Analysis of Deep Neural Networks Correlations with Human Subjects on a Perception Task

Loann Giovannangeli, Romain Giot, David Auber, Jenny Benois-Pineau, Romain Bourqui

▶ **To cite this version:**

Loann Giovannangeli, Romain Giot, David Auber, Jenny Benois-Pineau, Romain Bourqui. Analysis of Deep Neural Networks Correlations with Human Subjects on a Perception Task. IEEE, pp.129-136, 10.1109/IV53921.2021.00029 . hal-03259690

HAL Id: hal-03259690
https://hal.science/hal-03259690v1

Submitted on 14 Jun 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Analysis of Deep Neural Networks Correlations with Human Subjects on a Perception Task

Loann Giovannangeli*, Romain Giot*, David Auber*, Jenny Benois-Pineau* and Romain Bourqui*
*LaBRI, UMR CNRS 5800, University Bordeaux, Talence, France
{first}.{last}@u-bordeaux.fr

***Abstract*—In information visualization, it has become mandatory to assess visualization techniques efficiency either to write a survey, optimize a technique or even design a new one. To do so, the common way is to conduct user evaluations through which human subjects are asked to solve a task on different visualization techniques while their performances are measured to assess which technique is the most efficient. These evaluations can be complex to design and setup in order not to be biased and, in the end, their results can become contestable when the evaluation methods standards evolve. To overcome these flaws, new evaluation methods are emerging, mostly making use of modern and efficient computer vision techniques such as deep learning. These new methods rely on a strong assumption that has not been studied deeply enough yet: humans and deep learning models performances can be correlated.**

This paper explores the performances of both a state-of-the-art deep neural network and human subjects on an outlier detection task taken from a previous experiment of the literature. The objective is to study whether the machine and humans behaviors were different or if some correlations can be observed. Our study shows that their results are significantly correlated and a machine learning model efficiently learned to predict human performances using deep neural network metrics as input. Hence, this work presents a use case where using a deep neural network to assess human subjects performances is efficient.

***Index Terms*—Information visualization, Deep learning, User evaluation, Correlations, Automated evaluation**

I. Introduction

Since many years, information visualization has established itself as an efficient way to explore complex data [1]. Victim of its own success, the number of visualization techniques greatly increased. Now that the scope of techniques is very large, the need to evaluate them has emerged to answer questions such as "*which technique fits best to my data and task?*" or "*is my visualization technique better than existing ones?*". As it is complicated to quantify visualization techniques efficiency, it is common to compare them to assess their *relative* efficiency (*i.e.*, "VIS_A is better than VIS_B when *[condition]*"). Usually, these comparisons are conducted with user evaluations [2]–[4]. If user evaluations methods can be different, the main process they rely on is to ask a population of human subjects to solve a task on data represented with different visualization techniques, and measure their performances. The most efficient technique to solve the task will be the one which led to the best performances during the evaluation.

If user evaluations are the main (if not the only) accepted method to evaluate visualization techniques, it is principally because the targeted end-users of these techniques are human beings. Hence, evaluating the techniques efficiency on human subjects makes the experimental results generalization to all potential end-users (*i.e.*, humankind) straightforward. However, they also comprise limiting constraints and drawbacks, some of which are described in the following. (i) The experiments with users involvement can be complicated and time-consuming to setup and conduct. (ii) Special care must be taken in order not to bias the experiment with preliminary design choices [5] (*e.g.*, selected trials conditions, evaluation duration, subjects population). (iii) Subjects environments must all be equally conducive for conducting the experiment. (iv) It is complicated to recruit a sufficient number of subjects. These drawbacks make evaluations difficult to reproduce. Although it is possible to minimize them, there most of the time remain some subjective design choices that can be contested or become contestable when user evaluation standard methods evolve.

Lately, the great performances of Deep Learning (DL) on image analysis tasks led to an increase of the information visualization community interest in these techniques. More specifically, some researchers are studying how accurately DL techniques can assess representations readability and how well they can model human perception. Therewith, we expect to be able to overcome user evaluations flaws by using DL-based automated approaches. If some works have already shown how DL techniques could be used to help visualization techniques evaluations [6]–[9], only a few measured how Deep Neural Networks (DNNs) are correlated with humans on perception tasks [6], [10].

In this paper, we make use of an existing experiment data from Giovannangeli *et al.* [7] to explore the relations between DL techniques and human behaviors on perception tasks. In their study, the authors built a *difficulty metric* based on the performances of a Deep Neural Network to statistically assess human perception of an outlier detection task difficulty. The experiment itself and its data are detailed later in Section III.

The contributions in this paper are the qualitative and quantitative study of the correlations between a DNN and human subjects performances when requested to solve the same visual outlier detection task. In the qualitative study, we compare the DNN and humans known strategies to solve the task. The quantitative study measures the correlation coefficients between some DNN metrics and human subjects performances. These correlation coefficients are shown to be strong, and we further increase the understanding of the DNN results to assess human performances by fitting a Machine

Learning (ML) model to predict human performances from various DNN metrics. The model effectively approximated a function relating the DNN performances to human subjects performances and improved our understanding of how and when we can trust the DNN results to assess human behaviors. More generally, this work contributes to the *ML4VIS* [11] field by validating an application of DL techniques to assess human behaviors on perception tasks and shows an example where DNN results are strongly correlated with human performances.

The remainder of the paper is organized as follows. Section II presents related works on *ML4VIS* and existing comparisons of DNNs with human capabilities. Section III presents the original experiment from which we use the DNN and human subjects data to study if they are correlated. Section IV qualitatively presents the relations between the DNN and human subjects strategies to solve the task, while Section V presents the quantitative study of their performances correlations. Finally, we conclude the study in Section VI and present leads for future work.

II. Related works

A. Deep Neural Networks for Information Visualization

Since Deep Neural Networks (DNNs) have proven to be efficient computer vision techniques, several works made use of them to assess humans capabilities to solve a task. In a survey on quality metrics for information visualization, Behrisch *et al.* [9] stated that DNNs were a promising direction for evaluating representations qualities. Some studies continued exploring this thematic, comparing human subjects to DNNs on the interpretation of simple graphical elements [10] or even complex data structures representations [6], [8]. This field of research is getting more and more interest in the recent years as shown in the Wang *et al.* [11] survey on the application of Machine Learning (ML) techniques to various Information Visualization domains, which they called *ML4VIS*. Some studies are already defining generic workflows to make use of Deep Learning (DL) models to address some information visualization problems [6], [12]. These methods rely on the assumption that ML techniques could model the perception of graphical content by humans. Yet, we still lack knowledge about this cornerstone assumption and we currently know more about the differences between humans and ML techniques than we know about their correlations. This lack of knowledge makes it ambiguous how well these methods are suited for the problems they address and makes information visualization experts skeptical about them.

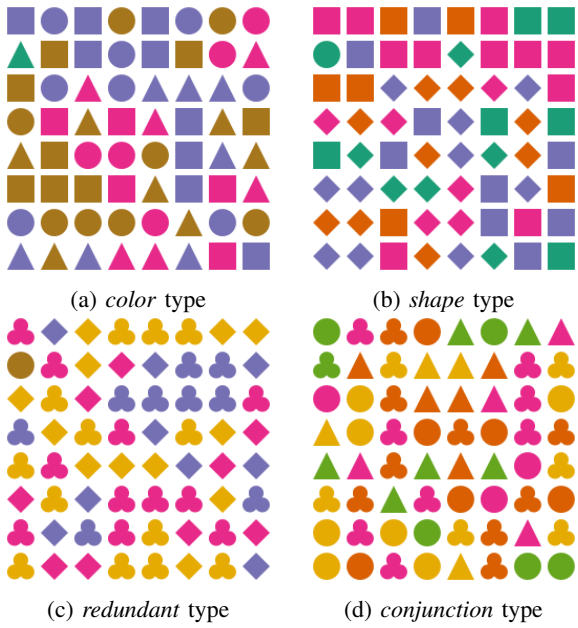
B. Deep Neural Networks and Humans

The comparison of computer vision techniques performances and human being capabilities were initially studied to motivate and drive the research axes of these computer programs. For example, Fleuret *et al.* [13] studied how some ML techniques performed compared to human subjects on a scene categorization task. They concluded that ML techniques performances remained worse than their subjects ones for this task. Stabinger *et al.* [14] extented Fleuret study by comparing two state-of-the-art DNNs (namely, LeNet and GoogLeNet) to Fleuret subjects performances on the same task and data. Again, they found human subjects to outperform these computer programs both in terms of accuracy and training time. They claimed that the better performances of human subjects were due to their abstraction capabilities and prior knowledge. Recently, Dodge and Karam [15] compared Convolutional Neural Networks (CNNs) and humans performances when asked to classify dogs on images with varying distortion. They found that DNNs accuracy was greater or equal to human accuracy on non-distorted images. As the distortion increases, their correlations decreases. Finally, CNNs were not able to efficiently solve the classification task on images with high distortion, while human subjects could still. They imputed this to humans capabilities to consider the global image and abstract its meaning even with very pronounced distortions. On the contrary, it becomes difficult for the CNNs convolution operations to extract features as images becomes noisy. They conclude that DNNs remain interesting tools despite their performances on distorted images since they are much faster than humans to solve the task. For example, they state that humans take one minute in average to classify an image in the ImageNet dataset, and are likely to lose performances as they fatigue. Dodge and Karam [16] continued this evaluation with another study where their human subjects were only exposed the images to classify for 100ms. Still, human subjects outperformed DNNs on distorted images. They concluded that humans higher accuracy was not favored by high-level interpretation capabilities since their better performances are reached using early human visual system only.

For the most part, these works compared humans and DNNs accuracy on classification tasks to study which was best to solve them under various circumstances. However, only a few of these works studied the *correlations* that could occur between them and none of them studied information visualization related tasks. Haehn *et al.* [10] reproduced existing perception studies with CNNs and compared their performances to human subjects ones. The studies consisted in evaluating a set of elementary graphical perceptual tasks across various encoding. They found that CNNs are not good models for human graphical perception and that their strengths are the opposite of humans ones. CNNs are better at solving tasks that require to interpret simple elements many times, whereas humans are more capable of interpreting complex representations requiring a higher level of abstraction or prior knowledge. Finally, Giovannangeli *et al.* [6] reproduced two information visualization evaluations of graph representations (*i.e.*, representations of complex data) with CNNs. They were able to draw the same conclusions as the experiments they were reproducing and that were conducted on human subjects. They concluded that the CNNs and humans performances might be correlated, although such a claim would require further investigations.

III. Original Experiment: Humans and ResNet Data

In this section, we describe the experiment we used the data of to study correlations between human performances solving



(a) *color* type (b) *shape* type

(c) *redundant* type (d) *conjunction* type

Fig. 1: Examples of images of the original experiment [7]. Each image has a different *Type*, has exactly 4 *#colors*, 3 *#shapes*, and the outlier is at position 8 (second row, first column).

an outlier detection task in abstract graphical content and the same task performed by ResNet [17], one of the most popular and efficient DNN frameworks.

A. Description of the Original Experiment

The Deep Neural Network (DNN) and human subjects data used to study correlations are taken from the Giovannangeli *et al.* [7] study of the capacity limit of *color* and *shape* visual attributes. They evaluated a task that consisted in locating an outlier stimulus in a 8×8 grid of colored shapes. The experiment parameters space comprized various possibilities of stimuli aspects. The task difficulty was defined by the outlier encoding Type and the display heterogeneity. *Type* is the parameter that encodes the visual attributes which make the outlier unique in the grid and was shown to be the major condition of the task difficulty. It has four values: *color* means the outlier color is unique in a grid; *shape* means the outlier shape is unique; *redundant* means its color *and* its shape are both unique; and *conjunction* means its combination color-shape is unique. The experiment parameters space is presented in Table I and example images can be seen in Figure 1.

B. DNN-based Metric

In their experiment [7], authors used a DNN-based metric to assess how variations of the outlier encoding and the display heterogeneity affected the task difficulty. The process consisted in training a DNN model architecture (in this case, ResNet [17]) to solve the task and statistically study how the variations of some conditions affected its performances, which is then interpreted as a *difficulty metric*. With this process, they were able to evaluate a broad parameters space (3290 configurations

Visual attribute values			Image		
Shape	**Color**	**Position**	**Type (Outlier encoding)**	**#colors**	**#shapes**
▲	#1B9E77	0	color	1	1
●	#D95F02	⋮		2	2
■	#7570B3	63	shape	3	3
♣	#E7298A			4	4
◆	#66A61E		redundant (red.)	5	5
	#E6AB02			6	
	#A6761D		conjunction (conj.)	7	

TABLE I: Parameters space of the original experiment [7]. Every stimulus was defined by a Shape, a Color and a Position corresponding to the row-major order of the grid. Each image was given an outlier encoding *Type*, number of colors (*#colors*) and number of shapes (*#shapes*). Underlined values were the ones kept and uniformly distributed after the DNN-based parameters space reduction for the user evaluation.

repeated 64 times each) and the resulting metric was designed to their specific task and data. The metric was finally used to refine their hypotheses and to sub-sample the parameters space of a user evaluation (see Table I underline parameter values). Finally, the latter user evaluation studied these hypotheses and measured the capacity limits of color and shape.

Their method rely on the assumption that the DNN and human subjects performances would be affected by the same conditions (variations in display heterogeneity and outlier encoding). However, they also supposed that the behaviors of humans and DNNs were not expected to be strictly correlated, and their study did not aim at exploring their hypothetical correlations. In this work, we want to verify this assertion.

C. Experimental Data

In this paper we aim at studying how human performances and DNN decision making are correlated. The DNN used for outlier detection task is ResNet [17] and the data are taken from the previous experiment of Giovannangeli *et al.* [7] later referred to as *original experiment*.

Human subjects data are made of the 22 subjects answers to the 44 trials of the evaluation. More specifically, subjects Response Time (RT) on each trial was measured and we can compute subjects Error Rates (ER) out of their answers. For each trial, a time limit was set to 30 seconds. If a subject did not answer within the 30 allotted seconds, a wrong answer is registered but its RT is not accounted. In this study, subjects performances were measured with their ER (between 0 and 1), their mean RT (between 0 and 30 seconds) and these measures standard deviations.

ResNet data are taken from the trained *ResNet50* [17] model performances on a *test dataset* (*i.e.*, data the model has not seen during its training), which corresponds to 21 056 samples. The model was pretrained on ImageNet and then fine-tuned with the objective to locate the outlier on the dataset of graphical images from [7]. Since the correlations between human subjects and DL techniques have not been much studied on perception tasks, we do not have any *a priori* about the ResNet performance metrics that would best enable to study these correlations. The intuition

is that as the task becomes more difficult, human subjects performances and ResNet performances will decrease. But while human subjects performances are commonly measured with ER and RT, there are more considerable metrics for DNN classifiers, each of them measuring different aspects of a model performances. Hence, the metrics we selected are among the most common ones with classification tasks in ML community:

- *Confidence*: Max value of the softmax prediction vector, *i.e.*, the predicted class probability
- *Categorical Cross Entropy (CCE)*: quantifies how accurately the model predicted the correct class based on the logarithm of the correct class predicted probability
- *Area Under the Curve (AUC)*: quantifies how well a model is able to distinguish correct from incorrect classes based on the Receiver Operator Characteristic (ROC) curve which plots TP rate against FP rate.
- *Categorical Hinge*: makes sure that the correct class probability is greater than the sum of incorrect classes probabilities by a safety margin
- $Error\,Rate\,(ER) = 1-(TP+TN)/(TP+FP+FN+TN)$
- $Recall = TP/(TP+FN)$
- $Precision = TP/(TP+FP)$
- $F1\ score = 2*(Recall*Precision)/(Recall+Precision)$

where TP and TN are respectively True Positive/Negative predictions; FP and FN are respectively False Positive/Negative predictions.

IV. ResNet and Human Subjects Strategies

In this section, we discuss a qualitative evaluation of the strategies employed by ResNet and human subjects to solve the task of outlier detection in graphical images and how these strategies can tell us about their correlations or uncorrelations.

A. Human Subjects Strategy: Perception and Visual Search

The Visual Search and Perception literature are cornerstones to understand how human brain would process the outlier detection task on such representations. According to this literature, and mainly the Treisman and Gelade *Feature-Integration Theory of Attention* [18], the human brain strategy will be the following. First, the display is *pre-attentively* processed: it is read as a whole (or several regions, *i.e.*, texture segregation) and, if the outlier *pops out*, the task is solved very easily. This corresponds to the combination of bottom-up attention (attraction by contrasts, outstanding colour our singularities in the orientation, [18]) and top-down one, as the visual search task was defined and subjects were instructed. Then, a *top-down* processing takes place: subjects look from most general to more specific elements of the display in a task-driven visual search. Finally, if the task is not solved yet, human brain goes through a *top-down* process of visual search: every element is serially processed until the visual task is solved. This behavior was confirmed in the original experiment [7] through the qualitative study of subjects' answers to a questionnaire. The absence of gaze fixation data in the original experiment unable us to verify the impact of other visual attention theories such as central fixation bias [19] on subjects performances.

B. ResNet Strategy

By design, Convolutional Neural Networks (CNNs) model a bottom-up process of their inputs. The first convolution layer(s) of a CNN detect pixel-wise information while later convolutions only work on more and more abstracted views of the input. Although the ResNet architecture is made of residual blocks to enable the model to work on different levels of abstraction at once, it still follows a bottom-up process. We assume that ResNet and human subjects performances would only be correlated on complex cases that require subjects to go through a serial process of the graphical patterns.

Built around convolution operations, CNNs work best at identifying objects outlines. Hence, we can expect ResNet to perform better when the outlier can be identified by its unique *shape*. Such assumption was mentioned as a limitation in the original experiment since it led to major differences between ResNet and subjects error rates. In this study, we assume that we will observe lower correlations between ResNet and humans on this condition.

Although ResNet and humans used different means to solve the task, the outcome of their respective strategies might still be correlated (*i.e.*, both strategies can be affected by the same conditions). It is the intuition on which *ML4VIS* is based on.

V. (Cor)Relations Analysis

In this section, we study the correlations and relationships between ResNet and human subjects experimental results. Since we do have more ResNet prediction samples than human subjects results, we compare their performances aggregated by various parameters. As stated in the original experiment, the parameters that affect most the task difficulty are *Type*, number of colors (*#colors*) and number of shapes (*#shapes*). We compare ResNet and the subjects averaged performances for each *Type-#colors-#shapes* combinations. This leads to 44 performances aggregations that are computed from 21 056 samples for ResNet and $44*22$ samples for human subjects. In addition to the subjects Error Rates (ER) and mean Response Times (RT), we consider their standard deviations since two conditions that have the same mean RT might very well not be of the same difficulty if their standard deviations are significantly different.

A. Correlation Coefficients

The first approach to study ResNet and subjects correlations is with usual correlation coefficients such as Pearson [20], which captures linear correlations. Since the Pearson correlation coefficient (commonly referred to as R) works best on normally distributed ensembles [21], we used the Kurtosis and Skewness statistics [22] to determine whether our data were normally distributed or not. When Kurtosis and Skewness are both between -2 and $+2$, data are considered normally distributed [5]. In the following, all considered data were found to be normally distributed, meaning that Pearson correlation coefficients and their corresponding *p-values* are reliable. As stated in the original experiment,

		Subjects Performances			
		ER	**ER STD**	**RT**	**RT STD**
ResNet Metrics	**Confidence**	-0.851	-0.77	-0.91	-0.7
	Area Under the Curve	-0.846	-0.756	-0.916	-0.676
	Categorical Cross Entropy	0.827	0.761	0.912	0.69
	Error rate	0.806	0.759	0.903	0.699
	Recall	-0.806	-0.759	-0.902	-0.699
	Precision	-0.806	-0.759	-0.902	-0.699
	Categorical Hinge	0.785	0.753	0.894	0.7
	F1 Score	-0.788	-0.75	-0.893	-0.687

TABLE II: Pearson correlation coefficients between ResNet and human subjects metrics (Error Rate, Error Rate standard deviation, Mean Response Time and Mean Response Time standard deviation). All correlation tests were significant ($p\text{-}value \ll 0.001$). Data are normally distributed according to Kurtosis and Skewness statistics, making Pearson coefficients reliable. ResNet metrics are ordered from most to less correlated to subjects performances in average.

Type is a parameter that has a strong impact on the task difficulty. According to *Type* definition (see Section III-A) and our assumption on ResNet strategy to solve the task (see Section IV-B), we expect ResNet–subjects correlations to be different on the various *Type* values. In the next, correlations will be studied *Overall* and *per Type*. Acceptance thresholds for Pearson tests *p-value* are $\alpha = 0.05$ Overall and $\alpha = 0.025$ per Type (due to a Bonferroni correction).

Table II presents the Overall correlation coefficients between ResNet metrics and subjects performances. ResNet metrics are ordered from strongest average coefficients with human performances to weakest ones. If all subjects performances metrics follow the order *lower is better*, it is not the case for all the ResNet metrics. Hence, we expect to observe negative correlation coefficients (*i.e.*, anticorrelations). For example, as the task becomes easier, we expect ResNet Confidence to increase whereas other metrics such as Error Rates are expected to decrease. Coefficients signs is not what we are interested in so we interpret both -1 and 1 as *strong* coefficients, whereas -0.1 and 0.1 are *weak*. As we can see, all the coefficients reveal strong correlations between ResNet and the subjects on all pairs of metrics, the weakest coefficient being -0.676 which remains relatively acceptable. ResNet metrics strongest correlations are with subjects mean Response Times where the coefficients are all stronger than 0.89. To compare, the correlation between subjects ER and subjects RT is $0.912; p\text{-}value \ll 0.001$. It means that the ResNet metrics correlations with subjects RT is of the same order of magnitude as the subjects ER one. ResNet metrics are also strongly correlated with the subjects ER, the mean of these coefficients absolute values being 0.814. Finally, subjects performances standard deviations correlations with ResNet metrics are weaker that the previous we observed, although they remain acceptable. Correlation coefficients with subjects ER STD are between 0.753 and 0.77, and are slightly even weaker with subjects RT STD where they range between 0.676 and 0.7 (considering their absolute values).

		Type			
	Overall	***color***	***shape***	***redundant***	***conjunction***
ER	-0.851	-0.798	-0.86	–	-0.89
ER STD	-0.77	-0.806	-0.881	–	-0.847
RT	-0.91	-0.861	**-0.487**	-0.762	-0.913
RT STD	-0.7	-0.813	**-0.414**	-0.821	-0.722

TABLE III: ResNet Confidence Pearson correlation coefficient with human subjects performances. Results are shown *Overall* and *per Type* (see Section III-A). Bold coefficients mean their corresponding *p-value* was higher than the acceptance threshold. The acceptance threshold was $\alpha = 0.05$ Overall, and $\alpha = 0.025$ per Type (Bonferroni correction). Correlations could not be computed on Type *redundant* ER and ER STD since the subjects never made any error under this condition (*i.e.*, their sequence of errors is constant).

Table II shows that ResNet Confidence is the metric that is the most correlated with subjects performances in average. In the following, we detail the correlations *per Type* between the subjects performances and ResNet Confidence only. This choice was made for the sake of readability and because all ResNet metrics follow the same trend as *Type* varies. The results of these correlation tests are presented in Table III. In the Overall column, one can see the results presented in Table II while other columns present the ResNet Confidence correlation coefficients with subjects performances for each Type value. On Type *color*, ResNet Confidence is strongly correlated with all the subjects performances ($R \leq -0.798$). As opposed to Overall, the correlations with subjects ER STD are not weaker than with subjects ER. Only a slight decrease can be observed with subjects RT STD compared to with subjects RT. On Type *shape*, ResNet Confidence is strongly correlated with both subjects ER and subjects ER STD. However, the coefficients of correlations with subjects RT and RT STD are weak and the corresponding *p-values* are above the acceptance threshold. With both low coefficients and high *p-values*, we can conclude that ResNet Confidence is uncorrelated with subjects RT and RT STD. On Type *redundant*, no correlation coefficient could be computed on ER and ER STD since subjects never made any errors on this condition. Nevertheless, we see that ResNet Confidence is strongly correlated with subjects RT and RT STD. It is noteworthy that the correlation with RT STD is stronger than with RT, which is the opposite of the trend we have seen on Overall results. Finally, on Type *conjunction* condition, ResNet Confidence correlation with subjects performances is always stronger than Overall. Again, correlations with subjects performances standard deviations are weaker than on their very performances.

In general, we observed that ResNet and human subjects performances are strongly correlated. It is encouraging and justifies the use of ResNet in the original experiment, although we have seen that ResNet ER was not the most optimal measure to base the *difficulty metric* on. But, these correlation coefficients might not tell us *what* is the relationship between ResNet and the subjects performances. Moreover, although

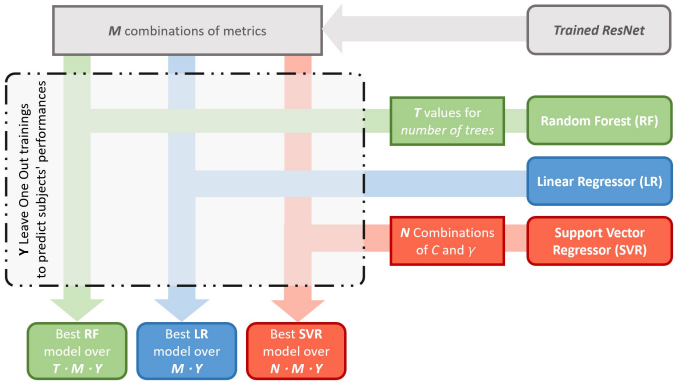


Fig. 2: Pipeline of the best regressor model selection. The three regressors are trained with all ResNet metrics combinations. In addition, some hyper-parameters Random Forest and SVR are fine-tuned. The pipeline results are presented in Table IV.

the correlations are strong, they might not be strong enough to be used as is to assess human performances. In fact, we believe that there are more chances that the relationship between ResNet and subjects performances is complex and that standard correlation coefficients which work on one-to-one dimensions may not be suited to this problem. For this reason, we will study how the combinations of several ResNet metrics can be related to human subjects performances.

B. Macine Learning Regression

Regression is a common way to find many-to-many or many-to-one relations. Our assumption is that if there is a function such that $f : X \rightarrow Y$ where X are ResNet performances and Y are human subjects performances, then there is a relation between the ResNet and the subjects performances. We look for *relations* here (as opposed to *correlations*) since, if such a relation was to exist, we could not assume that it admits a reciprocal. The existence of this relation alone would be sufficient since, in most *ML4VIS* use cases, Deep Learning techniques are used to assess human behaviors and not the other way around.

Since the performances we study are computed on 44 samples, Deep Learning techniques are not suited to this task as they would require more samples. We preferred them other state-of-the-art Machine Learning (ML) techniques that works well with fewer data. We tested three ML algorithms to search for the relation function. We used the Scikit-Learn python library [23] implementation of Linear Regressor, Random Forest and SVR (Support Vector Regression). We selected Linear Regressor for its simplicity and Random Forest and SVR as they are state-of-the-art and well proven models. We used Leave One Out cross-validation [24] to make reliable estimations of ML techniques with few samples and avoid sample-based bias. We tested two methods to predict subjects performances: (1) fit one regressor to predict all four subjects performance measures, and (2) fit four regressors to predict one subject measure each. We assumed that the second method could be more effective since each regressor could be optimized to predict its dedicated measure. However, the first method appeared to be more accurate by a slight margin and is the one that is discussed in the following. Figure 2 summarizes the best regressor selection pipeline. As shown in the figure, we tuned some Random Forest and SVR hyper-parameters. For Random Forest, the number of trees varied between 1 and 400. For SVR, the two hyper-parameters C and γ varied as proposed in [25]: $C \in \{2^{-5}; 2^{-3}; ...; 2^{15}\}$ and $\gamma \in \{2^{-15}; 2^{-13}; ...; 2^{3}\}$. No hyper-parameters was tuned for Linear Regressor models. The best selected models hyper-parameters are shown in Table IV. The three methods hyper-parameters that have not been addressed where left to their default value in the Scikit-Learn python library [23]. As we cannot *a priori* assert which ResNet metrics can be related to human subjects performances, each ML method was fit with all the (255) combinations of metrics. In the end, the best model is evaluated on the mean of 44 Leave One Out trainings Mean Absolute Error (MAE). Although the process is computationally expensive, it returns the best model for each ML regressor according to all the hyper-parameters and ResNet metrics combinations.

The best regressor models for each ML techniques are shown in Table IV. The coefficients of determination (R^2 scores) are defined by the square of Pearson correlation coefficients between a set of ground truths and predictions. We used R^2 scores to assess how well the models learned as it is a very common Machine Learning regression metric. As we can see, the ML technique that learned best to predict human performances from ResNet metrics is Random Forest, with an R^2 score of 0.956. The model used ResNet Confidence, AUC and Recall metrics to make its predictions and obtained low MAE scores for each performance to predict. In average, its prediction are only 4.4% away from the subjects ER, and 1.4s from their mean RT (which is low since RT ranges to up to 30 seconds). In view of the subjects performances, these MAEs are significantly small enough so that we can consider the model has found a relationship between the ResNet metrics it used and the subjects performances.

To verify that the regression approach finds a more accurate relation between ResNet and the subjects performances than standard one-to-one correlations, we computed the Pearson correlation coefficients between the subjects performances and the best regression model predictions (*i.e.*, the best Random Forest). We expect these correlations to be higher than the best ResNet metric correlations with subjects performances (*i.e.*, Confidence) presented in Table III. The results of the Pearson correlation coefficients on the best Random Forest model are shown in Table V. Overall, we see that the ML model predictions are strongly correlated with the subjects results, and are significantly stronger than the ResNet Confidence correlations with subjects performances. On Type *color* condition, the ML model predictions are strongly and almost perfectly correlated with subjects ER and RT ($R > 0.96$). As it was observed in Section V-A, we can see that correlations with subjects performances standard deviations are slightly weaker than with their very performances, but they remain above 0.89 here. On Type *shape*, the correlations between

				MAE between ML predictions and ground truths			
	R^2	Hyper-parameters	ResNet Metrics Used	ER	ER STD	RT	RT STD
Linear Regressor	0.715	–	Confidence, CCE, AUC, Recall	0.06	0.085	1.865	1.516
Random Forest*	0.956	9 trees	Confidence, AUC, Recall	0.044	0.059	1.418	1.166
SVR	0.753	$C = 2^9$; $\gamma = 2^3$	Confidence, CCE, AUC, Hinge	0.059	0.121	2.586	1.962

TABLE IV: Performances of the best model for each Machine Learning regressor. For each regressor, its best model R^2 score, hyper-parameters and ResNet metrics used to achieve its results are shown. The remaining columns correspond to the MAE between each regressor best model predictions and the corresponding subjects performances ground truths. Subjects ER range between 0 and 1 while subjects RT range between 0 and 30 (seconds), meaning that observed MAEs are low. The best regressor (noted with the symbol *) is the Random Forest.

		Type			
	Overall	*color*	*shape*	*redundant*	*conjunction*
ER	0.914	0.964	0.885	–	0.983
ER STD	0.809	0.892	0.893	–	0.983
RT	0.934	0.982	**0.557**	0.841	0.988
RT STD	0.786	0.925	**0.312**	0.872	0.966

TABLE V: Pearson correlation coefficients between human subjects performances and the best Random Forest model predictions of the subjects performances. Refer to Table III caption for more information.

the ML model predictions and human subjects performances are not significantly stronger than those observed in Table III with ResNet Confidence. No significant improvement is observed on correlations with subjects ER and ER STD; and no correlations can be observed between the model predictions and the subjects RT and RT STD. Again, on Type *redundant*, no correlation could be computed with subjects ER and ER STD (see Section V-A). Nevertheless, the ML model MAE on these two conditions is 0, meaning that the model captured the human perception of the task difficulty on this condition. Correlations are significantly stronger between the ML model predictions and subjects RT and RT STD than with ResNet Confidence. Finally, on Type *conjunction*, the Random Forest predictions are very close to being perfectly correlated with human subjects performances ($R \geq 0.966$).

Considering how accurately the best ML model approximated human subjects performances from ResNet metrics, we conclude there is a relationship between the ResNet and human subjects performances to solve the perception task. The only significant uncorrelation that was initially observed with standard one-to-one correlation coefficients studies and remained after the improvements found with the ML model regression is between ResNet performances and subjects Response Time on the Type *shape* condition. This uncorrelation is probably due to CNNs capabilities to detect outlines as presented in Section IV-B. With these results, we understand better how we must interpret ResNet performances to assess human performances in the original experiment task: we know that they are strongly correlated on all conditions but Type *shape* where subjects RT and RT STD should *not* be assessed from ResNet results.

VI. Conclusion

In this paper, we studied the correlations between a Deep Neural Network model (ResNet) and human subjects performances on a *perception task* consisting in identifying an outlier defined by its color and/or shape in a 8×8 grid of colored shapes. Although we know that humans and DNN models intrinsic strategies differ, their resulting performances could be correlated.

First, we showed that ResNet and human subjects results are strongly correlated with standard Pearson correlation coefficients between various ResNet metrics and subjects Error Rates and mean Response Time. To deepen the study of correlations, we used Machine Learning techniques to fit a model to predict human subjects performances on the main aggregations of conditions for the task difficulty. We expected that fitting a model to predict human performances with ResNet metrics as inputs would lead to stronger relations than those we observed with Pearson correlation coefficients on pairs of metrics. The best performing model shown that ResNet and human subjects performances to solve the task are strongly related, except when the outlier must be identified by its unique shape only. On this latter condition, their respective strategies efficiency might have been so different in favor of ResNet that it is not possible to find a relationship with human subjects Response Time.

These results show that the original study [7] assumption to use ResNet results as a difficulty metric that assesses the trends of difficulty that should be observed when humans solve the task was not conceptually wrong. However, our results also show that ResNet Error Rate was not the best metric to build such a metric on.

It is important to note that this paper has presented a study of a Deep Neural Network correlations with human subjects on a perception task. Hence, this work cannot be extended for designing an evaluation since it needs the human subjects data to compute correlation coefficients. This work rather aims at validating the use that was made of a DNN results to assess human performances in an evaluation. If experts remain skeptical about the use of DNNs to predict humans performances, these results are a step forward to trust Deep Learning techniques for predicting human performances on perception tasks. We believe this study benefits the *ML4VIS* field of research by participating in the clarification of the humans–DNNs correlations on visual interpretations tasks.

As our results have shown, the correlation is dependent on the visual attributes that were used to encode the data in the representations. As we can also assume that the correlations would be dependant on the DNN model architecture and the task considered, our major lead for future work is to propose a taxonomy based on DNN architecures, perception tasks and visual attributes (that are used to encode data). This taxonomy would guide experts on the DNN architecture to choose and inform them on how to read their trained DNN results to assess the task difficulty for human subjects and refine their experiments.

REFERENCES

[1] J. J. Thomas and K. A. Cook, Eds., *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. IEEE Computer Society, 2006.

[2] Sansen, Joris and Bourqui, Romain and Pinaud, Bruno and Purchase, Helen, "Edge visual encodings in matrix-based diagrams," in *2015 19th International Conference on Information Visualisation*. IEEE, 2015, pp. 62–67.

[3] Auber, David and Huet, Charles and Lambert, Antoine and Renoust, Benjamin and Sallaberry, Arnaud and Saulnier, Agnes, "Gospermap: Using a gosper curve for laying out hierarchical data," *IEEE transactions on visualization and computer graphics*, vol. 19, no. 11, pp. 1820–1832, 2013.

[4] Ghoniem, Mohammad and Fekete, J-D and Castagliola, Philippe, "A comparison of the readability of graphs using node-link and matrix-based representations," in *IEEE symposium on information visualization*. IEEE, 2004, pp. 17–24.

[5] H. Purchase, *Experimental Human-Computer Interaction: A Practical Guide with Visual Examples*. Cambridge University Press, 2012.

[6] Giovannangeli, Loann and Bourqui, Romain and Giot, Romain and Auber, David, "Toward automatic comparison of visualization techniques: Application to graph visualization," *Visual Informatics*, vol. 4, no. 2, pp. 86–98, 2020.

[7] Giovannangeli, Loann and Giot, Romain and Auber, David and Bourqui, Romain, "Impacts of the Numbers of Colors and Shapes on Outlier Detection: from Automated to User Evaluation," *arXiv preprint arXiv:2103.06084*, 2021.

[8] Haleem, Hammad and Wang, Yong and Puri, Abishek and Wadhwa, Sahil and Qu, Huamin, "Evaluating the readability of force directed graph layouts: A deep learning approach," *IEEE computer graphics and applications*, vol. 39, no. 4, pp. 40–53, 2019.

[9] Behrisch, Michael and Blumenschein, Michael and Kim, Nam Wook and Shao, Lin and El-Assady, Mennatallah and Fuchs, Johannes and Seebacher, Daniel and Diehl, Alexandra and Brandes, Ulrik and Pfister, Hanspeter and others, "Quality metrics for information visualization," in *Computer Graphics Forum*, vol. 37, no. 3. Wiley Online Library, 2018, pp. 625–662.

[10] Haehn, Daniel and Tompkin, James and Pfister, Hanspeter, "Evaluating 'graphical perception'with CNNs," *IEEE transactions on visualization and computer graphics*, vol. 25, no. 1, pp. 641–650, 2018.

[11] Wang, Qianwen and Chen, Zhutian and Wang, Yong and Qu, Huamin, "Applying Machine Learning Advances to Data Visualization: A Survey on ML4VIS," *arXiv preprint arXiv:2012.00467*, 2020.

[12] Wu, Aoyu and Wang, Yun and Shu, Xinhuan and Moritz, Dominik and Cui, Weiwei and Zhang, Haidong and Zhang, Dongmei and Qu, Huamin, "Survey on Artificial Intelligence Approaches for Visualization Data," *arXiv preprint arXiv:2102.01330*, 2021.

[13] Fleuret, François and Li, Ting and Dubout, Charles and Wampler, Emma K and Yantis, Steven and Geman, Donald, "Comparing machines and humans on a visual categorization test," *Proceedings of the National Academy of Sciences*, vol. 108, no. 43, pp. 17 621–17 625, 2011.

[14] Stabinger, Sebastian and Rodríguez-Sánchez, Antonio and Piater, Justus, "25 years of cnns: Can we compare to human abstraction capabilities?" in *International Conference on Artificial Neural Networks*. Springer, 2016, pp. 380–387.

[15] Dodge, Samuel and Karam, Lina, "A study and comparison of human and deep learning recognition performance under visual distortions," in *2017 26th international conference on computer communication and networks (ICCCN)*. IEEE, 2017, pp. 1–7.

[16] ——, "Can the early human visual system compete with Deep Neural Networks?" in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 2798–2804.

[17] K. He and X. Zhang and S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 770–778.

[18] A. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychology*, vol. 12, no. 1, pp. 97–136, 1980.

[19] Tatler, Benjamin W, "The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions," *Journal of vision*, vol. 7, no. 14, pp. 4–4, 2007.

[20] Lee Rodgers, Joseph and Nicewander, W Alan, "Thirteen ways to look at the correlation coefficient," *The American Statistician*, vol. 42, no. 1, pp. 59–66, 1988.

[21] Kowalski, Charles J, "On the effects of non-normality on the distribution of the sample product-moment correlation coefficient," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 21, no. 1, pp. 1–12, 1972.

[22] Zwillinger, Daniel and Kokoska, Stephen, *CRC standard probability and statistics tables and formulae*. Crc Press, 1999.

[23] Pedregosa, Fabian and Varoquaux, Gaël and Gramfort, Alexandre and Michel, Vincent and Thirion, Bertrand and Grisel, Olivier and Blondel, Mathieu and Prettenhofer, Peter and Weiss, Ron and Dubourg, Vincent and others, "Scikit-learn: Machine learning in Python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[24] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," *Statistics surveys*, vol. 4, pp. 40–79, 2010.

[25] Hsu, Chih-Wei and Chang, Chih-Chung and Lin, Chih-Jen and others, "A practical guide to support vector classification," 2003.