrée par Nicola Luminati, Bertrand Le Saux, et moi

même, réalisé en partenariat entre Alteia et l'ONERA.

3.3.1 DISIR

Au vu des raisons décrites ci-dessus, l'objectif est donc d'être capable d'interagir avec les clics d'un utilisateur pour produire une carte de segmentation sémantique précise.

3.3.1.1 Approches existantes

Cet objectif n'est pas nouveau puisque c'était déjà plus ou moins le cas avec GrabCut [189] en 2004. Cependant, GrabCut conduit à une segmentation au sens historique [202, 203, 204] et pas à une segmentation sémantique (notamment capable de gérer des zones texturées). De plus, l'objectif est de minimiser le nombre de clics. Or, GrabCut impose a minima 1 clic par objet puisque les objets n'ont pas de sémantique a priori.

Aussi, la minimisation des clics invite à considérer un modèle sémantisé (capable d'effectuer une prédiction seul) mais aussi capable d'exploiter des clics utilisateur. Cela introduit immédiatement une limitation : il est alors nécessaire de disposer d'une petite pré-base d'apprentissage qui permet d'aligner le modèle sur une sémantique. Cette limite sera traitée après.

Néanmoins, à supposer qu'on dispose d'une petite base d'apprentissage (qu'on souhaite étendre), [225] propose un tel *modèle sémantisé mais aussi capable de gérer des clics*. Ce paradigme appelé DISIR inspiré de DIOS [130] consiste à apprendre à un réseau de neurones :

— À prédire seul en l'absence d'annotation utilisateur.
— Mais aussi à raffiner en présence d'annotation.

Cette double capacité illustrée en figure 3.6 dépasse à la fois GrabCut en proposant une sémantique mais aussi une simple correction a posteriori d'une prédiction pré-existante puisque le modèle apprend à tirer parti du clic comme résumé en table 3.1.

3.3.1.2 Apprendre simultanément sémantique et gestion des clics

Au coeur de ce paradigme, l'idée centrale est de représenter les clics de l'utilisateur comme une entrée supplémentaire mais modifiable du réseau : chaque classe dispose d'une entrée et l'utilisateur est amené à effectuer des clics *colorés* au sens où un clic pour une classe est encodé dans la couche correspondante. Le paradigme de DISIR peut ainsi s'illustrer par la figure 3.7 qui apporte une implémentation pratique des objectifs de la figure 3.6.

Ce paradigme a été intensivement testé sur de nombreux datasets académiques de télédétection (POTSDAM, AIRS, INRIA) et permet un affinage efficace de la carte de prédiction produite. Ainsi, l'IoU est améliorée de 3% sur AIRS, 3.7% sur INRIA et 4.2% sur POTSDAM avec 120 clics localisés dans les grandes zones d'erreurs. Cela représente en moyenne 5000 pixels corrigés par clics. Ce résultat est remarquable surtout qu'il reste valide pour des niveaux

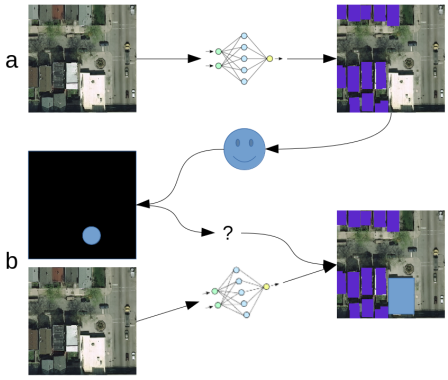


FIGURE 3.6 – Les deux objectifs du système d'annotation : être capable de prédire une sémantique (a) mais aussi de pouvoir l'affiner à l'aide d'entrées *utilisateur* (b).

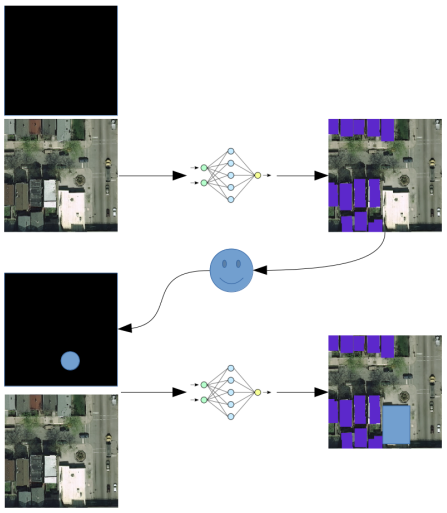


FIGURE 3.7 – DISIR : les objectifs du système d'annotation (illustrés en figure 3.6) sont réalisés en considérant un réseau profond qui prend en entrée des cartes initialement vierges capables d'encoder les clics de l'utilisateur.

gestion du clic	sans modèle	avec modèle
sans sémantique		GrabCut [189]
avec sémantique	filtrage d'une prédiction	DISIR

TABLE 3.1 – DISIR combine l'utilisation d'une sémantique et l'utilisation d'un modèle pour prendre en compte les clics utilisateurs.

de performances initiales différents (très haut sur AIRS, moins sur POTSDAM qui est multi-classes). Par ailleurs, ce résultat reste valide pour des objets de tailles variables : il est peu sensible à des changements de résolution (a minima sur AIRS où le gain reste stable d'une résolution de 7.5cm à une résolution de 50cm).

Il faut cependant noter que cette implémentation suppose de trouver le bon équilibre entre les deux tâches du réseau. C'est justement en trouvant un tel compromis que [225] atteint la performance de 5000 pixels annotés en un clic. Ce compromis est réalisé par un encodage efficace des clics ainsi que par une procédure d'apprentissage couplant les deux tâches. Durant l'apprentissage, des clics associés à des classes sont émulés aléatoirement dans les images à partir de la vérité terrain de façon éparse. Ainsi, le réseau voit deux types de pixels. Il voit des zones sans clic ce qui le force à apprendre à capturer la corrélation entre l'image et la vérité terrain de segmentation. Mais, il voit aussi des zones associés à un clic (qui donne au niveau du pixel correspondant la vérité terrain).

Je me permets d'insister sur le fait que cet équilibre n'est pas trivial : notamment, en présence de trop peu de données, le réseau n'a aucun intérêt à apprendre à utiliser les clics (qui sont intermittents) là où l'image est a priori toujours disponible. C'est d'ailleurs ce qu'on peut observer quand on réduit la taille de la base d'apprentissage. Une alternative évidente est alors de supprimer l'image de façon intermittente. Cependant, cela tend à complexifier l'optimisation.

Ce point est particulièrement problématique car il amplifie la limite déjà identifiée que cette approche nécessite une pré-base d'apprentissage (qui doit donc être conséquente).

3.3.2 DISCA

Une première réponse à ce problème serait de pouvoir utiliser une approche semblable à DISIR sur une base de données *proche mais différente* de la base utilisée pour régler le modèle. Si l'objectif est d'annoter une base de toits de bâtiments sur une ville donnée, il est envisageable d'utiliser un modèle DISIR appris sur une base de données de segmentation de toit académique comme INRIA ou AIRS.

Ce cas d'application est d'ailleurs pertinent d'un point de vue industriel car les modèles d'apprentissage profond nécessitent des exemples diversifiés afin d'être capables de gérer un changement de ville [215]. Mais plus largement, disposer d'une base de données relativement proche de la donnée qu'on souhaite annoter est une hypothèse bien plus faible que celle de devoir disposer d'une pré base.

Il se trouve que DISIR se transfère difficilement [3] ce qui a conduit à l'algo-

3. On peut noter un gain de performance de 20% lors du transfert d'un modèle AIRS vers POTSDAM en 120 clics. Précision importante, l'expérience DISIR sur POTSDAM est multi classes et commence avec une performance initiale forte. Inversement, dans le transfert AIRS vers POTSDAM, on est sur une segmentation de toits avec une performance moyenne de 58% d'IoU seulement. Mais, elle est facilement améliorable car au bout de 120 clics, la performance

	sans gestion du clic	avec gestion du clic
poids fixes		DISIR
poids variable	WTP [110]	DISCA

TABLE 3.2 – DISCA combine l'apprentissage faiblement supervisé sur les points cliqués (comme WTP qui n'atteint pas le niveau de performance voulu) et la propagation de ces clics (comme DISIR qui ne traite que des données relativement proches du domaine initial).

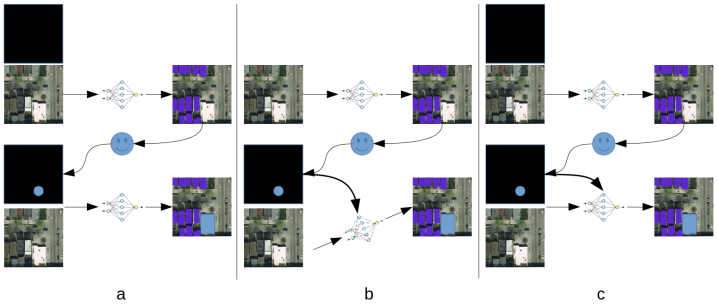


FIGURE 3.8 – Illustration de la méthode DISCA (c) qui combine une actualisation des poids comme dans WTP (b) et DISIR (a).

rithme DISCA [226].

3.3.2.1 Actualiser un modèle DISIR à la volée

DISCA est une amélioration du paradigme DISIR qui combine DISIR et un léger alignement des poids du modèle effectué sur les points cliqués (dont la vérité terrain est fournie par l'utilisateur).

Ce changement est important car il permet en théorie de gérer un transfert d'un jeu de données connu vers d'un jeu de données inconnu sans la nécessité d'avoir les mêmes classes, le même capteur, la même résolution...

Cependant, il convient de rappeler que l'apprentissage faiblement supervisé sur des points a déjà été évalué en section 3.2 et que la conclusion n'était pas satisfaisante. Mais justement, ici, le modèle ne cherche **pas** à faire un apprentissage faiblement supervisé sur des points comme *what's the point* (WTP) [110], le modèle cherche *juste* à actualiser un modèle DISIR comme l'indique la table 3.2 et/ou la figure 3.8.

Comparé à DISIR, DISCA offre des gains même sur des données relativement éloignées. Sur l'exemple de transfert AIRS vers POTSDAM, DISCA atteint très rapidement des performances élevées puis rattrape doucement celles de l'apprentissage supervisé classique dans le domaine cible comme le montre la

est remontée autour de 85%.

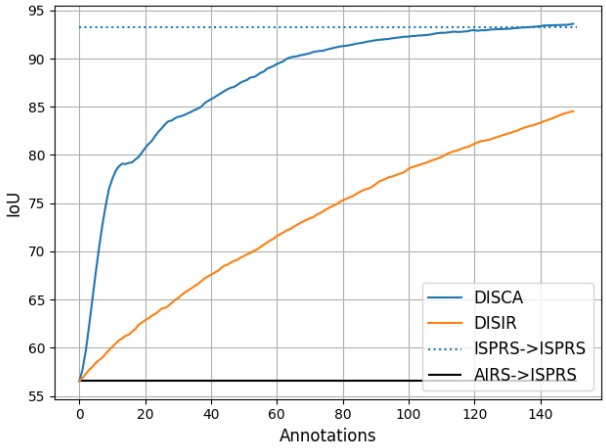


FIGURE 3.9 – Illustration de l'efficacité de l'approche DISCA : un modèle appris sur AIRS est transféré sur POTSDAM. Quelques clics suffisent pour grandement augmenter la pertinence des cartes produites.

figure 3.9. Inversement, on voit que DISIR a du mal à rattraper la performance de l'apprentissage supervisé et que sa performance augmente bien plus lentement que DISCA.

Par ailleurs, qualitativement, DISCA offre l'opportunité (et la menace voir 3.3.2.3) de modification de large amplitude. Par exemple, dans le jeu de données POTSDAM (test), il y a un seul bâtiment parking avec des voitures sur le toit (voir figure 3.10). Cependant, comme ce n'est pas le cas de tous les autres bâtiments et que toutes les autres voitures sont sur de la route, les modèles appris sur POTSDAM (apprentissage) vont difficilement pouvoir traiter ce cas. Ce problème persiste même avec DISIR qui n'apporte qu'une correction très locale dans ce cas (tant la classe indiquée par l'utilisateur est contraire à la classe prédite). Par contre, DISCA est capable de réaligner ses poids en quelques clics ce qui permet de traiter rapidement ce cas (et potentiellement tous les cas similiaires qui seraient présents dans les données cibles).

Ces résultats sont significatifs tant d'un point de vue académique qu'industriel. D'ailleurs, ces deux logiciels DISIR et DISCA sont utilisés par l'entreprise *Alteia* sur des problèmes privés, et, ils sont par ailleurs en cours d'intégration dans une chaîne de production d'annotation au sein de l'entreprise *CSgroup*.

3.3.2.2 La stabilisation de DISCA

Enfin, des travaux plus récents [218] utilisent DISCA avec de nouvelles classes. Cette approche (appelée ICSS dans [218]) est conceptuellement identique à DISCA (à partir du moment où on réactualise les poids, on peut rajouter une classe) mais nécessite d'utiliser des régularisations beaucoup plus avancées

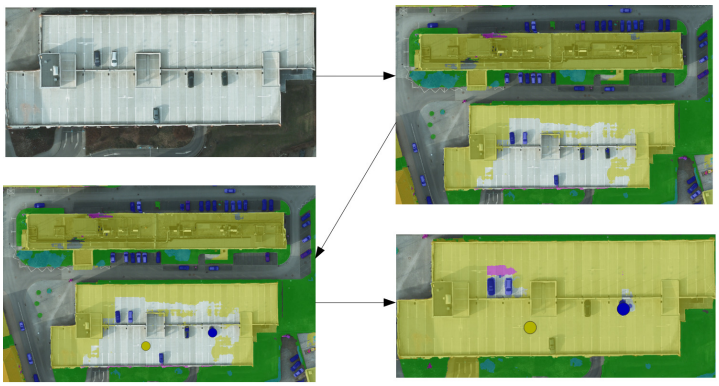
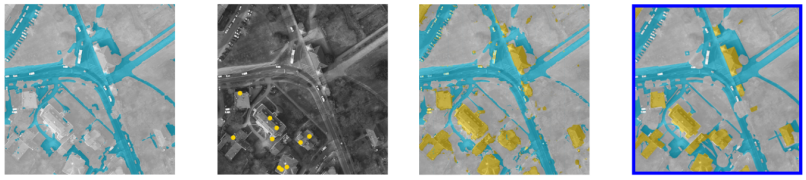


FIGURE 3.10 – Ce bâtiment est le seul de POTSDAM avec des voitures sur le toit, ce qui conduit à des erreurs - cependant rattrapables en deux clics avec DISCA.



(a) Initial prediction (b) New class annotations (c) New prediction (d) Ground-truth

FIGURE 3.11 – Illustration de l'utilisation de ICSS pour annoter la classe toit à partir d'un modèle préappris uniquement avec la classe route. Plus qualitativement, la performance d'un apprentissage bi-classe est de 84% alors que l'approche ICSS apprenant la classe toit à la volée arrive néanmoins à 80% d'IoU.

pour éviter une divergence du modèle.

Cette version de DISCA, qui peut conduire à des résultats intéressants comme ceux de la figure 3.11, est aussi l'occasion de souligner quelques instabilités persistantes dans DISCA.

Ce problème d'instabilité tient au fait que l'apprentissage de DISIR équilibre deux objectifs : produire une segmentation sémantique et être capable d'exploiter les points cliqués. Mais, l'actualisation des poids dans DISCA (ou ICSS) tend à favoriser uniquement la production d'une carte de segmentation sémantique (par ailleurs focalisée sur les derniers exemples vus). Ce biais est traité localement par l'utilisation de régularisations importantes sans lesquelles les approches DISCA et ICSS sont fortement sous-performantes (d'ailleurs une partie importante des efforts de développement nécessaires à [226, 218] portent sur la sélection et le réglage des bonnes régularisations). Cependant, cette solution locale ne peut tenir longtemps dans la durée, et au bout d'un grand nombre de clics, il est clair que DISCA tend à diverger.

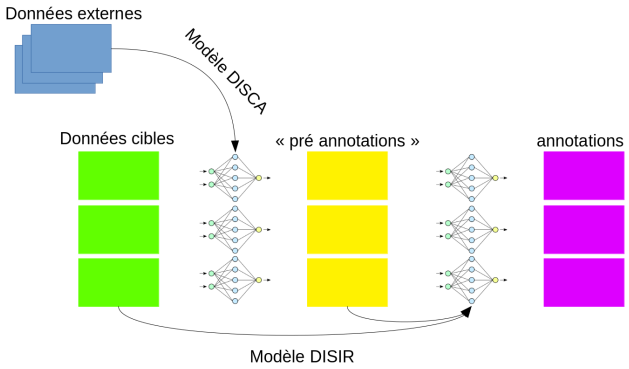


FIGURE 3.12 – DISCA souffre d'instabilité et ne peut pas traiter un trop gros volume de données en un seul passage. Cependant, une approche système consistant à utiliser DISCA puis DISIR sur le jeu de données ainsi créé peut gérer ces instabilités.

Cependant, l'important est plus de profiter de la fenêtre de stabilité de DISCA pour corriger les erreurs principales de sorte à ce qu'on puisse considérer les sorties comme une pré-vérité terrain. En fonction de la qualité des annotations désirées, cette étape peut être suffisante : on peut penser aux nombreuses bases de données académiques bruitées pourtant très utilisées comme INRIA (voir figure 3.3).

Mais, si la qualité des annotations reste insuffisante, il est alors toujours possible de réaliser un deuxième passage par exemple avec DISIR. Ainsi, via une telle approche *système* (comme par exemple celle illustrée en figure 3.6), on peut largement espérer traiter les potentielles instabilités de DISCA. DISIR de son côté n'a aucune instabilité car seules les entrées sont modifiées.

Actuellement, cet aspect système peut difficilement être évalué à l'aide de bases de données académiques car les performances obtenues avec DISCA sur la plupart des bases de données tombent dans le bruit de mesure due à l'imprécision de la vérité terrain. Précisément, on peut renouveler ici l'interrogation de la pertinence d'utiliser une métrique globale type IoU pour mesurer les performances sachant que deux humains annotant une même donnée auront probablement une IoU de seulement 90%. Dans ce contexte, il est difficile d'évaluer si la sortie de DISCA est de qualité suffisante pour être considérée comme une annotation (ce qui est notre objectif final). Cependant, les évaluations de DISIR sur INRIA montrent qu'il est capable de supporter une vérité terrain bruitée. Ce qui tend à montrer que l'approche système *DISCA puis DISIR* de la figure 3.12 est une approche viable. On voit donc que ce problème d'instabilité de DISCA n'est pas forcément problématique dans une approche système (même si une telle approche n'est pas aussi simple et/ou élégante qu'on pourrait l'espérer).

FIGURE 3.13 – Illustration d'une limite de l'approche DISCA : il est possible d'annoter une base proche de bases existantes (par exemple de POTSDAM [163] au centre vers DOTA [87] à gauche) mais ce transfert est douteux pour des images plus lointaines (comme des images de propergol [244]).

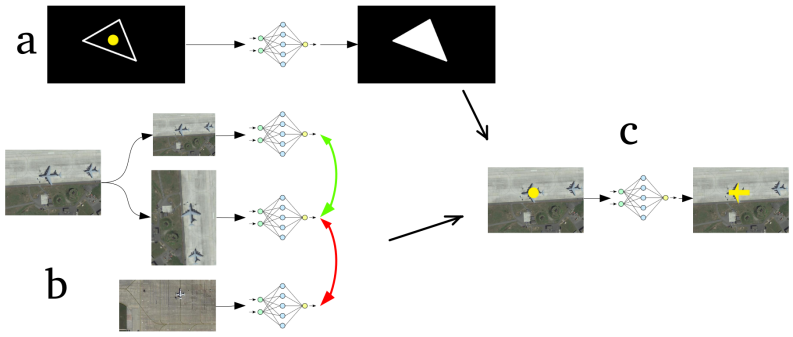
3.4 Perspectives d'industrialisation

Les résultats présentés précédemment démontrent l'efficacité des approches DISIR, DISCA (et ICSS) [225, 226, 218] pour tirer parti de clics de l'utilisateur. Ces travaux devraient conduire à des réalisations industrielles dans un contexte d'annotation d'images de télédétection.

Cependant, l'utilisation de ces logiciels est actuellement limitée par la nécessité de disposer d'un pré-apprentissage suffisamment proche des données cibles. Certes, ICSS permet d'ajouter une classe dans le processus, et, DISCA a montré une capacité à se réaligner sur des nouvelles données. Néanmoins, la différence de domaine entre POTSDAM [163] et un autre jeu de télédétection (par exemple DOTA [87]) est nettement moins grande qu'entre POTSDAM et les images de segmentation de propergol du chapitre 1 [244] comme souligné par la figure 3.13. Dit autrement, ce qui fait la force de l'outil est de combiner une prédiction déjà pertinente avec une entrée utilisateur. Mais, le revers de cette médaille et la nécessité de disposer d'une telle prédiction, ce qui cantonne aujourd'hui ces approches à des domaines suffisamment traités par la communauté (comme la télédétection).

Pour aller plus loin, il faudrait donc réussir à fusionner les approches type GrabCut (segmentation) et DISIR/DISCA (segmentation sémantique). Plus précisément, on voudrait un outil qui a le comportement d'un GrabCut avant la première interaction avec l'utilisateur mais qui converge vers un comportement type DISIR/DISCA après un certain nombre de clics. On veut donc un système suffisamment flexible pour faire coexister une notion de segmentation (avec un degré de sémantique variable) et la gestion du clic utilisateur.

Cependant, il n'est pas trivial de disposer d'un modèle assez malléable pour gérer plusieurs degrés de sémantique tout en s'adaptant à des données quelconques. Par exemple, une idée serait d'entraîner un réseau de neurones à prédire la sortie de GrabCut sur les données cibles. Ainsi, l'algorithme se prépare à avoir un comportement initial à la GrabCut tout en ayant le potentiel de tendre



La figure utilise des images googlemap (Ramstein, Allemagne et Andrews Air force, États-Unis). La figure b illustre le paradigme du contrastive learning : le modèle cherche à produire une représentation latente plus ou moins invariante à des *data transformations* (cette proximité est illustrée par la double flèche verte) mais cependant distantes des autres images (cette distance est illustrée par la double flèche rouge).

FIGURE 3.14 – Perspective sur la mécanisation de l'annotation en un clic : l'utilisation d'un pré-apprentissage couplant la gestion des clics (a) et des méthodes d'auto-supervision (b) permettrait de tirer parti des clics de l'utilisateur (c) comme dans DISIR/DISCA sans avoir besoin d'une base préexistante.

vers DISIR/DISCA. Cependant, il n'est pas clair que l'espace latent construit par une supervision qui émule GrabCut soit suffisamment déplié pour ensuite venir capturer la sémantique et donc que le modèle puisse converger vers un comportement de type DISIR/DISCA.

Mais, des techniques récentes d'auto-supervision - *self supervised learning* en anglais - semblent offrir la possibilité d'apprendre un tel espace latent.

Quelques mots d'abord sur l'auto-supervision qui apparaît comme une des perspectives les plus saillantes de la vision par ordinateur. L'utilisation de données non annotées en apprentissage est quelque choses de classique puisque les approches pré apprentissage profond étaient souvent basées [169, 172] sur des dictionnaires calculés à l'aide de méthodes de regroupement de type K moyenne [205]. Par ailleurs, l'auto-encodage a été testé avec des réseaux de neurones [157] dès leur apparition. Cette technique consiste à optimiser un modèle pour qu'il régresse sa propre entrée avec la contrainte de passer par un code de dimension inférieure. Cela le force à sélectionner l'information permettant la meilleure reconstruction moyenne. Cette idée d'auto-encodage peut ainsi être utilisée en soutien à une approche supervisée [29]. Cependant, ce type d'approche ne s'est pas imposé.

Alternativement, l'idée de l'auto-supervision est de considérer un problème basé uniquement sur la donnée (comme l'auto-encodage) mais *plus difficile* (forçant ainsi le modèle à construire un espace latent efficace). Un exemple simple en traitement vidéo, est celui de la technique de régression de l'image suivante [136] - *next frame prediction* en anglais - dans lequel le modèle essaye de ré-

gresser l'image suivante à partir des images passées. Il s'agit d'un problème difficile qui nécessite de capturer les corrélations spatio-temporelles au sein des vidéos naturelles. Typiquement, l'algorithme gagne à construire une représentation sous-jacente d'estimation de flot optique [12] connu pour être utile en classification sémantique de vidéos avec [116] ou sans apprentissage profond [160]. Récemment, une approche [30] dite de *contrastive learning* (illustrée en figure 3.14.b) a produit des résultats intéressants. Cette approche consiste à apprendre une représentation pour chaque image de sorte que ces représentations soient à la fois distantes entre-elles mais peu sensibles à des transformations simples des images (dites *data augmentation* en anglais). Cette approche produit des représentations très pertinentes pour la sémantique. Le fait de poursuivre l'apprentissage (*finetuning*) à partir de ces représentations permet même d'améliorer la performance par rapport à un apprentissage supervisé classique.

Cette auto-supervision parait pertinente pour dépasser la limite d'une simple supervision GrabCut probablement trop faible (comme pour l'auto-encodage simple) pour conduire une représentation latente qui converge au fil des annotations vers un comportement de type DISIR/DISCA. Cette perspective peut s'illustrer par la figure 3.14 et devrait donner lieu à un sujet de thèse.

3.5 Synthèse

Mes travaux sur l'apprentissage faiblement supervisé (apprendre avec des clics [230], apprendre à partir de vidéos [212], apprendre à partir de mélanges de base [215]) sont connexes à l'approche DISIR/DISCA issue de la thèse de Gaston Lenczner [226, 241, 225, 218] qui est intéressante car elle promet un déploiement industriel court terme. Cependant, ces travaux doivent se voir comme un tout à côté des travaux du chapitre précédent : dans les 2 cas, il s'agit d'augmenter l'acceptabilité des méthodes d'apprentissage profond : dans le chapitre 2 en diminuant le sentiment que ces algorithmes sont des boites noires, et, dans ce chapitre, en rendant plus facile la constitution des bases de données nécessaires pour utiliser ce type d'algorithme.

Actuellement, cette capacité à créer plus facilement des bases de données sera sûrement utile pour démultiplier l'utilisation d'approches d'apprentissage profond sur des données scientifiques éventuellement mieux contraintes que des données de type ImageNet. Ce point est au coeur de mon projet à 5 ans qui est présenté dans le prochain chapitre.

Chapitre 4

Perspectives

4.1 Les données scientifiques pour l'apprentissage profond

4.1.1 De la jungle à la loi

Le chapitre 1 présente des applications de méthodes d'apprentissage profond à différents problèmes dans les domaines de la mécanique des fluides, des matériaux ou de l'observation de la Terre. Ce sont donc des exemples de *comment l'apprentissage peut aider d'autres sciences*. Inversement, les connaissances et données scientifiques pourraient-elles aider l'apprentissage par ordinateur ? Pour cela, il faudrait que ces données apportent une contribution que n'offrent pas les données classiquement utilisées de type ImageNet (i.e. les données du numérique et/ou des réseaux sociaux). En effet, ces données du numérique sont aujourd'hui le moteur des augmentations de performance des méthodes d'apprentissage. Cependant, ces données sont aussi volatiles, bruitées et de pertinences très variables. Certains parlent de données *dans la jungle* (*in the wild* [178]) pour marquer leur manque de cohérence. C'est ce manque de cohérence (pas forcément représentatif des applications industrielles) qui pousse à perdre en *confiance* pour gagner en performances.

Inversement, des données mieux contraintes (par exemple par des lois physiques) et/ou moins volatiles pourraient permettre d'aboutir à des résultats plus stables, plus explicables, et ainsi, aider la communauté *apprentissage par ordinateur* à mieux comprendre ces algorithmes d'apprentissage profond. Par exemple, si on considère la donnée d'un champ de vitesse à l'intérieur d'un moteur d'avion, on a alors la certitude que ces données répondent alors à certaines lois et la possibilité d'échantillonner l'espace des possibles de façon comprehensible.

Dit autrement, aujourd'hui le bac à sable de la vision par ordinateur est ImageNet ou CIFAR10 [177], mais pour créer des IA de confiance, il serait peut être plus pertinent de se donner un bac à sable mieux cadré par exemple avec

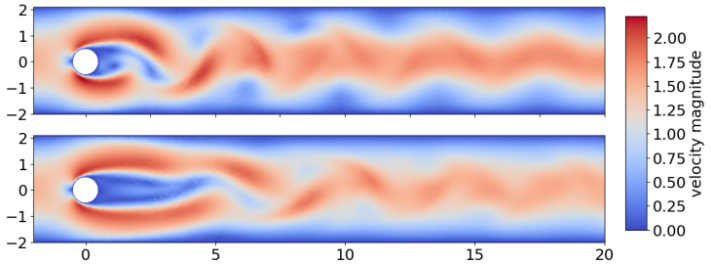


FIGURE 4.1 – Image extraite de [41] illustrant le contrôle d'écoulement : la première ligne est l'écoulement non contrôlé et la seconde l'écoulement contrôlé.

des données issues des sciences physiques.

Mon projet de recherche est d'ailleurs tourné sur l'intérêt que ce type de données offre pour augmenter notre compréhension des mécanismes internes des approches d'apprentissage profond, pour les éprouver, les calibrer, et donc, pour construire une confiance dans leur comportement. Cet axe de recherche est orienté par l'écosystème ONERA qui génère de grandes quantités de données scientifiques (et des données d'excellence dans ses domaines de prédilection comme la mécanique des fluides). Cette perspective est donc à la fois individuelle et collective, et je participe actuellement à la coordination d'un projet (principalement) interne à l'ONERA sur ce sujet. Le nom du projet *MASSIPH* pour *Maîtriser l'Apprentissage Statistique par la SImulation et les sciences PHysique* illustre bien cet objectif.

4.1.2 Un exemple concret de mécanique des fluides

Afin d'illustrer les propos précédents, je propose ici de décrire un travail prévu dans le cadre du projet *MASSIPH*. Il fait suite à des résultats ONERA récents [16] avec l'utilisation de méthodes de renforcement par apprentissage profond - *deep reinforcement learning* en anglais - appliquée à la mécanique des fluides. Quelques mots pour contextualiser ces travaux : il y a généralement des capteurs de pression, et des souffleurs à l'entrée des moteurs des avions. Ces souffleurs sont contrôlables et peuvent souffler ou aspirer de l'air. Évidemment, le volume d'air qui peut être soufflé ou aspiré est totalement négligeable par rapport au volume d'air entrant. Cependant, comme il s'agit d'un écoulement hautement non linéaire, l'action du souffleur peut empêcher des décollements et ainsi diminuer les frottements internes au moteur. Il est alors possible d'obtenir des économies de carburant significatives à l'aide d'un contrôle efficace des souffleurs, ce qui est évidemment souhaitable pour l'aviation (peut-être même plus en terme d'image qu'économiquement). L'optimisation du comportement du souffleur en fonction de l'état de l'écoulement (partiellement observable via les capteurs de pression) peut se voir comme un problème de renforcement illustré en figure 4.1.

Le paradigme de l'apprentissage par renforcement n'est pas particulièrement décrit dans ce manuscrit (bien que j'ai participé à des contributions sur ce sujet [223, 222] dans un contexte robotique ou de vision par ordinateur [220]) mais un point d'entrée peut être trouvé ici [195]. Dans les grandes lignes, le renforcement est une supervision de séquences. En effet, dans un problème de classification supervisé, chaque prédiction est indépendante des autres, et, à l'apprentissage, chaque prédiction du modèle est mise en regard de celle de l'humain. Dans un problème de renforcement, des séquences d'actions sont considérées (les séquences sont indépendantes mais au sein d'une séquence les prédictions impactent les suivantes). De plus, la supervision considérée est moins forte qu'en classification : chaque décision est uniquement associée à une récompense - *reward* en anglais - et l'optimisation de la politique consiste à maximiser l'espérance de récompense. Dans le cas du contrôle d'écoulement - *flow control* en anglais - on contrôle l'action du souffleur à chaque pas de temps et l'objectif est de maximiser le rendement du moteur, sachant que les actions précédentes influeront sur l'écoulement futur.

On peut alors utiliser différentes familles de modèles pour encoder la politique issue du renforcement. En particulier, le renforcement par apprentissage profond consiste à utiliser des réseaux de neurones profonds pour approximer la politique optimale [156]. C'est ce type d'approche qui a été testée pour le contrôle d'écoulement dans [16] et les performances obtenues dépassent significativement les performances d'un modèle classique (linéaire) plus simple.

Cependant, ces résultats ne sont pas forcément *acceptables* en l'état : ils gagnent à être améliorés par des approches d'IA de confiance et notamment d'explicabilité. Être capable de mieux comprendre la représentation interne d'un tel modèle notamment pour valider que cette représentation a un sens physique est indispensable pour faire accepter ce type de méthode à la fois à la communauté mécanique des fluides et aussi aux instances réglementaires (quand bien même un tel module de contrôle d'écoulements n'est pas critique puisqu'il n'agit que sur la consommation de carburant). Mais inversement, ce contexte, où on connaît les lois qui régissent l'évolution de l'état, est justement pertinent pour appliquer des méthodes d'explicabilité puisque on a un a priori sur la forme que devraient prendre les explications.

Il y a donc bien dans cette application un intérêt à la fois pour la mécanique des fluides (de disposer de modèles performants) et pour l'apprentissage (de disposer d'un cadre propice à l'application de méthode d'explicabilité).

Au delà de ce cas particulier, qui devrait conduire à un sujet de thèse (en collaboration avec des chercheurs de l'ONERA de mécanique des fluides), la perspective qui se dégage est bien la pertinence des données scientifiques mieux contraintes pour accélérer le développement des méthodes dites d'IA de confiance.

4.2 Télédétection et IA de confiance

4.2.1 La disponibilité des données en télédétection

Indépendamment de la stabilité des données scientifiques, une autre source de données pourrait être mieux explorée pour l'IA de confiance : il s'agit des données de télédétection dont la disponibilité et la structuration sont des atouts.

En effet, de plus en plus de données satellites passent en *open source* comme par exemple les satellites Sentinel-2 pour lesquels il est virtuellement possible d'avoir accès à toutes les données acquises depuis le lancement et ainsi avoir une banque de données représentatives de la distribution. Inversement, seuls les géants du numériques disposent de base de données aussi représentatives de leurs problèmes (mais ces données ne seront pas ouvertes aux académiques).

Ainsi, toutes les approches visant à estimer les biais de tirage et/ou les performances de généralisation gagneraient à utiliser des banques de données satellites plutôt que les 1000000 images d'Imagenet pour lesquels des travaux posent déjà la question de leur représentativité [64].

Un autre avantage est que ces données sont spatialisées géographiquement (et temporellement) et l'erreur d'intérêt est justement une erreur uniforme vis-à-vis de la géographie (et de la saison). Ainsi, évaluer les impacts de biais de tirage est alors possible en considérant des tirages plus ou moins bien répartis. Un sujet de thèse pourrait être lancé sur ce point avec le LNE [1].

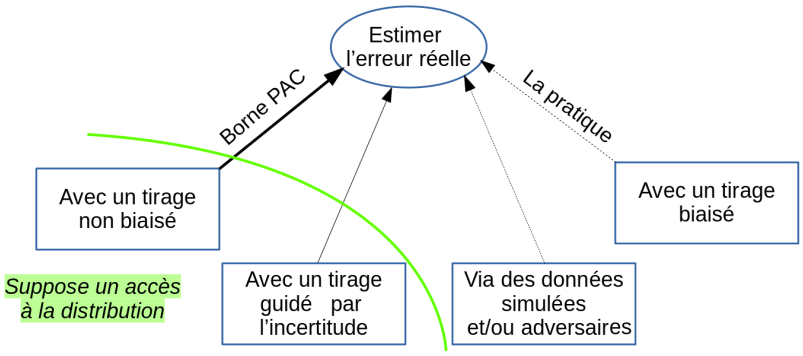
4.2.2 Évaluer les biais de tirages

Comme discuté en chapitre 2, sans hypothèse supplémentaire, il est nécessaire d'avoir un tirage identiquement distribué selon P pour évaluer l'erreur commise par un modèle $\int_{x,y(x)\neq \text{sign}(f(x))} P(x)\mathrm{d}x$. Cependant, dans la pratique, on ne cherche pas à traiter l'ensemble des problèmes y, P et on peut espérer avoir une approximation *acceptable* de l'erreur avec un tirage biaisé. Or comme discuté précédemment, en télédétection, on peut justement tirer selon P (puisqu'on peut disposer d'une base de données représentative de P) et aussi simuler facilement des tirages biaisés (notamment géographiquement) grâce à la structuration géographique des données. C'est donc un contexte où on peut évaluer empiriquement l'impact des biais de tirage.

D'ailleurs, il existe un continuum illustré dans la figure 4.2 entre le tirage identiquement distribué et le tirage naïf : on peut notamment penser aux approches de tirages dirigés [106] et/ou aux approches de synthèse de données. Mais, la télédétection est pertinente sur tout ce continuum.

Par exemple, la télédétection permet de simuler des tirages géographiquement biaisés mais aussi d'estimer la possibilité d'utiliser l'incertitude pour diriger des tirages. L'idée centrale est que (pour les distributions intéressantes) la valeur absolue de f corrèle (relativement) avec la probabilité de ne pas faire

1. Le rapport Villani a proposé que LNE devienne l'organisme chargé de l'autorisation de mise sur marché des IA.



Dans le cadre le plus contraint, on n'a qu'une campagne d'essai mais ni annotation ni donnée supplémentaire. Dans ce cas, la seule solution pour venir explorer le comportement sur d'autres données, c'est d'en générer. Parfois, on dispose de plus de données, et il est alors possible de réfléchir à un tirage dirigé voire à un tirage identiquement distribué si l'on peut disposer d'annotations.

FIGURE 4.2 – L'évaluation d'un modèle appris : l'accès à la distribution des données (facilité par exemple en télédétection) est nécessaire pour évaluer l'écart entre la pratique et les garanties théoriques.

d'erreurs. Cette corrélation se généralise et/ou transfère éventuellement mieux que la décision (associée au signe de f). On peut alors chercher des données qui pourraient être mal classées parmi les plus incertaines. Cette idée présente dès [199] vient à la fois des vecteurs supports (qui définissent l'hyperplan séparateur à eux seuls dans [198]), de l'apprentissage actif [180] - *active learning* en anglais - où l'incertitude est souvent utilisée comme approximation de la fonction d'acquisition [190] (on la retrouve aussi dans l'estimation des événements rares [137] voir annexe 5.4.6). Par ailleurs, une évaluation de différentes estimations de l'incertitude dans un cadre de télédétection a été réalisée dans le cadre de la thèse de Gaston Lenczner [209].

Autre exemple, la génération de données est une façon de disposer de données supplémentaires sur lesquelles on peut évaluer le comportement du réseau. Cependant, la question de savoir si les données ainsi produites sont plus générales que les données nécessaires pour apprendre ces générateurs n'est pas close et l'utilisation d'un contexte de télédetection peut doublement aider à répondre à cette question. En effet, la télédétection met parfois en oeuvre des données qu'il est potentiellement plus facile à simuler que des images naturelles (dont la distribution est probablement très creuse). Notamment, ce type d'approches appliquées à la simulation électromagnétique devrait être considéré dans le projet MASSIPH (l'expertise de l'ONERA en simulation électromagnétique étant établie). À noter que l'idée de coupler GAN [147] et simulation est présente dans [104]. La question de savoir si des GAN seuls sont suffisants est difficile. J'ai participé à des travaux préliminaires avec le LNE sur la génération de données

FIGURE 4.3 – Frise chronologique de mes travaux et encadrements passés et à venir.

à la fois réaliste et incertaine [227] dans une perspective de qualification. Cependant, ces travaux sont totalement inapplicables aujourd'hui en production même si d'autres travaux ONERA [8, 7] basés sur [54] et appliqués à des images de télédétection ont conduit à des résultats prometteurs.

Ainsi, la télédétection est pertinente pour évaluer empiriquement les écarts de mesure de performances sur l'ensemble du continuum de la figure 4.2 : que ce soit pour disposer d'un tirage uniforme, de tirages dirigés et/ou de simulation de données, la télédétection offre un cadre pertinent pour comparer ces évaluations vis-à-vis de l'évaluation naïve ou de l'évaluation non biaisée.

4.3 Conclusion

Ce chapitre clot ce manuscrit. Il en ressort mon projet de recherche à 5 ans notamment en 4 sujets de thèse qui sont rappelés de façon synthétique en annexe (5.4) qui émargent à l'IA de confiance et/ou à l'utilisation des données scientifiques ou de télédétection notamment produites à l'ONERA. Ce projet se construit sur des travaux antérieurs notamment associés à l'IA de confiance ou à des applications de méthodes d'apprentissage profond au sein de l'ONERA présentées dans les autres chapitre (SALAD publié à WACV et DISIR/DISCA qui représente une avancée potentiellement importante pour aider à l'annotation de problèmes de niche).

C'est à la lumière de la cohérence de ce projet à 5 ans et des travaux passés, illustrée par la figure 4.3, que je me présente à l'habilitation à diriger des recherches.

Chapitre 5

Bibliographie et Annexes

5.1 Références

(Articles dont je ne suis pas co-auteurs.)

[1] Diego Granziol, Mingtian Zhang et Nicholas Baskerville. *A Practical PAC-Bayes Generalisation Bound for Deep Learning*. 2022. url : `https://openreview.net/forum?id=mYaOK2og0tf`.

[2] Zhe Li, Josue Ortega Caro, Evgenia Rusak, Wieland Brendel, Matthias Bethge, Fabio Anselmi, Ankit B Patel, Andreas S Tolias et Xaq Pitkow. « Robust deep learning object recognition models rely on low frequency information in natural images ». In : *bioRxiv* (2022).

[3] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell et Saining Xie. « A ConvNet for the 2020s ». In : *arXiv preprint arXiv : 2201.03545* (2022).

[4] Ali Alqahtani, Xianghua Xie et Mark W Jones. « Literature Review of Deep Network Compression ». In : *Informatics*. T. 8. 4. Multidisciplinary Digital Publishing Institute. 2021, p. 77.

[5] Benoit Bonnet, Teddy Furon et Patrick Bas. « Generating Adversarial Images in Quantized Domains ». In : *IEEE Transactions on Information Forensics and Security* (2021).

[6] Sebastien Bubeck et Mark Sellke. « A universal law of robustness via isoperimetry ». In : *Advances in Neural Information Processing Systems* 34 (2021).

[7] Javiera Castillo-Navarro, Bertrand Le Saux, Alexandre Boulch et Sebastien Lefevre. « Classification and Generation of Earth Observation Images using a Joint Energy-Based Model ». In : *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*. IEEE. 2021, p. 2098-2101.

[8] Javiera Castillo-Navarro, Bertrand Le Saux, Alexandre Boulch et Sebastien Lefevre. « Energy-based models in earth observation : From generation to semi-supervised learning ». In : *IEEE Transactions on Geoscience and Remote Sensing* (2021).

[9] Liangyu Chen, Tong Yang, Xiangyu Zhang, Wei Zhang et Jian Sun. « Points as queries : Weakly semi-supervised object detection by points ». In : *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, p. 8823-8832.

[10] Arthur Claviere, Eric Asselin, Christophe Garion et Claire Pagetti. « Safety Verification of Neural Network Controlled Systems ». In : *51st Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops, DSN Workshops 2021, Taipei, Taiwan, June 21-24, 2021*. IEEE, 2021, p. 47-54.

[11] Herve Delseny et al. « White paper machine learning in certified systems ». In : *arXiv preprint arXiv :2103.10529* (2021).

[12] Pierre Godet, Alexandre Boulch, Aurélien Plyer et Guy Le Besnerais. « STaRFlow : A SpatioTemporal Recurrent Cell for Lightweight Multi-Frame Optical Flow Estimation ». In : *2020 25th International Conference on Pattern Recognition (ICPR)*. 2021, p. 2462-2469.

[13] Benjamin Kellenberger, Devis Tuia et Dan Morris. « Introducing AIDE : a Software Suite for Annotating Images with Deep and Active Learning Assistance ». In : *EGU General Assembly Conference Abstracts*. 2021, EGU21-12065.

[14] Yan Liu, Zhijie Zhang, Li Niu, Junjie Chen et Liqing Zhang. « Mixed supervised object detection by transferring mask prior and semantic similarity ». In : *Advances in Neural Information Processing Systems* 34 (2021).

[15] Juliette Mattioli, Francois Terrier, Loic Cantat, Julien Chiaroni, Michel Barreteau, Yannick Bonhomme, Christophe Guettier et Christophe Alix. « IA de confiance : condition necessaire pour le déploiement de l'IA dans les systemes de defense ». In : *APIA (Conference Nationale sur les Applications Pratiques de l'Intelligence Artificielle)*. 2021.

[16] Romain Paris, Samir Beneddine et Julien Dandois. « Robust flow control and optimal sensor placement using deep reinforcement learning ». In : *Journal of Fluid Mechanics* 913 (2021), A25.

[17] Albert Pumarola, Enric Corona, Gerard Pons-Moll et Francesc Moreno-Noguer. « D-nerf : Neural radiance fields for dynamic scenes ». In : *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, p. 10318-10327.

[18] Elias Ramzi, Nicolas Thome, Clément Rambour, Nicolas Audebert et Xavier Bitot. « Robust and Decomposable Average Precision for Image Retrieval ». In : *Advances in Neural Information Processing Systems* 34 (2021).

[19] S. Ravuri, K. Lenc et M. et al. Willson. « Skilful precipitation nowcasting using deep generative models of radar. » In : *Nature 597, 672–677 (2021)*. 2021.

[20] Wojciech Sirko et al. « Continental-scale building detection from high resolution satellite imagery ». In : *arXiv preprint arXiv :2107.12283* (2021).

[21] Paul Viallard, Eric Guillaume VIDOT, Amaury Habrard et Emilie Morvant. « A PAC-Bayes Analysis of Adversarial Robustness ». In : *Advances in Neural Information Processing Systems* 34 (2021).

[22] Pedro Stefanin Volpiani, Morten Meyer, Lucas Franceschini, Julien Dandois, Florent Renac, Emeric Martin, Olivier Marquet et Denis Sipp. « Machine learning-augmented turbulence modeling for RANS simulations of massively separated flows ». In : *Physical Review Fluids* 6.6 (2021), p. 064607.

[23] Chia-Hung Yuan et Shan-Hung Wu. « Neural tangent generalization attacks ». In : *International Conference on Machine Learning*. PMLR. 2021, p. 12230-12240.

[24] Yue Zhou, Xiaofang Hu, Jiaqi Han, Lidan Wang et Shukai Duan. « High frequency patterns play a key role in the generation of adversarial examples ». In : *Neurocomputing* 459 (2021), p. 131-141.

[25] Lea Berthomier, Bruno Pradel et Lior Perez. « Cloud Cover Nowcasting with Deep Learning ». In : *2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA)*. 2020, p. 1-6.

[26] Lucas Beyer, Olivier J Henaff, Alexander Kolesnikov, Xiaohua Zhai et Aaron van den Oord. « Are we done with imagenet ? » In : *arXiv preprint arXiv :2006.07159* (2020).

[27] Alexandre Boulch. « ConvPoint : Continuous convolutions for point cloud processing ». In : *Computers & Graphics* 88 (2020), p. 24-34.

[28] Andrew Brown, Weidi Xie, Vicky Kalogeiton et Andrew Zisserman. « Smooth-ap : Smoothing the path towards large-scale image retrieval ». In : *European Conference on Computer Vision*. Springer. 2020, p. 677-694.

[29] Javiera Castillo-Navarro, Bertrand Le Saux, Alexandre Boulch et Sebastien Lefevre. « On auxiliary losses for semi-supervised semantic segmentation ». In : *ECML PKDD 2020 : European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*. 2020.

[30] Ting Chen, Simon Kornblith, Mohammad Norouzi et Geoffrey Hinton. « A simple framework for contrastive learning of visual representations ». In : *International conference on machine learning*. PMLR. 2020, p. 1597-1607.

[31] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi et Geoffrey Hinton. « Big Self-Supervised Models are Strong Semi-Supervised Learners ». In : *arXiv preprint arXiv :2006.10029* (2020).

[32] European Commission. « White Paper on Artificial Intelligence : A European approach to excellence and trust ». In : *Com (2020) 65 Final* (2020).

[33] Yukun Ding, Jinglan Liu, Jinjun Xiong et Yiyu Shi. « Revisiting the Evaluation of Uncertainty Estimation and Its Application to Explore Model Complexity-Uncertainty Trade-Off ». In : *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2020, p. 4-5.

[34] Alexey Dosovitskiy et al. « An image is worth 16x16 words : Transformers for image recognition at scale ». In : *arXiv preprint arXiv :2010.11929* (2020).

[35] Vivien Sainte Fare Garnot, Loic Landrieu, Sebastien Giordano et Nesrine Chehata. « Satellite image time series classification with pixel-set encoders and temporal self-attention ». In : *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, p. 12325-12334.

[36] Yiding Jiang et al. « Neurips 2020 competition : Predicting generalization in deep learning ». In : *arXiv preprint arXiv :2012.07976* (2020).

[37] Fabian Kuppers, Jan Kronenberger, Amirhossein Shantia et Anselm Haselhoff. « Multivariate confidence calibration for object detection ». In : *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2020, p. 326-327.

[38] Alexander Levine et Soheil Feizi. « Deep partition aggregation : Provable defense against general poisoning attacks ». In : *arXiv preprint arXiv :2006.14768* (2020).

[39] Junnan Li, Caiming Xiong, Richard Socher et Steven Hoi. « Towards noise-resistant object detection with noisy annotations ». In : *arXiv preprint arXiv :2003.01285* (2020).

[40] Trisha Mahoney, Kush R Varshney et Michael Hind. *How to Measure and Reduce Unwanted Bias in Machine Learning*. O'Reilly, 2020.

[41] Jean Rabault, Feng Ren, Wei Zhang, Hui Tang et Hui Xu. « Deep reinforcement learning in fluid mechanics : A promising method for both active flow control and shape optimization ». In : *Journal of Hydrodynamics* 32.2 (2020), p. 234-246.

[42] Mengmeng Xu, Chen Zhao, David S. Rojas, Ali Thabet et Bernard Ghanem. « G-TAD : Sub-Graph Localization for Temporal Action Detection ». In : *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020.

[43] Sukmin Yun, Jongjin Park, Kimin Lee et Jinwoo Shin. « Regularizing class-wise predictions via self-knowledge distillation ». In : *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, p. 13876-13885.

[44] Hanwei Zhang, Yannis Avrithis, Teddy Furon et Laurent Amsaleg. « Smooth adversarial examples ». In : *EURASIP Journal on Information Security* 2020.1 (2020), p. 1-12.

[45] Zhuo Zheng, Yanfei Zhong, Junjue Wang et Ailong Ma. « Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery ». In : *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, p. 4096-4105.

[46] Cem Anil, James Lucas et Roger Grosse. « Sorting out Lipschitz function approximation ». In : *International Conference on Machine Learning*. PMLR. 2019, p. 291-301.

[47] Samir Beneddine. « Comparison of Deep Learning strategies for flow field reconstruction from wall measurements ». In : *APS Division of Fluid Dynamics Meeting Abstracts*. 2019, p. L23-004.

[48] Maxime Bucher, Tuan-Hung Vu, Matthieu Cord et Patrick Perez. « Zero-shot semantic segmentation ». In : *Advances in Neural Information Processing Systems* 32 (2019).

[49] Qi Chen, Lei Wang, Yifan Wu, Guangming Wu, Zhiling Guo et Steven L Waslander. « Aerial imagery for roof segmentation : A large-scale dataset towards automatic mapping of buildings ». In : *ISPRS Journal of Photogrammetry and Remote Sensing*. T. 147. Elsevier. 2019, p. 42-55.

[50] Jeremy M Cohen, Elan Rosenfeld et J. Zico Kolter. *Certified Adversarial Robustness via Randomized Smoothing*. 2019. arXiv : `1902.02918 [cs.LG]`.

[51] Charles Corbiere, Nicolas Thome, Avner Bar-Hen, Matthieu Cord et Patrick Perez. « Addressing failure prediction by learning model confidence ». In : *Advances in Neural Information Processing Systems*. 2019, p. 2902-2913.

[52] Maxime Ferrera, Vincent Creuze, Julien Moras et Pauline Trouve-Peloux. « AQUALOC : An underwater dataset for visual–inertial–pressure localization ». In : *The International Journal of Robotics Research* 38.14 (2019), p. 1549-1559.

[53] Leopoldina Fortunati, Giuseppe Lugano et Anna Maria Manganelli. « European perceptions of autonomous and robotized cars ». In : *International Journal of Communication* 13.2 (2019), p. 2728-2747.

[54] Will Grathwohl, Kuan-Chieh Wang, Jorn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi et Kevin Swersky. « Your classifier is secretly an energy based model and you should treat it like one ». In : *arXiv preprint arXiv :1912.03263* (2019).

[55] Dan Hendrycks et Thomas Dietterich. « Benchmarking neural network robustness to common corruptions and perturbations ». In : *arXiv preprint arXiv :1903.12261* (2019).

[56] Yanping Huang et al. « Gpipe : Efficient training of giant neural networks using pipeline parallelism ». In : *Advances in neural information processing systems* 32 (2019).

[57] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran et Aleksander Madry. « Adversarial examples are not bugs, they are features ». In : *Advances in neural information processing systems* 32 (2019).

[58] Longlong Jing, Yucheng Chen et Yingli Tian. « Coarse-to-fine semantic segmentation from image-level labels ». In : *IEEE Transactions on Image Processing* 29 (2019), p. 225-236.

[59] Benjamin Kellenberger, Diego Marcos et Devis Tuia. « When a few clicks make all the difference : Improving weakly-supervised wildlife detection in UAV images ». In : *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2019.

[60] Xi Li, Huimin Ma et Xiong Luo. « Weaklier supervised semantic segmentation with only one image level annotation per category ». In : *IEEE Transactions on Image Processing* 29 (2019), p. 128-141.

[61] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding et Shilei Wen. « Bmn : Boundary-matching network for temporal action proposal generation ». In : *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, p. 3889-3898.

[62] Chenri Ni, Nontawat Charoenphakdee, Junya Honda et Masashi Sugiyama. « On the Calibration of Multiclass Classification with Rejection ». In : *Advances in Neural Information Processing Systems*. 2019, p. 2586-2596.

[63] Matthieu Nugue. « Outils pour l'etude conjointe par simulation et traitement d'images experimentales de la combustion de particules d'aluminium utilisees dans les propergols solides ». Thèse de doct. Universite Paris-Saclay (ComUE), 2019.

[64] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt et Vaishaal Shankar. « Do imagenet classifiers generalize to imagenet ? » In : *International Conference on Machine Learning*. PMLR. 2019, p. 5389-5400.

[65] Jerome Revaud, Jon Almazan, Rafael S Rezende et Cesar Roberto de Souza. « Learning with average precision : Training image retrieval with a listwise loss ». In : *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, p. 5107-5116.

[66] Aniruddha Saha, Akshayvarun Subramanya, Koninika Patil et Hamed Pirsiavash. « Adversarial patches exploiting contextual reasoning in object detection ». In : *arXiv preprint arXiv :1910.00068* (2019).

[67] Mingxing Tan et Quoc Le. « Efficientnet : Rethinking model scaling for convolutional neural networks ». In : *International conference on machine learning*. PMLR. 2019, p. 6105-6114.

[68] Ze Yang, Shaohui Liu, Han Hu, Liwei Wang et Stephen Lin. « Reppoints : Point set representation for object detection ». In : *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, p. 9657-9666.

[69] M Mutlu YAPICI, Adem Tekerek et Nurettin TOPALOĞLU. « Literature review of deep learning research areas ». In : *Gazi Muhendislik Bilimleri Dergisi (GMBD)* 5.3 (2019), p. 188-215.

[70] Runhao Zeng, Wenbing Huang, Mingkui Tan, Yu Rong, Peilin Zhao, Junzhou Huang et Chuang Gan. « Graph Convolutional Networks for Temporal Action Localization ». In : *ICCV*. 2019.

[71] Anish Athalye, Nicholas Carlini et David Wagner. « Obfuscated Gradients Give a False Sense of Security : Circumventing Defenses to Adversarial Examples ». In : *Proceedings of the 35th International Conference on Machine Learning*. T. 80. PMLR, 2018, p. 274-283.

[72] Alexandre Boulch, Pauline Trouve, Elise Koeniguer, Fabrice Janez et Bertr Le Saux. « Learning speckle suppression in SAR images without ground truth : application to sentinel-1 time-series ». In : *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*. IEEE. 2018, p. 2366-2369.

[73] Maxime Bucher, Stephane Herbin et Frederic Jurie. « Semantic bottleneck for computer vision tasks ». In : *Asian Conference on Computer Vision*. Springer. 2018, p. 695-712.

[74] Yu Wei Chao, Sudheendra Vijayanarasimhan, Bryan Seybold, David A. Ross, Jia Deng et Rahul Sukthankar. « Rethinking the Faster R-CNN Architecture for Temporal Action Localization ». In : *CoRR* abs/1804.07667 (2018).

[75] Sungil Choi, Seungryong Kim, Kihong Park et Kwanghoon Sohn. « Learning descriptor, confidence, and depth estimation in multi-view stereo ». In : *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2018, p. 276-282.

[76] Alexandra Chouldechova et Aaron Roth. « The frontiers of fairness in machine learning ». In : *arXiv preprint arXiv :1810.08810* (2018).

[77] Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt et Aleksander Madry. « A rotation and a translation suffice : Fooling cnns with simple transformations ». In : (2018).

[78] Todd Huster, Cho-Yu Jason Chiang et Ritu Chadha. « Limitations of the Lipschitz constant as a defense against adversarial examples ». In : *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2018, p. 16-29.

[79] Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam et Dmitry Kalenichenko. « Quantization and training of neural networks for efficient integer-arithmetic-only inference ». In : *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, p. 2704-2713.

[80] Mathieu Laroze, Romain Dambreville, Chloe Friguet, Ewa Kijak et Sebastien Lefevre. « Active learning to assist annotation of aerial images in environmental surveys ». In : *2018 international conference on content-based multimedia indexing (CBMI)*. IEEE. 2018, p. 1-6.

[81] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang et Ming Yang. « BSN : Boundary Sensitive Network for Temporal Action Proposal Generation ». In : *CoRR* abs/1806.02964 (2018).

[82] Shigang Liu, Jun Zhang, Yu Wang, Wanlei Zhou, Yang Xiang et Olivier De Vel. « A data-driven attack against support vectors of svm ». In : *Proceedings of the 2018 on Asia Conference on Computer and Communications Security*. 2018, p. 723-734.

[83] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras et Adrian Vladu. « Towards Deep Learning Models Resistant to Adversarial Attacks ». In : *International Conference on Learning Representations*. ICLR. 2018.

[84] Antonio Polino, Razvan Pascanu et Dan Alistarh. « Model compression via distillation and quantization ». In : *arXiv preprint arXiv : 1802.05668* (2018).

[85] Ali Shafahi, W Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras et Tom Goldstein. « Poison frogs ! targeted clean-label poisoning attacks on neural networks ». In : *arXiv preprint arXiv :1804.00792* (2018).

[86] Eric Wong et Zico Kolter. « Provable defenses against adversarial examples via the convex outer adversarial polytope ». In : *International Conference on Machine Learning*. PMLR. 2018, p. 5286-5295.

[87] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo et Liangpei Zhang. « DOTA : A large-scale dataset for object detection in aerial images ». In : *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, p. 3974-3983.

[88] Navaneeth Bodla, Bharat Singh, Rama Chellappa et Larry S Davis. « Soft-NMS–improving object detection with one line of code ». In : *Proceedings of the IEEE international conference on computer vision*. 2017, p. 5561-5569.

[89] Tom B Brown, Dandelion Mane, Aurko Roy, Martin Abadi et Justin Gilmer. « Adversarial patch ». In : *arXiv preprint arXiv : 1712.09665* (2017).

[90] Shyamal Buch, Victor Escorcia, Bernard Ghanem, Li Fei-Fei et Juan Carlos Niebles. « End-to-End, Single-Stream Temporal Action Detection in Untrimmed Videos ». In : *Proceedings of the British Machine Vision Conference (BMVC)*. 2017.

[91] Shyamal Buch, Victor Escorcia, Chuanqi Shen, Bernard Ghanem et Juan Carlos Niebles. « SST : Single-Stream Temporal Action Proposals ». In : *CVPR*. 2017.

[92] Noelie Cherrier, Thibaut Castaings et Alexandre Boulch. « Deep sequence-to-sequence neural networks for ionospheric activity map prediction ». In : *International Conference on Neural Information Processing*. Springer. 2017, p. 545-555.

[93] Xiyang Dai, Bharat Singh, Guyue Zhang, Larry S. Davis et Yan Qiu Chen. « Temporal Context Network for Activity Localization in Videos ». In : *The IEEE International Conference on Computer Vision (ICCV)*. 2017.

[94] Jiyang Gao, Zhenheng Yang, Kan Chen, Chen Sun et Ram Nevatia. « TURN TAP : Temporal Unit Regression Network for Temporal Action Proposals ». In : *The IEEE International Conference on Computer Vision (ICCV)*. 2017.

[95] Jiyang Gao, Zhenheng Yang et Ram Nevatia. « Cascaded Boundary Regression for Temporal Action Detection ». In : *BMVC*. 2017.

[96] F. C. Heilbron, W. Barrios, V. Escorcia et B. Ghanem. « SCC : Semantic Context Cascade for Efficient Action Detection ». In : *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, p. 3175-3184.

[97] Rui Hou, Rahul Sukthankar et Mubarak Shah. *Real-Time Temporal Action Localization in Untrimmed Videos by Sub-Action Discovery*. 2017.

[98] Guy Katz, Clark Barrett, David L Dill, Kyle Julian et Mykel J Kochenderfer. « Reluplex : An efficient SMT solver for verifying deep neural networks ». In : *International Conference on Computer Aided Verification*. Springer. 2017, p. 97-117.

[99] Jian-Hao Luo, Jianxin Wu et Weiyao Lin. « Thinet : A filter level pruning method for deep neural network compression ». In : *Proceedings of the IEEE international conference on computer vision*. 2017, p. 5058-5066.

[100] Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat et Pierre Alliez. « Can Semantic Labeling Methods Generalize to Any City ? The INRIA Aerial Image Labeling Benchmark ». In : *International Geoscience and Remote Sensing Symposium*. IEEE. 2017, p. 3226-3229.

[101] Luis Munoz-Gonzalez, Battista Biggio, Ambra Demontis, Andrea Paudice, Vasin Wongrassamee, Emil C Lupu et Fabio Roli. « Towards poisoning of deep learning algorithms with back-gradient optimization ». In : *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*. ACM. 2017.

[102] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Zeynel Berkay Celik et Ananthram Swami. « Practical black-box attacks against machine learning ». In : *Proc. ACM Asia Conference on Computer and Communications Security*. 2017.

[103] Zheng Shou, Jonathan Chan, Alireza Zareian, Kazuyuki Miyazawa et Shih-Fu Chang. « CDC : Convolutional-De-Convolutional Networks for Precise Temporal Action Localization in Untrimmed Videos ». In : *CVPR*. 2017.

[104] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang et Russell Webb. « Learning from simulated and unsupervised images through adversarial training ». In : *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, p. 2107-2116.

[105] Tatiana Tommasi, Novi Patricia, Barbara Caputo et Tinne Tuytelaars. « A deeper look at dataset bias ». In : *Domain adaptation in computer vision applications*. Springer, 2017, p. 37-55.

[106] Van-Tinh Tran. « Selection Bias Correction in Supervised Learning with Importance Weight ». Thèse de doct. Université de Lyon, juill. 2017.

[107] Huijuan Xu, Abir Das et Kate Saenko. « R-C3D : Region Convolutional 3D Network for Temporal Activity Detection ». In : *CoRR* abs/1703.07814 (2017).

[108] Zehuan Yuan, Jonathan Stroud, Tong Lu et Jia Deng. « Temporal Action Localization by Structured Maximal Sums ». In : juill. 2017, p. 3215-3223.

[109] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang et Dahua Lin. « Temporal Action Detection with Structured Segment Networks ». In : *ICCV*. 2017.

[110] Amy Bearman, Olga Russakovsky, Vittorio Ferrari et Li Fei-Fei. « What's the point : Semantic segmentation with point supervision ». In : *European conference on computer vision*. Springer. 2016, p. 549-565.

[111] Maxime Bucher, Stephane Herbin et Frederic Jurie. « Improving semantic embedding consistency by metric learning for zero-shot classiffication ». In : *European Conference on Computer Vision*. Springer. 2016, p. 730-746.

[112] Gong Cheng, Peicheng Zhou et Junwei Han. « Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images ». In : *IEEE Transactions on Geoscience and Remote Sensing* 54.12 (2016), p. 7405-7415.

[113] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth et Bernt Schiele. « The cityscapes dataset for semantic urban scene understanding ». In : *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, p. 3213-3223.

[114] Aurelien Ducournau et Ronan Fablet. « Deep learning for ocean remote sensing : an application of convolutional neural networks for super-resolution on satellite-derived SST data ». In : *2016 9th IAPR Workshop on Pattern Recogniton in Remote Sensing (PRRS)*. IEEE. 2016, p. 1-6.

[115] Victor Escorcia, Fabian Heilbron, Juan Carlos Niebles et Bernard Ghanem. « DAPs : Deep Action Proposals for Action Understanding ». In : t. 9907. Oct. 2016, p. 768-784.

[116] Christoph Feichtenhofer, Axel Pinz et Andrew Zisserman. « Convolutional two-stream network fusion for video action recognition ». In : *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, p. 1933-1941.

[117] Yarin Gal et Zoubin Ghahramani. « Dropout as a bayesian approximation : Representing model uncertainty in deep learning ». In : *international conference on machine learning*. 2016, p. 1050-1059.

[118] Hayit Greenspan, Bram Van Ginneken et Ronald M Summers. « Guest editorial deep learning in medical imaging : Overview and future promise of an exciting new technique ». In : *IEEE transactions on medical imaging* 35.5 (2016), p. 1153-1159.

[119] Kaiming He, Xiangyu Zhang, Shaoqing Ren et Jian Sun. « Deep residual learning for image recognition ». In : *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, p. 770-778.

[120] Fabian Heilbron, Juan Carlos Niebles et Bernard Ghanem. *Fast Temporal Activity Proposals for Efficient Detection of Human Actions in Untrimmed Videos*. 2016.

[121] Paul Henderson et Vittorio Ferrari. « End-to-end training of object class detectors for mean average precision ». In : *Asian conference on computer vision*. Springer. 2016, p. 198-213.

[122] Pavlo Molchanov, Stephen Tyree, Tero Karras, Timo Aila et Jan Kautz. « Pruning convolutional neural networks for resource efficient inference ». In : *arXiv preprint arXiv :1611.06440* (2016).

[123] Cathy O'neil. *Weapons of math destruction : How big data increases inequality and threatens democracy*. Broadway Books, 2016.

[124] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik et Ananthram Swami. « The limitations of deep learning in adversarial settings ». In : *Security and Privacy (EuroS&P), 2016 IEEE European Symposium on*. IEEE. 2016, p. 372-387.

[125] Xiaojuan Qi, Zhengzhe Liu, Jianping Shi, Hengshuang Zhao et Jiaya Jia. « Augmented feedback in semantic segmentation under image level supervision ». In : *European conference on computer vision*. Springer. 2016, p. 90-105.

[126] Joseph Redmon, Santosh Divvala, Ross Girshick et Ali Farhadi. « You only look once : Unified, real-time object detection ». In : *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, p. 779-788.

[127] A. Richard et J. Gall. « Temporal Action Detection Using a Statistical Language Model ». In : *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, p. 3131-3140.

[128] Zheng Shou, Dongang Wang et Shih-Fu Chang. « Temporal Action Localization in Untrimmed Videos via Multi-stage CNNs ». In : *CVPR*. 2016.

[129] Gurkirt Singh et Fabio Cuzzolin. « Untrimmed Video Classification for Activity Detection : submission to ActivityNet Challenge ». In : *CoRR* abs/1607.01979 (2016).

[130] Ning Xu, Brian Price, Scott Cohen, Jimei Yang et Thomas Huang. « Deep Interactive Object Selection ». In : *Conference on Computer Vision and Pattern Recognition*. IEEE. 2016, p. 373-381.

[131] J. Yuan, B. Ni, X. Yang et A. A. Kassim. *Temporal Action Localization with Pyramid of Score Distribution Features*. 2016.

[132] Barret Zoph et Quoc V Le. « Neural architecture search with reinforcement learning ». In : *arXiv preprint arXiv :1611.01578* (2016).

[133] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem et Juan Carlos Niebles. « Activitynet : A large-scale video benchmark for human activity understanding ». In : *Proceedings of the ieee conference on computer vision and pattern recognition*. 2015, p. 961-970.

[134] Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen et al. « Xgboost : extreme gradient boosting ». In : *R package version 0.4-2* 1.4 (2015), p. 1-4.

[135] Ross Girshick. « Fast r-cnn ». In : *Proceedings of the IEEE international conference on computer vision*. 2015, p. 1440-1448.

[136] Michael Mathieu, Camille Couprie et Yann LeCun. « Deep multi-scale video prediction beyond mean square error ». In : *arXiv preprint arXiv :1511.05440* (2015).

[137] Jerome Morio et Mathieu Balesdent. *Estimation of rare event probabilities in complex aerospace and other systems : a practical approach*. Woodhead publishing, 2015.

[138] Anh Nguyen, Jason Yosinski et Jeff Clune. « Deep neural networks are easily fooled : High confidence predictions for unrecognizable images ». In : *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, p. 427-436.

[139] Sandeep Paul, Lotika Singh et al. « A review on advances in deep learning ». In : *2015 IEEE Workshop on Computational Intelligence : Theories, Applications and Future Directions (WCI)*. IEEE. 2015, p. 1-6.

[140] Shaoqing Ren, Kaiming He, Ross Girshick et Jian Sun. « Faster r-cnn : Towards real-time object detection with region proposal networks ». In : *Advances in neural information processing systems* 28 (2015).

[141] Olaf Ronneberger, Philipp Fischer et Thomas Brox. « U-net : Convolutional networks for biomedical image segmentation ». In : *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, p. 234-241.

[142] Olga Russakovsky, Li-Jia Li et Li Fei-Fei. « Best of both worlds : human-machine collaboration for object annotation ». In : *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, p. 2121-2131.

[143] Olga Russakovsky et al. « Imagenet large scale visual recognition challenge ». In : *International journal of computer vision* 115.3 (2015), p. 211-252.

[144] Serena Yeung, Olga Russakovsky, Greg Mori et Li Fei-Fei. « End-to-end Learning of Action Detection from Frame Glimpses in Videos ». In : *CoRR* abs/1511.06984 (2015).

[145] Francesco Comaschi, Sander Stuijk, Twan Basten et Henk Corporaal. « A tool for fast ground truth generation for object detection and tracking from video ». In : *2014 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2014, p. 368-372.

[146] Ross Girshick, Jeff Donahue, Trevor Darrell et Jitendra Malik. « Rich feature hierarchies for accurate object detection and semantic segmentation ». In : *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, p. 580-587.

[147] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville et Yoshua Bengio. « Generative adversarial nets ». In : *Advances in neural information processing systems* 27 (2014).

[148] Mingyuan Jiu, Christian Wolf, Graham Taylor et Atilla Baskurt. « Human body part estimation from depth images via spatially-constrained deep learning ». In : *Pattern Recognition Letters* 50 (2014), p. 122-129.

[149] Shrinu Kushagra. « Niceness Assumptions for Learning Algorithms ». Mém. de mast. University of Waterloo, 2014.

[150] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár et C Lawrence Zitnick. « Microsoft coco : Common objects in context ». In : *European conference on computer vision*. Springer. 2014, p. 740-755.

[151] Dan Oneata, Jakob Verbeek et Cordelia Schmid. *The LEAR submission at Thumos 2014*. 2014.

[152] Karen Simonyan et Andrew Zisserman. « Very deep convolutional networks for large-scale image recognition ». In : *arXiv preprint arXiv : 1409.1556* (2014).

[153] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow et Rob Fergus. « Intriguing properties of neural networks ». In : *International Conference on Learning Representations* (2014).

[154] Limin Wang, Yu Qiao et Xiaoou Tang. « Action Recognition and Detection by Combining Motion and Appearance Features ». In : *THUMOS Action Recognition challenge*. 2014, p. 1-6.

[155] Benoit Frenay et Michel Verleysen. « Classification in the presence of label noise : a survey ». In : *IEEE transactions on neural networks and learning systems* 25.5 (2013), p. 845-869.

[156] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra et Martin Riedmiller. « Playing atari with deep reinforcement learning ». In : *arXiv preprint arXiv :1312.5602* (2013).

[157] Ilya Sutskever, James Martens, George Dahl et Geoffrey Hinton. « On the importance of initialization and momentum in deep learning ». In : *International conference on machine learning*. PMLR. 2013, p. 1139-1147.

[158] Alex Tamkin, Iain Usiri et Chala Fufa. *Deep CNNs for diabetic retinopathy detection*. Rapp. tech. Stanford University, Tech. Rep, 2013.

[159] Ruth Urner et Shai Ben-David. « Probabilistic lipschitzness a niceness assumption for deterministic labels ». In : *Learning Faster from Easy Data-Workshop NIPS*. T. 2. 2013, p. 1.

[160] Heng Wang, Alexander Klaser, Cordelia Schmid et Cheng-Lin Liu. « Dense trajectories and motion boundary descriptors for action recognition ». In : *International journal of computer vision* 103.1 (2013), p. 60-79.

[161] Adam Kepecs et Zachary F Mainen. « A computational framework for the study of confidence in humans and animals ». In : *Philosophical Transactions of the Royal Society B : Biological Sciences* 367.1594 (2012), p. 1322-1337.

[162] Alex Krizhevsky, Ilya Sutskever et Geoffrey E Hinton. « Imagenet classification with deep convolutional neural networks ». In : *Advances in neural information processing systems*. 2012, p. 1097-1105.

[163] Franz Rottensteiner, Gunho Sohn, Jaewook Jung, Markus Gerke, Caroline Baillard, Sebastien Benitez et Uwe Breitkopf. « The ISPRS benchmark on urban object classification and 3D building reconstruction ». In : *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*. T. 1, no. 1. Copernicus GmbH, 2012, p. 293-298.

[164] Corina Vaduva, Inge Gavat et Mihai Datcu. « Deep learning in very high resolution remote sensing image information mining communication concept ». In : *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*. IEEE. 2012, p. 2506-2510.

[165] Angela Yao, Juergen Gall, Christian Leistner et Luc Van Gool. « Interactive object detection ». In : *2012 IEEE conference on computer vision and pattern recognition*. IEEE. 2012, p. 3242-3249.

[166] Sangmin Oh et al. « A large-scale benchmark dataset for event recognition in surveillance video ». In : *CVPR 2011*. IEEE. 2011, p. 3153-3160.

[167] Dong Yu, Jinyu Li et Li Deng. « Calibration of confidence measures in speech recognition ». In : *IEEE Transactions on Audio, Speech, and Language Processing* 19.8 (2011), p. 2461-2473.

[168] Bruno Berberian, Patrick Le Blaye, Victorien Marchand et Jean-Christophe Sarrazin. « A preliminary experiment on the concepts of authority sharing and agency in UAS supervisory control ». In : *2nd HUMOUS (HUMans Operating Unmanned Systems), Toulouse, France* (2010).

[169] Yuanqing Lin et al. « Imagenet classification : fast descriptor coding and large-scale svm training ». In : *Large scale visual recognition challenge* (2010).

[170] Johannes Magenheim, Wolfgang Reinhardt, Alexander Roth, Matthias Moi et Dieter Engbring. « Integration of a video annotation tool into a coactive learning and working environment ». In : *IFIP International Conference on Key Competencies in the Knowledge Society*. Springer. 2010, p. 257-268.

[171] Vinod Nair et Geoffrey E Hinton. « Rectified linear units improve restricted boltzmann machines ». In : *Icml*. 2010.

[172] Florent Perronnin, Yan Liu, Jorge Sanchez et Herve Poirier. « Large-scale image retrieval with compressed fisher vectors ». In : *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE. 2010, p. 3384-3391.

[173] Hong Cao et Alex C Kot. « Accurate detection of demosaicing regularity for digital image forensics ». In : *IEEE Transactions on Information Forensics and Security* 4.4 (2009), p. 899-910.

[174] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li et Li Fei-Fei. « ImageNet : A large-scale hierarchical image database ». In : *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009, p. 248-255.

[175] Derek Hoiem, Santosh K Divvala et James H Hays. « Pascal VOC 2008 challenge ». In : *World Literature Today* 24 (2009).

[176] Kevin Jarrett, Koray Kavukcuoglu, Marc'Aurelio Ranzato et Yann LeCun. « What is the best multi-stage architecture for object recognition ? » In : *2009 IEEE 12th international conference on computer vision*. IEEE. 2009, p. 2146-2153.

[177] Alex Krizhevsky, Vinod Nair et Geoffrey Hinton. « Cifar-10 and cifar-100 datasets ». In : *techical report* 6.1 (2009), p. 1.

[178] Jingen Liu, Jiebo Luo et Mubarak Shah. « Recognizing realistic actions from videos "in the wild" ». In : *2009 IEEE conference on computer vision and pattern recognition*. IEEE. 2009, p. 1996-2003.

[179] Marcin Marszalek, Ivan Laptev et Cordelia Schmid. « Actions in Context ». In : *IEEE Conference on Computer Vision & Pattern Recognition*. 2009.

[180] Burr Settles. *Active learning literature survey*. 2009.

[181] Pedro Felzenszwalb, David McAllester et Deva Ramanan. « A discriminatively trained, multiscale, deformable part model ». In : *2008 IEEE conference on computer vision and pattern recognition*. Ieee. 2008, p. 1-8.

[182] Bryan C Russell, Antonio Torralba, Kevin P Murphy et William T Freeman. « LabelMe : a database and web-based tool for image annotation ». In : *International journal of computer vision* 77.1 (2008), p. 157-173.

[183] Pascal Vincent, Hugo Larochelle, Yoshua Bengio et Pierre-Antoine Manzagol. « Extracting and composing robust features with denoising autoencoders ». In : *Proceedings of the 25th international conference on Machine learning*. 2008, p. 1096-1103.

[184] Amiran Ambroladze, Emilio Parrado-Hernandez et John Shawe-Taylor. « Tighter pac-bayes bounds ». In : *Advances in neural information processing systems* 19 (2006).

[185] Michael Donoser et Horst Bischof. « Efficient maximally stable extremal region (MSER) tracking ». In : *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*. T. 1. Ieee. 2006, p. 553-560.

[186] Qiang Zhu, Mei-Chen Yeh, Kwang-Ting Cheng et Shai Avidan. « Fast human detection using a cascade of histograms of oriented gradients ». In : *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*. T. 2. IEEE. 2006, p. 1491-1498.

[187] John Langford et Robert Schapire. « Tutorial on practical prediction theory for classification. » In : *Journal of machine learning research* 6.3 (2005).

[188] Nilesh Dalvi, Pedro Domingos, Sumit Sanghai et Deepak Verma. « Adversarial classification ». In : *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2004, p. 99-108.

[189] Carsten Rother, Vladimir Kolmogorov et Andrew Blake. « " GrabCut" interactive foreground extraction using iterated graph cuts ». In : *ACM transactions on graphics (TOG)* 23.3 (2004), p. 309-314.

[190] Zhiqiang Zheng et Balaji Padmanabhan. « On active learning for data acquisition ». In : *2002 IEEE International Conference on Data Mining, 2002. Proceedings*. IEEE. 2002, p. 562-569.

[191] Larry R Medsker et LC Jain. « Recurrent neural networks ». In : *Design and Applications* 5 (2001), p. 64-67.

[192] Alexey L Pomerantsev. « Confidence intervals for nonlinear regression extrapolation ». In : *Chemometrics and Intelligent Laboratory Systems* 49.1 (1999), p. 41-48.

[193] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999.

[194] Francois Denis et Remi Gilleron. « PAC learning under helpful distributions ». In : *International Workshop on Algorithmic Learning Theory*. Springer. 1997, p. 132-145.

[195] Mance E Harmon et Stephanie S Harmon. *Reinforcement Learning : A Tutorial*. 1997.

[196] Oded Maron et Tomas Lozano-Perez. « A framework for multiple-instance learning ». In : *Advances in neural information processing systems* 10 (1997).

[197] David H Wolpert. « The lack of a priori distinctions between learning algorithms ». In : *Neural computation* 8.7 (1996), p. 1341-1390.

[198] Corinna Cortes et Vladimir Vapnik. « Support-vector networks ». In : *Machine learning* 20.3 (1995), p. 273-297.

[199] David D Lewis et Jason Catlett. « Heterogeneous uncertainty sampling for supervised learning ». In : *Machine learning proceedings 1994*. Elsevier, 1994, p. 148-156.

[200] Dale Griffin et Amos Tversky. « The weighing of evidence and the determinants of confidence ». In : *Cognitive psychology* 24.3 (1992), p. 411-435.

[201] Yann LeCun, Bernhard Boser, John Denker, Donnie Henderson, Richard Howard, Wayne Hubbard et Lawrence Jackel. « Handwritten digit recognition with a back-propagation network ». In : *Advances in neural information processing systems* 2 (1989).

[202] Shimon D Yanowitz et Alfred M Bruckstein. « A new method for image segmentation ». In : *Computer Vision, Graphics, and Image Processing* 46.1 (1989), p. 82-95.

[203] Robert M Haralick et Linda G Shapiro. « Image segmentation techniques ». In : *Computer vision, graphics, and image processing* 29.1 (1985), p. 100-132.

[204] Guy Barrett Coleman et Harry C Andrews. « Image segmentation by clustering ». In : *Proceedings of the IEEE* 67.5 (1979), p. 773-785.

[205] John A Hartigan et Manchek A Wong. « Algorithm AS 136 : A k-means clustering algorithm ». In : *Journal of the royal statistical society. series c (applied statistics)* 28.1 (1979), p. 100-108.

5.2 Présentation du candidat à l'habilitation à diriger des recherches

5.2.1 Parcours

- 2007 - 2011 : École Polytechnique
- 2011 - 2014 : thèse
  - dirigée par Catherine Achard
  - ED SMAER, Sorbonne université et CEA LIST
  - sujet : Segmentation supervisée d'actions à partir de primitives haut niveau dans des flux vidéos
- 2014 - aujourd'hui : Chercheur à l'ONERA - charge contractuelle ≈ 50%

5.2.2 Publications

- Articles de journaux internationaux avec comité de lecture : 9 [1]
- Articles de conférences internationales avec comité de lecture : 24 [2]
- Article publiées dans des journaux nationaux ou conférences nationales : 8

Journaux internationaux avec comité de lecture

[206] Adrien Chan-Hon-Tong. « A New Algorithm for Linear Programming in Critical Systems ». In : *SN Computer Science* 4.1 (2023), p. 1-10.

[207] Magdeleine Airiau, Adrien Chan-Hon-Tong, Robin W Devillers et Guy Le Besnerais. « Regressing Image Sub-Population Distributions with Deep Learning ». In : *Sensors* 22.23 (2022), p. 9218.

[208] Aurélie Bouchard, Magalie Buguet, Adrien Chan-Hon-Tong, Jean Dezert et Philippe Lalande. « Comparison of different forecasting tools for short-range lightning strike risk assessment ». In : *Natural Hazards* (2022), p. 1-37.

[209] Gaston Lenczner, Adrien Chan-Hon-Tong, Bertrand Le Saux, Nicola Luminari et Guy Le Besnerais. « DIAL : Deep Interactive and Active Learning for Semantic Segmentation in Remote Sensing ». In : *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* (2022).

[210] Andrea Desantis, Adrien Chan-Hon-Tong, Thérèse Collins, Hinze Hogendoorn et Patrick Cavanagh. « Decoding the temporal dynamics of covert spatial attention using multivariate EEG analysis : contributions of raw amplitude and alpha power ». In : *Frontiers in human neuroscience* 14 (2020), p. 430.

[211] Adrien Chan-Hon-Tong. « An algorithm for generating invisible data poisoning using adversarial noise that breaks image classification deep learning ». In : *Machine Learning and Knowledge Extraction* 1.1 (2019), p. 192-204.

[212] Juliette Chataigner, Stephane Herbin et Adrien Chan-Hon-Tong. « Pertinence of Video for Single Image Deep Network ». In : *International Journal of Machine Learning and Computing* 7 (2017), p. 238-242.

1. dont 1 durant la thèse et dont 3 en premier auteur
2. dont 3 durant la thèse et dont 7 en premier auteur

[213] Manuel Campos-Taberner et al. « Processing of extremely high-resolution Lidar and RGB data : outcome of the 2015 IEEE GRSS data fusion contest–part a : 2-D contest ». In : *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 9.12 (2016), p. 5547-5559.

[214] Adrien Chan-Hon-Tong, Catherine Achard et Laurent Lucat. « Simultaneous segmentation and classification of human actions in video streams using deeply optimized Hough transform ». In : *Pattern Recognition* 47.12 (2014), p. 3807-3818.

Conférences internationales avec comité de lecture

[215] Safa Bousbih, Adrien Chan-Hon-Tong et Gaston Lenczner. « What could we learn from many datasets in remote sensing roof semantic segmentation ? » In : *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*. IEEE. 2022, p. 999-1002.

[216] Adrien Chan-Hon-Tong. « Solving Linear Programming While Tackling Number Representation Issues ». In : *Proceedings of the 11th International Conference on Operations Research and Enterprise Systems - ICORES,* INSTICC. SciTePress, 2022, p. 40-47.

[217] Pol Labarbarie, Adrien Chan-Hon-Tong, Stéphane Herbin et Milad Leyli-Abadi. « Benchmarking and deeper analysis of adversarial patch attack on object detectors ». In : *Workshop Artificial Intelligence Safety-AI Safety (IJCAI-ECAI conference)*. 2022.

[218] Gaston Lenczner, Adrien Chan-Hon-Tong, Nicola Luminari et Bertrand Le Saux. « Weakly-supervised continual learning for class-incremental segmentation ». In : *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*. IEEE. 2022, p. 4843-4846.

[219] Adrien Chan-Hon-Tong, Gaston Lenczner et Plyer Aurélien. « Demotivate adversarial defense in remote sensing ». In : *IGARSS 2021-2021 IEEE International Geoscience and Remote Sensing Symposium*. IEEE. 2021.

[220] Guillaume Vaudaux-Ruth, Adrien Chan Hon Tong et Catherine Achard. « ActionSpotter : Deep Reinforcement Learning Framework for Temporal Action Spotting in Videos ». In : *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE. 2021, p. 631-638.

[221] Guillaume Vaudaux-Ruth, Adrien Chan Hon Tong et Catherine Achard. « SALAD : Self-Assessment Learning for Action Detection ». In : *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2021, p. 1269-1278.

[222] Thomas Chaffre, Julien Moras, Adrien Chan-Hon-Tong, Julien Marzat, Karl Sammut, Gilles Le Chenadec et Benoit Clement. « Learning-based vs Model-free Adaptive Control of a MAV under Wind Gust ». In : *International Conference on Informatics in Control, Automation and Robotics*. Springer. 2020, p. 362-385.

[223] Thomas Chaffre., Julien Moras., Adrien Chan-Hon-Tong. et Julien Marzat. « Sim-to-Real Transfer with Incremental Environment Complexity for Reinforcement Learning of Depth-based Robot Navigation ». In : *Proceedings of the 17th International Conference on Informatics in Control, Automation and Robotics - ICINCO,* INSTICC. SciTePress, 2020, p. 314-323.

[224] Adrien CHAN-HON-TONG. « Symmetric adversarial poisoning against deep learning ». In : *2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA)*. 2020, p. 1-5.

[225] G. Lenczner, B. Le Saux, N. Luminari, A. Chan-Hon-Tong et G. Le Besnerais. « DISIR : DEEP IMAGE SEGMENTATION WITH INTERACTIVE REFINEMENT ». In : *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* V-2-2020 (2020), p. 877-884.

[226] Gaston Lenczner, Adrien Chan-Hon-Tong, Nicola Luminari, Bertrand Le Saux et Guy Le Besnerais. « Interactive Learning for Semantic Segmentation in Earth Observation ». In : *Maclean Workshop*. 2020.

[227] Jordan Platon, Guillaume Avrin et Adrien Chan-Hon-Tong. « Generating corner cases for crashtesting deep networks ». In : *ECAI*. 2020.

[228] Rodrigo Caye Daudt, Adrien Chan-Hon-Tong, Bertrand Le Saux et Alexandre Boulch. « Learning to understand earth observation images with weak and unreliable ground truth ». In : *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*. IEEE. 2019, p. 5602-5605.

[229] Robin Devillers, Matthieu Nugue, Adrien Chan Hon Tong, Guy Le Besnerais et Julien Pichillou. « Experimental analysis of aluminum-droplet combustion in solid-propellant conditions using deep learning ». In : *EUCASS 2019*. MADRID, Spain, juill. 2019.

[230] Adrien Chan-Hon-Tong et Nicolas Audebert. « Object detection in remote sensing images with center only ». In : *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*. IEEE. 2018, p. 7054-7057.

[231] Adrien Chan-Hon-Tong. et Stephane Herbin. « Practical Scheduling of Computer Vision Functions ». In : *Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 6 : VISAPP, (VISIGRAPP 2017)*. INSTICC. SciTePress, 2017, p. 347-352.

[232] Zehira Haddad, Adrien Chan Hon Tong et Jaime Lopez Krahe. « Image Resolution Enhancement based on Curvelet Transform ». In : *International Conference on Computer Vision Theory and Applications*. T. 5. SCITEPRESS. 2017, p. 167-173.

[233] A. Chan-Hon-Tong R. Tripathi et A. Boulch. « Off the shelf deep learning pipeline for remote sensing applications ». In : *Proceedings of the 2017 conference on Big Data from Space*. 2017.

[234] Adrien Chan-Hon-Tong et Stephane Herbin. « Tracking based sparse box proposal for time constraint detection in video stream ». In : *2015 International Conference on Image Processing Theory, Tools and Applications (IPTA)*. IEEE. 2015, p. 81-86.

[235] Adrien Lagrange, Bertrand Le Saux, Anne Beaupere, Alexandre Boulch, Adrien Chan-Hon-Tong, Stéphane Herbin, Hicham Randrianarivo et Marin Ferecatu. « Benchmarking classification of earth-observation data : From learning explicit features to convolutional networks ». In : *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE. 2015, p. 4173-4176.

[236] Adrien Chan-Hon-Tong, Catherine Achard et Laurent Lucat. « Deeply optimized hough transform : Application to action segmentation ». In : *International Conference on Image Analysis and Processing*. Springer. 2013, p. 51-60.

[237] Adrien Chan-Hon-Tong, Nicolas Ballas, Catherine Achard, Bertrand Delezoide, Laurent Lucat, Patrick Sayd et Françoise J Prêteux. « Skeleton Point Trajectories for Human Daily Activity Recognition. » In : *VISAPP*. 2013, p. 520-529.

[238] Thierry Chesnais, Thierry Chateau, Nicolas Allezard, Yoann Dhome, Boris Meden, Mohamed Tamaazousti et Adrien Chan-Hon-Tong. « A Region Driven and Contextualized Pedestrian Detector. » In : *VISAPP*. 2013, p. 796-799.

Autres articles

[239] F. Boniol, A. Chan-Hon-Tong, A. Eudes, S. Herbin, G. Le Besnerais, C. Pagetti et M. Sanfourche. « Défis pour la certification de systèmes basés vision par ordinateur pour l'aéronautique civile ». In : *Aerospace Lab journal* 15 (2020).

[240] E. Colin Koeniguer et al. « Exemples récents de contributions de l'apprentissage profond à des problèmes d'observation de la Terre ». In : *Aerospace Lab journal* 15 (2020).

[241] Gaston Lenczner, Bertrand Le Saux, Nicolas Luminari, Adrien Chan-Hon-Tong et Guy Le Besnerais. « Segmentation sémantique d'images aériennes avec améliorations interactives ». In : *RFIA 2020*. 2020.

[242] M. Nugue, J.-M. Roche, G. Le Besnerais, C. Trottier, R. W. Devillers, J. Pichillou, A. Chan-Hon-Tong, A. Boulch et A. Hurmane). « Accroître l'extraction d'informations à partir de données scientifiques grâce à l'apprentissage profond ». In : *Aerospace Lab journal* 15 (2020).

[243] Adrien Chan-Hon-Tong. « Exemples passeurs en apprentissage profond ». In : *RFIA 2018*. 2018.

[244] Matthieu Nugue, Robin W Devillers, Adrien Chan Hon Tong, Guy Le Besnerais et Julien Pichillou. « Classification automatique d'images de propergol solide en combustion par utilisation de réseaux de neurones convolutifs ». In : *GRETSI 2017*. 2017.

[245] Adrien Chan-Hon-Tong, Stephane Herbin et Alexandre Boulch. « Un générateur de boites englobantes parcimonieux pour la détection d'objets dans des vidéos ». In : *XXV colloque Gretsi 2015*. 2015.

[246] Adrien Chan-Hon-Tong, Catherine Achard et Laurent Lucat. « Transformée de Hough sans a priori pour la segmentation ». In : *XXIV Colloque Gretsi*. 2013.

5.2.3 Brevets et diffusion logicielle

— https ://data.inpi.fr/brevets/WO2019186073 *METHODE DE DETECTION DE CELLULES PRESENTANT AU MOINS UNE ANOMALIE DANS UN ECHANTILLLON CYTOLOGIQUE*, HADDAD ZEHIRA, HERBIN STÉPHANE, CHAN-HON-TONG ADRIEN

— https ://data.inpi.fr/brevets/EP3198523 *PROCEDE ET SYSTEME DE DETECTION D'EVENEMENTS DE NATURE CONNUE*, CHAN-HON-TONG ADRIEN, LUCAT LAURENT

Indépendamment des brevets, je cherche quand cela est possible à mettre les codes en accès libre à la fois dans un objectif de transparence et de reproductibilité mais aussi dans un objectif de diffusion logicielle. Les codes issues de la thèse de Gaston Lenczner sont sur github (à noter qu'ils sont utilisés dans au moins 3 entreprises). Les codes des travaux sur l'empoisonnement sont sur github (achanhon/AdversarialModel). Je maintiens un dépôt générique avec des codes orientés télédétection sont sur github (delta-onera/delta_tb). Les codes associés aux travaux propergol sont considérés comme sensibles. Ceux de la thèse de Guillaume Vaudaux-Ruth ne sont pas en accès libre car ils pourraient être valorisés avec la startup deeptimize.

5.2.4 Encadrements

5.2.4.1 Encadrement de stage

— Juliette Chataigner (M1 2017) [212]
— Thomas Chaffre (M2 2020) [222, 223]
— Jordan Platon (M1 2020) [227]

5.2.4.2 Co encadrement de thèse

— Matthieu Nugue 2017-2019
  — directeur de thèse : Guy Le Besnerais
  — financement CNES
  — sujet : Outils pour l'étude conjointe par simulation et traitement d'images expérimentales de la combustion de particules d'aluminium utilisées dans les propergols solides
— Gaston Lenczner 2020-2022
  — directeur de thèse : Guy Le Besnerais
  — financement Alteia (entreprise)
  — sujet : Interactive semantic segmentation of aerial images with deep neural networks
— Guillaume Vaudaux-Ruth 2019-2021
  — directrice de thèse : Catherine Achard
  — financement DGA-ONERA
  — sujet : Du repérage sémantique robuste d'actions vers leur détection dans les vidéos
— Magdeleine Airiau (deuxième année)
  — directeur de thèse : Guy Le Besnerais
  — financement CNES
  — sujet : couplage mesure par IA et simulation
— Pol Labarbarie (première année)
  — directeur de thèse : Stéphane Herbin
  — financement IRT confiance.ai
  — sujet : résistance aux exemples adversaires par patch

5.2.5 Enseignement

— Monitorat à l'université Paris Saclay de 2012 à 2014 en informatique
— Cours de M1 d'introduction au machine learning (partages des scéances 50%/50% avec Stéphane Herbin) à l'ENSTA de 2016 à aujourd'hui
— Tutoriel machine learning à l'EUROSAE de 2018 à aujourd'hui dans le cadre d'une formation de Laurent Chaudron

5.2.6 Rayonnement

— Area Chair à IGARSS 2021 (et potentiellement 2022)
— Area Chair à là conférence CAID (IA pour la défense) depuis 2018
— Relecture de 12 articles pour MDPI
— Relecture de 1 papier à ICCV (sous relecteur de Stéphane Herbin)
— Relecture de 1 papier à CVPR (sous relecteur de Stéphane Herbin)
— Relecture de 1 projet ANR

5.3 Résumé de mon projet de recherche à 5 ans : 4 thèses à venir

Pour donner une vision globale de mon projet de recherche à 5 ans, cette section auto-suffisante par rapport au reste du manuscrit (mais donc redondante) propose 4 sujets de thèse déjà introduits dans les chapitres 2, 3 et 4 et regroupés ici. On peut noter que l'IA de confiance et la télédétection y sont des fils directeurs.

5.3.1 Estimation du risque orageux par réseau de neurones

Récemment, l'entreprise *Deepmind* s'est penché sur l'utilisation de méthodes d'apprentissage profond pour l'estimation météo [19]. Cependant, ce sujet n'est pas nouveau, que ce soit à Météo France [25] ou à l'ONERA [92]. Plus précisément, l'ONERA ne s'intéresse pas à la météo mais à des états de l'atmosphère qui peuvent avoir des conséquences sur nos capacités de défense comme l'état de l'ionosphère (liaisons satellites) [92] ou le risque orageux (aviation en général). Notamment, un travail préexistant [208] a montré le potentiel de l'utilisation de ces méthodes pour estimer un risque orageux long terme.

Partant de ce contexte, l'objet de la thèse est d'essayer de gagner en précision temporelle en allant vers la prédiction court terme.

Pour cela, les modèles pourront s'appuyer sur des données satellites permettant de disposer de la quasi-totalité des éclairs passés : en effet les éclairs étant des évènements saillants, et les satellites GLM [3] sont capables de les détecter. Malheureusement, les éclairs sont des phénomènes intrinsèquement stochastiques et la connaissance des éclairs ayant eu lieu ne permet pas directement de déduire les éclairs futurs.

3. https ://ghrc.nsstc.nasa.gov/lightning/overview_glm.html

Il conviendra donc de trouver d'autres sources de données permettant de construire une représentation de l'atmosphère propre à la prédiction des orages. Cela pourra se faire en s'appuyant sur l'expertise de l'équipe *Foudre, Plasma et Application* de l'ONERA. Il conviendra aussi faire le lien avec la littérature *d'IA de confiance* notamment sur la production de probabilité d'incidences plutôt que sur une décision binaire *éclair/pas éclair* compte tenu du caractère stochastique des orages. Cela pourra se faire en s'appuyant sur l'expertise de l'équipe *Image vision apprentissage* de l'ONERA sur cette problématique de calibration [221].

5.3.2 Auto-supervision pour l'annotation en un clic

La segmentation sémantique est une tâche de vision par ordinateur offrant de nombreuses applications possibles sur une variété de données différentes [235, 239]. Cependant, la construction de bases de données de segmentation sémantique impose d'annoter au pixel près les images d'entraînement ce qui est une limitation pratique importante. Partant de ce constat, de nombreux travaux portent sur la mécanisation de l'annotation [142], travaux dans lesquels Alteia et l'ONERA ont contribué notamment via le développement de méthodes *d'annotation en un clic* [225, 226, 241] visant à raffiner une carte de pré-segmentation à l'aide de clics de l'utilisateur.

Si ces dernières approches sont très performantes, elles sont aujourd'hui limitées aux applications pour lesquelles des bases de données *proches* pré-existent. En effet, ces bases de données sont utilisées pour construire un modèle capable de pré-segmenter et de raffiner sa propre segmentation en utilisant des clics utilisateurs. Il est ainsi possible de traiter toutes sortes de problèmes de télédétection à partir de bases académiques [218].

Mais, il n'est aujourd'hui pas possible de traiter directement les données très distantes (comme celles de [239]) à partir du paradigme introduit dans [218].

Aussi, l'objectif de cette thèse est de supprimer le besoin de cette pré-segmentation sémantique (et donc d'une base de données pré-existante) à l'aide de méthode d'auto-supervision (comme le *contrastive learning* [30, 31]) qui offre la possibilité de créer des espaces latents propres à la sémantique (et donc maintenir des performances équivalentes à celles de [225]).

L'idée sous-jacente est de réussir à construire une représentation latente à la fois du clic mais aussi de la donnée cible pour disposer d'un modèle aligné sur les données avant même le premier clic de l'utilisateur. En ce sens, on gagnera potentiellement à s'inspirer des méthodes couplant texte et image grâce à un produit de représentations latentes [73] mais aussi de méthodes historiques de segmentation comme GrabCut [189].

5.3.3 Explicabilité appliquée à des données régies par des lois physiques

L'explicabilité est une thématique qui est désormais très populaire dans la communauté de vision par ordinateur [73]. Cependant, malgré un certain

nombres de succès, l'explicabilité reste difficile sur des objets de très grande dimension comme les images.

Inversement, au sein de l'ONERA, l'apprentissage profond est utilisé dans de nombreux domaines régis par des lois physiques par exemple la mécanique des fluides. Or, ces domaines sont probablement intéressants pour l'explicabilité car une bonne explication passe alors nécessairement par la reconstruction des lois sous-jacentes. Ainsi, d'une part, disposer de modèles explicables intéresse fortement les chercheurs de mécanique des fluides - et inversement - un problème de mécanique des fluides est un contexte favorable pour avancer sur cette idée d'explicabilité notamment en évaluant la pertinence physique des explications.

Précisément, l'objectif de cette thèse est de prolonger les travaux ONERA [16]. Dans ces travaux, un algorithme de renforcement a été appris pour faire du contrôle d'écoulements. Les résultats obtenus sont significativement supérieurs aux résultats classiques de contrôle d'écoulement obtenus avec des modèles typiquement linéaires. Cependant, et ce sera le coeur de cette thèse, ces résultats sont à consolider, notamment par la compréhension des mécanismes mis en oeuvre dans la politique de renforcement.

5.3.4 Utilisation du contexte de la télédétection pour estimer l'effet des biais de tirages lors de l'évaluation de réseaux de neurones

La question de la réglementation des algorithmes d'apprentissage par ordinateur est un sujet critique pour aller vers des applications industrielles majeures (comme les voitures autonomes) comme en témoigne les débats législatifs au niveau de l'Union Européenne [4] et/ou la publication de nombreux *white paper* sur le sujet ([11] pour DEEL, [32] pour l'Europe, ou [15] pour le grand défi IA de l'agence innovation défense).

Cette réglementation imposera notamment une évaluation empirique. Mais, ce type d'évaluation repose sur l'idée centrale d'un tirage identiquement distribué sur une distribution fixe. Cependant, les campagnes de mesures qui accompagnent la validation d'un algorithme conduiront probablement à des données non identiquement distribuées car elles mettent en jeu des moyens de mesure dont la disponibilité est faible.

Aussi, plusieurs instances normatives se posent la question de la possibilité d'autres modalités de notation. Le LNE [5] envisage par exemple une évaluation en deux étapes calquée sur une évaluation écrite puis orale : le pendant de *l'écrit* serait une évaluation empirique rigide basée sur une campagne de mesure (forcément biaisée car limitée), mais, elle serait complétée par une partie dynamique (représentée comme un *oral*) où le système de mesure viendrait chercher à piéger le modèle en générant dynamiquement des exemples difficiles (pas adversaires au sens où ce ne serait juste des perturbations de données déjà connues - mais

4. eur-lex.europa.eu/legal-content/EN/TXT/ ?uri=CELEX%3A52021PC0206
5. www.lne.fr/fr/communiques-de-presse/intelligence-artificielle-nouvelle-certification

adversaire dans le sens où le but est d'aller chercher des données mal traitées). Mais, rien ne permet aujourd'hui de valider l'efficacité de ce paradigme.

Cependant, la télédétection parait être le contexte le plus adapté pour calibrer ce type de procédure car des données y sont disponibles et bien réparties en espace et en temps. Ainsi, l'objectif de cette thèse en collaboration avec le LNE serait de profiter de la disponibilité de données de télédétection permettant de réaliser des évaluations empiriques quasiment identiquement distribuées, et ainsi, calibrer des méta-algorithmes qui chercheraient à estimer cette performance au moyens de données non identiquement distribuées.

Notamment, les zones où l'incertitude de l'algorithme est forte apparaissent intéressantes dans ce contexte. Cette idée qui a donné lieu à des travaux préliminaires [227, 209] pourra être approfondie. On pourra aussi envisager des approches plus théoriques ou intermédiaires comme celles de la compétition NeurIPS2020 sur la prédiction de la performance en généralisation [36].

5.4 Digressions

5.4.1 Un mot sur les temps de calcul

La cohérence de mes travaux est d'essayer d'avoir une expertise sur l'ensemble des problèmes limitant l'acceptabilité des méthodes d'apprentissage par ordinateur, et plus précisément, d'apprentissage profond. Or, un frein majeur à cette acceptabilité vient des temps de calcul associés. Ces temps de calcul conduisent d'ailleurs à la nécessité d'un *hardware* dédié i.e. de GPU *NVIDIA* [6]. En effet, dès 2012, il était 100 fois plus rapide de faire de l'apprentissage profond sur GPU que sur CPU. Ce frein est peut-être même plus important que celui des problématiques d'IA de confiance, et en tout cas bien plus que celui des coûts d'annotations. Cependant, il n'était pas pertinent d'en faire un chapitre car je n'ai pas de contribution méthodologique suffisamment significative sur la question (bien que ce sujet m'ait fortement occupé en 2015-2016 avec un projet *défense* dont un des objectifs était d'implémenter un détecteur vidéo temps réel, ce qui a motivé [234, 231, 216, 206]).

Par ailleurs, cette problématique est à la fois un domaine de recherche académique (méthodes de compression de réseaux [122, 84, 99], réseaux en nombres entiers [79, 4], ...), mais elle est aussi devenu un problème industriel notamment chez *NVIDIA*. D'ailleurs, ce frein (même s'il subsiste) a très fortement été diminué par l'augmentation spectaculaire de la puissance de calcul GPU disponible résumé par la figure 5.1.

Néanmoins, cette contrainte subsiste et pousse parfois à faire des choix contradictoires avec des objectifs *IA de confiance*. Par exemple, si on regarde les étapes historiques de la diffusion de l'apprentissage profond vers la détection d'objets, on a d'abord R-CNN [146] qui propose des zones via une méthode ad-hoc puis les filtres avec Alexnet. Mais, R-CNN était quand même trop lent par rapport aux méthodes antérieures. Ainsi, l'idée de proposer des zones à l'aide

6. Ce n'est pas une publicité pour ce fabriquant, c'est un constat.

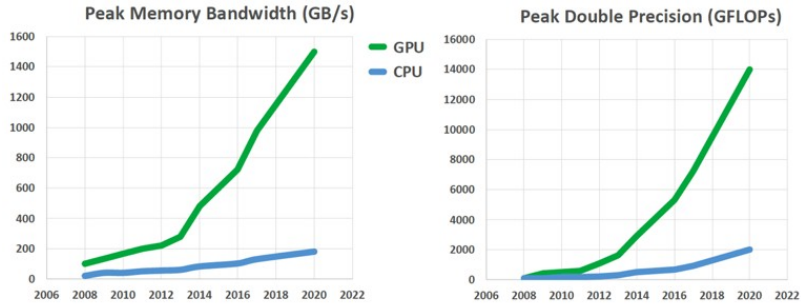


image extraite de
www.nextplatform.com/2019/07/10/a-decade-of-accelerated-computing-augurs-well-for-gpus

FIGURE 5.1 – Illustration de l'augmentation de la puissance GPU disponible sur une seule carte graphique.

de méthodes ad-hoc a été remplacé par un encodage global de l'image. Cela correspond à paralléliser les calculs entre les différentes couches plutôt qu'entre les différentes zones. De cette façon, Yolo [126] ou Fast-RCNN [135, 140] retrouvent des temps de calcul relativement faibles (notamment compatibles avec un traitement en flux de vidéo). Mais, les deux approches (R-CNN vs Fast-RCNN ou yolo) ne sont pas équivalentes. Par exemple, en présence d'une attaque par patch, l'encodage global de l'image est modifiée dans Yolo V4 entraînant une perte de performance si significative [66] qu'elle passe en dessous de celle de R-CNN structurellement insensible à l'attaque puisque les décisions concernant une zone n'utilise que cette zone et pas un encodage global. On voit ainsi que des caractéristiques pertinentes (être invariant au contexte) ont été négligées pour des raisons de temps de calcul [7].

D'ailleurs concernant R-CNN, rétrospectivement, la victoire d'Alexnet n'a pas immédiatement bousculé d'autres sous-domaines de la vision par ordinateur comme par exemple la détection d'objets. Cette latence est potentiellement juste mécanique car le domaine étant rythmé par la durée des thèses, mais on peut aussi penser que cette latence est due à des *a priori* négatifs sur cette technologie. Typiquement, une pensée assez commune en 2014 était de croire que l'apprentissage profond resterait inacceptable pour des données et/ou problèmes plus complexes que la classification d'images (car trop lourd en terme de calcul). Ainsi, en classification de vidéos, c'est une méthode à base de trajectoires denses de points d'intérêt encodées dans un sac de mots [160] qui reste en tête sur Hollywood2 [179] en 2014. De même en détection, c'est une méthode basée sur des templates dans l'espace des histogrammes de gradient (DPM [181]) qui reste en tête sur Pascal VOC [175] avant R-CNN. Mais [146] désamorcera cette idée (ou plutôt la compensera en démontrant 30% de performance supplémentaire).

7. Cette observation est issue des travaux menés dans la thèse de Pol Labarbarie.

Ce faisant, R-CNN a probablement autant participé à *la révolution de l'apprentissage profond* qu'Alexnet car il aura rendu plus acceptable ces méthodes en détection d'images.

5.4.2 Performances et applications critiques

La question de mesurer correctement les performances d'un modèle d'apprentissage par ordinateur (dans la perspective d'une application critique) est indépendante du niveau de cette performance. Maintenant, il faudra de fortes performances [8] pour des applications critiques : par exemple, même si on arrive à motiver la DGAC à autoriser des algorithmes critiques dans un avion, cela ne se fera pas en augmentant la tolérance de 10^{-9} crash par heure de vol, toutes causes confondues. Ce qui conduira à une exigence de performances très hautes sur tous les sous-systèmes critiques. On peut alors se demander si ces performances sont réellement atteignables aujourd'hui.

Cependant, cette tolérance de 10^{-9} crash par heure de vol ne prend pas en compte les erreurs humaines. Or, il se trouve que les accidents de la route ou de diagnostic sont majoritairement causés par le manque de concentration de l'opérateur et les incompréhensions homme-machine. Or, les algorithmes ne sont pas sujets à ce type d'erreurs. Ainsi, mettre en production à court terme des diagnostics médicaux ou des voitures autonomes pourrait aller dans le sens d'une réduction du nombre d'accidents. Non pas parce que ces algorithmes sont réellement meilleurs que le médecin ou le conducteur, mais parce qu'ils sont constants. Dit autrement, une voiture autonome respecte les limites de vitesse et n'est jamais ni fatiguée, ni alcoolisée, quand bien même elle serait moins fiable qu'un conducteur concentré. À noter que l'avion sans pilote n'est pas juste de la science-fiction : l'ONERA a travaillé sur un démonstrateur dans le projet ATTOL [9]. Un autre exemple tiré des travaux avec la startup VitaDx est qu'un anatomopathologiste ne peut consacrer en moyenne que 2 minutes à l'analyse d'une cytologie urinaire comportant en moyenne 50000 cellules urothéliales : dans ces conditions, l'exhaustivité apportée par l'analyse automatisée devient in fine plus performante que le praticien en routine quand bien même le praticien concentré serait plus fiable.

À noter qu'on pourrait apporter une autre réponse (douteuse) consistant à dire que puisque l'objectif est de remplacer un humain, il suffit d'être simplement meilleur que l'humain et que ce niveau est atteint. Cependant, ce dernier point est à débattre. Il y a eu beaucoup de bruit autour d'un post de Karpathy [10] [143] qui affirme en substance que la performance humaine sur Imagenet n'est que de 90% et que dès 2015, la performance des algorithmes de classification était passée au dessus de celle de l'humain. Cependant, il faut avoir en tête que la plupart des problèmes d'apprentissage réels sont ambiguës c'est-à-dire qu'il

8. Le médical est un peu à part car on a souvent des *gold standard* moyennement efficaces
9. www.airbus.com/en/newsroom/press-releases/2020-06-airbus-concludes-attol-with-fully-autonomous-flight-tests
10. karpathy.github.io/2014/09/02/what-i-learned-from-competing-against-a-convnet-on-imagenet

FIGURE 5.2 – Exemples d'ambiguïtés d'annotation (ou de signal) dans la base de données Imagenet trouvées par [26]

n'existe pas de fonction qui va de l'image vers le label et que même avec un humain concentré il n'est pas sûr que sa réponse soit la même que celle d'un autre humain concentré. Cela est notamment évident si l'humain doit donner 1 objet et qu'il y en a plusieurs dans l'image ce qui est le cas sur Imagenet comme l'illustre la figure 5.2. Ainsi, la réalité n'est pas que l'algorithme est meilleur que l'humain mais qu'on est dans le bruit de mesure où l'algorithme a juste mieux réussi à capturer les biais d'annotation. D'ailleurs, si on considère un problème non ambigu (il existe une fonction y de l'image vers la classe), structurellement, l'algorithme peut au mieux atteindre la performance de l'humain concentré puisqu'on suppose que tout humain concentré aura la même réponse. Donc, les algorithmes ne pourront pas (dans le cas d'une annotation manuelle) dépasser l'humain concentré. D'ailleurs, si on prend l'exemple de la détection (sans objet occulté - ce qui revient à une classification avec uniquement des images à 1 objet) alors les algorithmes sont bien moins bons que des humains concentrés (aujourd'hui). Mais, justement, la position de ce manuscrit est que la question de l'évaluation est pertinente indépendamment de la performance qu'on est aujourd'hui capable d'atteindre.

Enfin, il existe des exemples (comme le SCALP) où on refuse de passer d'un algorithme de vision par ordinateur bas niveau à un algorithme d'apprentissage profond pourtant certainement meilleur mais moins compréhensible. Cela illustre que la confiance n'est pas qu'une histoire de performance (surtout quand l'évaluation de la performance n'est pas triviale).

5.4.3 Le No Free Lunch Theorem

Le chapitre 2 écarte l'idée qu'on puisse avoir des bornes sur l'erreur réelle que ferait un algorithme d'apprentissage à partir d'un tirage partiellement biaisé sans autre hypothèse. En effet, on ne peut pas, sans hypothèse supplémentaire, traiter le problème y, P si on observe le problème y, Q où Q est la distribution (biaisée) du tirage.

Plus globalement, le *no free lunch theorem* [197] démontre que la moyenne

de l'erreur réelle (hors de la base d'entraînement) sur l'ensemble des problèmes y, P est la même pour tous les modèles. D'ailleurs [197] pointe que le tirage d'une base d'apprentissage ne donne aucune information certaine sur les données hors de cette base. Il démontre qu'il est impossible non seulement d'avoir une erreur en moyenne plus faible que $\frac{1}{2}$ mais aussi d'avoir une estimation garantie de l'erreur (sinon, on parcourt les modèles et on prend celui qui donne la meilleure garantie après tirage des données d'apprentissage). Par contre, avec des hypothèses supplémentaires, il est éventuellement possible d'estimer l'écart entre y et f, plus ou moins indépendemment de P (et donc même avec un tirage biaisé). De nombreux travaux explorent cette voie comme [106].

Cette idée de restreindre l'ensemble des problèmes y, P considérés pour obtenir des garanties mathématiques sur l'ensemble restreint est intéressante. C'est d'ailleurs une idée assez logique si on voit la vision par ordinateur comme une science dont les lois seraient des propriétés des y, P qu'on rencontre en pratique. Mais cet horizon où l'on saura formuler de façon explicite les lois qui régissent ces distributions semble lointain.

L'IA de confiance vise cependant à avancer dans cette voie de façon plus ou moins implicite et grossière. En effet, il est clair qu'il y aura des discussions (scientifiques mais aussi sociales) entre les producteurs d'IA et les instances chargées de les autoriser pour trouver un accord sur les bonnes approximations pour pouvoir considérer qu'un tirage biaisé peut être considéré comme suffisant.

Indépendamment, on peut noter qu'un autre problème de la borne PAC classique est de ne pouvoir démontrer que $P(\text{erreur}_{\text{reelle}} \leq \varepsilon) \leq 1 - \delta$ et non pas que $\text{erreur}_{\text{reelle}} \leq \varepsilon$ (à supposer qu'elle s'applique). Ici encore, sans hypothèse supplémentaire, il est impossible de passer d'une borne en probabilité vers une borne déterministe. Mais, une borne déterministe est éventuellement possible en restreignant l'ensemble des distributions [159, 149, 194]. Cependant de tels résultats sont aussi balbutiants.

5.4.4 Les défenses anti-adversaires

Le chapitre 2 souligne que les exemples adversaires invisibles à l'oeil sont problématiques au sens où ils montrent que les modèles n'apprennent pas comme les humains. Par contre, il n'est pas clair qu'on puisse les produire dans le monde physique, et on dispose de *défense* contre ces attaques.

La première des grandes familles de défense est constituée des défenses *certifiantes* au sens où ces défenses permettent en certains points de prouver une stabilité locale. On y trouve les méthodes formelles comme [98]. Cependant, ces méthodes passent difficilement à l'échelle. On y trouve aussi des approches qui construisent un polygone englobant la zone accessible dans l'espace latent correspondante à un bruit borné dans l'espace de départ. Un des premiers travaux de ce type est [86]. Dans ces travaux, plus l'approximation polygonale englobante colle à l'espace réellement accessible, plus le taux de points *certifiés stables* se rapproche du taux réel.

Ces travaux peuvent servir en test (pour démontrer la stabilité en un point) mais aussi à l'apprentissage, car substituer un point plus proche de la frontière

de décision à la place du point original force le réseau à accroître la marge (c'est-à-dire la distance entre les points connus et la frontière de décision). Cela permet ainsi d'entraîner des réseaux plus ou moins à vaste marge. Dans [86], un formalisme unifié permet même de réaliser cette opération implicitement.

D'ailleurs, cette idée peut être implémentée avec les points extrêmes d'un polygone englobant mais aussi avec des points résultants d'attaques adversaires (réalisées non pas par un hackeur mais par le concepteur pour durcir son modèle). On parle alors d'apprentissage adversaire introduit dès [153].

Ce type d'attaque peut se voir comme une approximation interne de l'espace accessible via des attaques (plutôt qu'une approximation englobante). Plus l'approximation interne est couvrante, plus on se rapproche de méthodes certifiantes. Ainsi, apprendre avec une attaque faible comme FGSM [124] (pour *Fast Gradient Sign Method*) ne permet pas de résister à une attaque forte comme PGD [83] (pour *projected gradient descent* qui est grossièrement une version itérative de FGSM). Mais, apprendre avec une attaque forte comme PGD procure une robustesse significative, même si cette robustesse reste potentiellement vulnérable à une attaque encore plus forte à la différence des approches englobantes. Cependant, ces approches par approximation interne passent aussi bien à l'échelle que l'apprentissage des réseaux eux-même, ce qui n'est pas le cas des méthodes englobantes. Par exemple, PGD ne demande pas plus de mémoire GPU que l'apprentissage classique (puisqu'il ne se base que sur le même gradient - même s'il demande K fois plus de temps de calcul où K est le nombre d'itérations dans le PGD). Inversement, [86] nécessite 28Go de mémoire GPU pour apprendre un VGG sur des paquets de 2 images CIFAR10 : c'est 30 fois plus que pour PGD.

Ces deux apprentissages sur données adversaires (soit englobants soit internes) sont les deux principales familles de défenses illustrées en figure 5.3.

En plus de ces deux familles de défense, il existe aussi :

— Des défenses visant à diminuer l'information exposée par le réseau [71]. L'attaquant doit alors se baser sur des méthodes dites *boites noires* notamment sans gradient à l'opposé des méthodes classiques dites *boites blanches*. Cependant, il est parfois possible de créer une version lissée du modèle attaqué [102].

— Des défenses visant à régulariser globalement le comportement du réseau (et pas simplement localement). C'est notamment l'objectif des réseaux Lipschitz. En effet, par définition un réseau 1-Lips vérifie $|f(x) - f(x')| \leq ||x - x'||$. Ainsi, le simple fait de savoir que $|f(x)| > \varepsilon$ (et que f est 1-Lipschitz) garantit que $\forall x'$, $||x - x'|| \leq \varepsilon \Rightarrow f(x)f(x') > 0$. Cependant il n'est pas du tout trivial de contraindre un réseau à être 1-Lipschitz [11]. Notamment, contraindre chaque couche à l'être conduit à un réseau trop régulier sous une activation ReLU : [78] montre qu'on ne peut pas encoder la valeur absolue dans un réseau ReLU dont chaque couche est

11. Bien entendu, on peut diviser par un grand nombre mais la marge est réduite autant que le coefficient de Lipschitz. Ce qui est difficile c'est d'avoir une marge relativement large tout en ayant un réseau 1-Lipschitz.

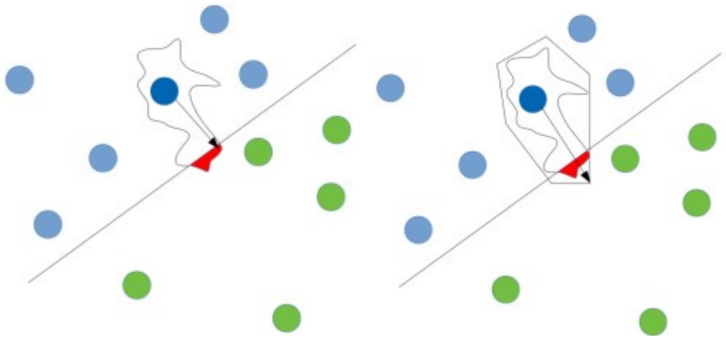


FIGURE 5.3 – Illustration des deux principales familles de défenses anti-adversaires : l'idée est de réaliser l'apprentissage non pas sur le point d'origine mais sur un point adversaire. Ce point peut être produit par une attaque volontaire (à gauche) : c'est une approximation interne de l'espace accessible via une perturbation facile à réaliser mais éventuellement vulnérable face à une attaque plus forte. Ce point peut aussi être le point extrême d'une approximation englobante de l'espace accessible.

1-Lipschitz. Cependant, l'introduction d'activations basées sur le réordonnancement des couches est potentiellement une solution [46]. Notons que [6] a récemment démontré un lien entre la sur-paramétrisation et le coefficient de Lipschitz minimal.

— Enfin, comme dans la théorie des jeux, les approches déterministes sont éventuellement incapables d'égaler les approches probabilistes. Ainsi, avoir non pas un réseau mais une distribution de réseaux permet d'augmenter la robustesse. C'est l'idée du *randomize smoothing* en anglais [50] qui permet une certification (probabiliste) des points.

L'ensemble de ces méthodes, notamment un simple apprentissage adversaire réalisé avec PGD, permet d'obtenir des réseaux relativement robustes aux attaques adversaires bornées en norme.

5.4.5 Les méthodes formelles et l'apprentissage ?

Le chapitre 2 ne parle pas du tout de l'utilisation de méthodes formelles pour vérifier des propriétés des algorithmes d'apprentissage par ordinateur alors que ce point est fortement mis en avant dans [11, 32, 15].

Il est clair que les méthodes formelles sont un moyen sûr de savoir si le réseau f est vulnérable à une attaque adversaire δ (borné par ε ou plus générale) en un point x (voir 5.4.4 ex : Reluplex [98]). Cependant, il y a une bonne raison pour laquelle on peut effectuer cette vérification : cela ne consiste *qu'à* savoir si $\exists \delta$ tel que $||\delta|| \leq \varepsilon$ et $f(x)f(x+\delta) < 0$. Ce qui est effectivement une assertion mathématique bien posée puisqu'on connaît ε, f, x. On peut alors réduire l'optimisation de δ à un énorme programme linéaire en nombre entier.

Mais, malheureusement, ce n'est pas de l'apprentissage par ordinateur au sens où il manque la fameuse fonction y qu'on ne connaît pas mais qu'un humain peut évaluer.

Ainsi, comme pour les bornes de type PAC, ces approches sont un des nombreux outils indispensables pour construire l'IA de confiance. Mais, ces méthodes ne sont pas de nature à donner des garanties sur des assertions qui impliquent y (par exemple une borne sur l'erreur réelle). Ainsi, elles ne peuvent pas seules produire des modèles sûrs.

D'ailleurs, même dans l'aviation où les codes critiques tendent à être vérifiés formellement, il convient de rappeler que cela n'implique pas une sûreté absolue car les méthodes formelles ne peuvent que prouver que le code valide ses spécifications. Mais une erreur de spécification est toujours possible. On peut penser au tragique incident de train d'atterrissage causé par un *bug* non pas dans le code mais dans les spécifications du code [12]. D'ailleurs, la DO-178 impose des processus de développement stricts que le code soit vérifié ou pas.

Dit autrement, une preuve mathématique consiste à *lier des hypothèses à une conclusion par de la logique*, mais on ne peut jamais prouver les hypothèses elles-même - sauf avec des hypothèses plus faibles mais il y a toujours des hypothèses primitives. Donc même avec des méthodes formelles, il pourra toujours y avoir des *bugs* (via les spécifications). D'ailleurs, dans le cas de l'apprentissage, le problème est justement qu'on ne connaît **pas** les spécifications i.e. y. Je me permets d'exclure l'utilisation de réseaux de neurones pour encoder une table connue [10]. Ces travaux peuvent avoir un intérêt industriel réel. Mais, ce n'est pas de l'apprentissage, c'est de la compression avec un modèle utilisé en apprentissage.

5.4.6 Détection d'évènements rares et apprentissage

Je n'ai pas parlé de détection d'évènements rares - *rare case event detection* en anglais - dans le chapitre 2. Il existe pourtant une littérature importante dans la communauté de la statistique sur ce sujet, dont l'ONERA est par ailleurs partie prenante (par exemple [137]). Cependant, ces travaux sur les cas rares ne sont généralement pas transposables à l'apprentissage par ordinateur.

Grossièrement, étant donnée une fonction ψ sur un compact X structuré, l'estimation de la probabilité des cas rares consiste à estimer $\frac{Vol(\{x,\psi(x)\geq\alpha\})}{Vol(X)} = \int\limits_{x\in X,\psi(x)\geq\alpha} \mathcal{U}(x)\mathrm{d}x$ (où $\mathcal{U}$ est une distribution uniforme sur X). Cette formulation est proche de celle de l'erreur réelle d'un modèle appris $\int\limits_{x\in X,\mathrm{sign}(f(x))\neq y(x)} P(x)\mathrm{d}x$ présentée en chapitre 2. Mais, il existe des différences majeures :

- Une première différence théorique est que ψ est décidable alors que l'évaluation de y nécessite un humain. Cependant, en pratique l'évaluation de ψ est potentiellement aussi coûteuse que celle de y.
- Par contre, une vraie différence est que ψ est supposée un minimum régulière, a minima continue, ce qui n'est pas le cas pour y.

12. www.reuters.com/article/uk-boeing-dreamliner-glitch-idUSLNE7A603L20111108

— Surtout, une autre différence est l'idée qu'on est sous un échantillonnage uniforme. Combiné avec le fait que ψ soit régulière, cela invite à considérer le guidage de l'échantillonnage dans X.
Mais, cela n'est pas possible en apprentissage car les distributions rencontrées peuvent être extrêmement creuses : créer une image ex-nihilo conduit presque sûrement à une image irréaliste. La question de savoir si les GAN sont une solution à ce problème est encore ouverte.

Par contre, la télédétection se rapproche un peu de ce cadre puisqu'on sera intéressé par une erreur uniforme vis-à-vis de la spatialisation, qu'on peut tirer de façon dirigée vis-à-vis de de cette spatialisation, ce qui ouvre la porte à un tirage guidé via le même type de méthode qu'en détection d'évènements rares. Cela dit ces techniques de détection d'évènements rares tournent autour d'un guidage via l'incertitude comme celui présenté en perspective (chapitre 4).

5.5 Un petit mot sur la soutenance

Contrairement au manuscrit, la soutenance se voulait *haut niveau* et s'est focalisée sur le message central : *l'IA de confiance c'est essayé de compenser l'impossibilité d'une mesure garantie du risque (quantitative) par une démonstration (qualitative et subjective) de notre maîtrise des méthodes d'apprentissage profond.*

En effet, depuis 1960 on pensait être en mesure d'estimer le risque de méthodes basées données via des théorèmes comme

$$P\left(\int \mathbf{1}_{-}(f(x)y(x))P(x)\mathrm{d}x \leq \sum_{k\in\{1,\dots,K\}} \mathbf{1}_{-}(f(x_k)y(x_k)) + \sqrt{\frac{-\log(\delta)}{2K}}\right) \geq 1-\delta$$

où $x_1, ..., x_K$ sont tirés i.i.d. selon P.

Malheureusement, l'apprentissage profond contredit cette idée ! En effet, même si tirer de façon i.i.d. selon P est évidemment impossible, l'hypothèse implicite de 1960 à 2012 était que *c'est pas trop grave de pas tirer exactement i.i.d. selon* P. Or, l'apprentissage profond a montré que les performances sont au contraire très sensibles aux changements de distributions (transferts, exemples adversaires et etc). Donc pour ces méthodes au moins (mais finalement la remise en question est plus globale), on ne peut **pas** utiliser des méthodes types Monte-Carlo pour estimer l'erreur réelle (sauf dans des cas très précis où on saurait borner l'erreur en norme L_1 entre P et la distribution du tirage). On n'a donc aucun moyen d'en fournir une borne garantie. Finalement P est un objet encore plus complexe que y qu'on peut évaluer avec un humain. Alors que P n'est ni décidable mathématiquement, ni évaluable via un humain. C'est le processus industriel de récolte des échantillons qui est soit suffisamment cadré de sorte qu'on puisse tirer selon P ou insuffisamment cadré et dans ce cas on a généralement pas d'idée de l'écart entre P et le tirage effectivement réalisé.

La *seule* alternative est de se contenter de démontrer *une maîtrise* (à court terme, car à long terme on sera peut-être en mesure de faire des hypothèses scientifiques sur les distributions qu'on peut rencontrer).

À coté de ce message très général, mon projet de recherche c'est d'utiliser des données mieux cadrées (par exemple contrainte par des lois physiques ou échantillonnable plus facilement car liés à des moyens d'acquisition statiques comme en satellitaire) pour augmenter puis démontrer (subjectivement) notre maîtrise.