On the stability, correctness and plausibility of visual explanation methods based on feature importance

Romain Xu-Darme, Jenny Benois-Pineau, Romain Giot, Georges Quénot, Zakaria Chihani, Marie-Christine Rousset, Alexey Zhukov

On the stability, correctness and plausibility of visual explanation methods based on feature importance

Romain Xu-Darme[1,3], Jenny Benois-Pineau[2], Romain Giot[2], Georges Quénot[3], Zakaria Chihani[1], Marie-Christine Rousset[3], and Alexey Zhukov[2]

[1] Université Paris-Saclay, CEA, List, Palaiseau, France
[2] LaBRI UMR CNRS 5800, University of Bordeaux, Talence, France
[3] Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, Grenoble, France
romain.xu-darme@cea.fr

Abstract. In the field of Explainable AI, multiples evaluation metrics have been proposed in order to assess the quality of explanation methods *w.r.t.* a set of desired properties. In this work, we study the articulation between the stability, correctness and plausibility of explanations based on feature importance for image classifiers. We show that the existing metrics for evaluating these properties do not always agree, raising the issue of what constitutes a good evaluation metric for explanations. Finally, in the particular case of stability and correctness, we show the possible limitations of some evaluation metrics and propose new ones that take into account the local behaviour of the model under test.

Keywords: XAI · Saliency maps · Evaluation metrics

1 Introduction

The permeation of Artificial Intelligence (AI) in an increasing range of everyday applications is such that it seems hard to overstate its potential impact. Increasing trust towards AI systems is crucial both for social acceptability and for ethical purposes, and in this regards explainability is probably the most important dimension for helping users to better understand and control complex AI models. Indeed, when interfacing with the user, Deep Neural Networks (DNN) are usually black boxes, and the link between their inner-workings and their results remains obscure, notably due to their increasing size [5]. Thus, a considerable research effort has been deployed to alleviate this problem [24,6,17,26,23,28,27,22]. However, this plethora of methods sometimes lacks clear and measurable ways of comparison, making it hard for a DNN developer to make informed decisions on which method to choose. It is this difficulty that we aim to abate in this paper. In particular, we target attribution methods assigning scores to every input dimension of the data (*e.g.* pixel in image and video classification tasks), in a "white-box" (model-agnostic) setting [4] that relies on an access to the model (for example to compute gradients [22] or analyse the features in convolutional layers [6].

An explanation is often multifaceted information, which should satisfy a set of domain specific [20] properties[15], such as *correctness*, *stability* and *plausibility*. Correctness [15] (*a.k.a.* fidelity [3], faithfulness [29]) evaluates the adequacy between the explanation and the model behaviour. Stability [3] (*a.k.a.* continuity [15], sensitivity [31]) evaluates how similar explanations are for similar inputs. Plausibility (*a.k.a.* coherency [15]) evaluates the credibility of an explanation from the user point of view. The degree of adequacy between an explanation and a given set of properties is evaluated qualitatively and/or quantitatively using dedicated metrics.

Our contribution This work studies the articulation between the stability, correctness and plausibility of explanations based on feature importance for image classifiers. It shows that existing metrics for evaluating these properties may disagree, raising the issue of what constitutes a good evaluation metric for explanations. Finally, in the particular case of stability and correctness, it shows the limitations of some existing metrics and proposes new ones that take into account the local behaviour of the model under test. The paper is organized as follows: Sec. 2 presents the related work; Sec. 3 presents new metrics for measuring the stability and correctness of explanation methods; Sec. 4 presents our experimental results; Sec. 5 contains our closing remarks.

2 Related work

We recall that this work is related to the *evaluation* of explanation methods rather than the explanation methods themselves (see [4,12,21] for various surveys on the subject). We focus on the relation between *correctness*, *plausibility* and *stability*, three properties that are highly relevant for explanation methods applied to image classifiers, as opposed to properties such as compactness or covariate complexity [15] that are more suited to models processing tabular data.

Correctness Correctness can be evaluated using parameter randomisation [1] - which assesses the causal relationship between the model parameters and a given explanation method. However, an explanation method sensitive to parameter randomisation does not necessarily reflect the correct model behaviour. Alternatively, deletion and insertion metrics [17] evaluate the ability of a method to correctly identify the relative importance of each pixel *w.r.t.* to the model decision. The Area Under the Deletion Curve (AUDC) measures the variations of the model output when masking out pixels in the original image, from the most to the least important pixels. Pixels can be removed incrementally [17] or individually [3]. Similarly, the Average Drop (AD), Average Increase (AI) [7] and Average Gain (AG) [32] metrics study the model output when masking out unimportant pixels from the image. However, such metrics use Out-of-Distribution (OoD) samples (*i.e.,* inputs that significantly differ from the distribution of training data) to evaluate the local behaviour of the model [10]. Finally, the Causal Local Explanation metric [18,31] (CLE) assumes that a correct explanation should also

accurately predict the behaviour of the model in the neighbourhood of a given input. However, this metric does not take into account the potential instability of the model.

Stability Stability metrics hypothesise that an explanation method should produce similar explanations for similar inputs. The most common method [3] measures the maximum discrepancy between explanations in the neighbourhood of a given input. However, this metric also does not take into account the local model behaviour. In particular, in a region of high model instability, which can be identified using adversarial attacks [16], this metric penalises correct explanation methods in favour of more stable methods. Hence, more recent proposals [2] propose an evaluation of stability *relative* to the model behaviour.

Plausibility Plausibility evaluates the adequacy between an explanation and the *mental model* [14] of a human user, *i.e.,* how this user *thinks* the model is behaving. As such, plausibility evaluation requires additional information provided by human agents. For computer vision applications, this information can take the form of Gaze-Fixation Density Maps (GFDMs) that capture the distribution of human attention inside each image of a dataset. Then, plausibility can be evaluated by measuring the Pearson Correlation Coefficient (PCC) [33] or the Similarity [30] (SIM) between a saliency map and the "ground-truth" information. However, plausibility does not necessarily entail correctness [15]: if the model decision is biased, then a correct explanation method will likely produce implausible explanations.

3 Improving stability and correctness metrics through local surrogate models

As discussed above, current stability and correctness metrics do not always take the underlying model into account and may penalize correct explanation methods when the model displays high variations in a local region of the input space. Additionally, current correctness metrics tend to evaluate local explanations using perturbed inputs that significantly differ from the original image, assuming that the model behaviour remains stable in large portions of the input space. However, as demonstrated by adversarial attacks [16], such hypothesis may not always hold. Therefore, in this section we propose new metrics that not only use perturbed samples in a restricted neighbourhood around the original image (as in [18]), but also take into account the model behaviour.

We consider a model $f : \mathcal{X} \to \mathcal{Y}$ trained on a visual classification task with K possible classes. $\mathcal{X}$ usually corresponds to the space of RGB images of a given size $H \times W$, and $\mathcal{Y} = \mathbb{R}^K$ represents the output logits of the model (before softmax normalisation). For $k \in [1 \dots K]$, we denote $f_k : \mathcal{X} \to \mathbb{R}$ the restriction of f to the output logit for class k: *i.e.,* $f(X) = \big(f_1(X), \dots f_K(X)\big)$. During inference, the model decision $p(X)$ corresponds to the index of the highest value in $f(X)$, which represents the most probable class for X, *i.e.,* $p(X) = arg\max_{k \in [1 \dots K]}(f_k(X))$.

Let $X_0 \in \mathcal{X}$. We study the behaviour of various methods for explaining the classification result $p_0 = p(X_0)$, which is equivalent to explaining the output of $f_{p_0}(X_0)$. For simplicity, we denote $g(X) = f_{p_0}(X)$ the sub-model of f related to the most probable class predicted for X_0. Without loss of generality, we consider that an explanation method based on feature importance $s : \mathcal{X} \rightarrow \mathcal{S}$ is a function taking an input image X and producing a *saliency map* containing the relative importance of each pixel of X *w.r.t.* the output of g. In the context of visual classification, such explanations help answer the question "which part of the image contributed the most to the decision?". Note that methods such as FEM [9] or ML-FEM [6] are class-agnostic and therefore produce the same saliency map regardless of the target output logit.

Building a surrogate model As in [31], we build a surrogate of the model g from a saliency map $s(X_0)$ as:

$$\forall X \in \mathcal{X},\ E_{X_0}(X) = s(X_0)^T(X - X_0) + g(X_0) \in \mathbb{R} \tag{1}$$

where elements in $\mathcal{S}$ and $\mathcal{X}$ can be seen as vectors in $\mathbb{R}^{H \times W \times 3}$. In particular, $E_{X_0}(X_0) = g(X_0)$. The construction of E_{X_0} is motivated by the piece-wise linear nature of ReLU networks (such as ResNet50 [11] or VGG [25]) and is based on the principles of explanation maps [7] or gradient $\odot$ input [23] (where $\odot$ represents the element-wise multiplication).

Measuring stability To measure the local stability of a given explanation method s in the neighbourhood of $X_0 \in \mathcal{X}$, [3] introduces a metric equivalent to the Lipschitz criterion (LIP), under the postulate that a stable method s should produce similar explanations for similar inputs:

$$LIP(X_0) = \max_{\|X_0 - \tilde{X}\|_2 < \epsilon} \frac{\left\| s\left(X_0\right) - s(\tilde{X}) \right\|_2}{\left\| X_0 - \tilde{X} \right\|_2} \tag{2}$$

for some $\epsilon > 0$. However, this metric does not take into account the possible *in*stability of the model g in the neighbourhood of X_0. Since saliency maps aim to capture the local behaviour of the model, a correct explanation method applied to an unstable model might return a high value $LIP(X_0)$ and be discarded in favour of a less correct, but seemingly more "stable" method (*e.g.,* an explanation method returning a constant value such as Fake-CAM [19] will yield $LIP(X) = 0,\ \forall X \in \mathcal{X}$). In this work, we state that *stable explanation methods should produce similar explanations for similar model behaviours*. Therefore, we propose a revised stability metric, as illustrated in Fig. 1. Let $\tilde{X} \in \mathcal{X}$ s.t. $\left\| X_0 - \tilde{X} \right\|_2 < \epsilon$, $m = \left(X_0 + \tilde{X}\right)/2$ be the middle point between X_0 and $\tilde{X}$, and

$$D\left(X_0, \tilde{X}\right) = \left| E_{X_0}(m) - E_{\tilde{X}}(m) \right| \tag{3}$$

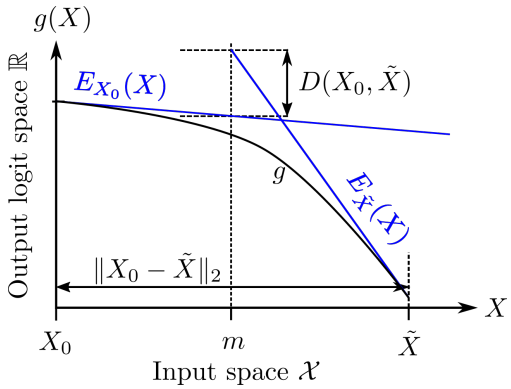

Fig. 1: Our LSS metric. For a given pair of inputs $\left(X_0, \tilde{X}\right)$ and an explanation method s, we build two surrogates E_{X_0} and $E_{\tilde{X}}$ approximating the local behaviour of the model. Our proposed metric measures how these surrogate models match at the mid-point between X_0 and $\tilde{X}$.

We define the *Local Surrogate Stability* (LSS) of the explanation method s around X_0 as

$$LSS(X_0) = \max_{\|X_0 - \tilde{X}\|_2 < \epsilon} \frac{D\left(X_0, \tilde{X}\right)}{\left\|X_0 - \tilde{X}\right\|_2} \tag{4}$$

In particular, for a constant explanation method s.t. $\forall X \in \mathcal{X}$, $s(X) = k$, then $D\left(X_0, \tilde{X}\right) = \left|k \times \left(\tilde{X} - X_0\right) + f\left(X_0\right) - f(\tilde{X})\right|$, *i.e.*, $D\left(X_0, \tilde{X}\right) \geq 0$.

Measuring correctness To measure the correctness of a given explanation method in the neighbourhood of X_0, [18] introduces the *Causal Local Explanation* metric (CLE):

$$CLE(X_0) = \mathbb{E}_{\left\|X_0 - \tilde{X}\right\|_2 < \epsilon} C\left(X_0, \tilde{X}\right) \tag{5}$$

where $\mathbb{E}$ is the expectation operator and $C\left(X_0, \tilde{X}\right)$ measures the local prediction error of the surrogate model E_{X_0} for an input $\tilde{X}$ and is equal to $\left|E_{X_0}(\tilde{X}) - g(\tilde{X})\right|$. However, as for the LIP metric, CLE does not take into account the possible instability of the model g in the neighbourhood of X_0. We propose to measure this prediction error, but *relative to* the changes in the model output between X_0 and $\tilde{X}$ (see Fig. 2). More precisely, we define the *Local Relative Correctness* (LRC)

$$LRC(X_0) = \mathbb{E}_{\left\|X_0 - \tilde{X}\right\|_2 < \epsilon} \frac{C\left(X_0, \tilde{X}\right)}{\Delta_g\left(X_0, \tilde{X}\right) + \eta^2} \tag{6}$$

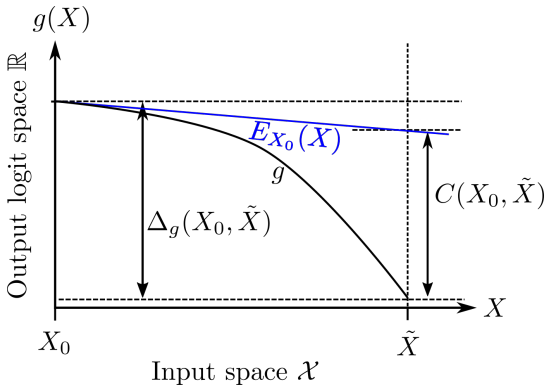

Fig. 2: Our LRC metric. For a given pair of inputs $\left(X_0, \tilde{X}\right)$ and an explanation method s, we build the surrogate model E_{X_0}, then measure the prediction error $C(X_0, \tilde{X})$ of the surrogate at $\tilde{X}$, but relative to the changes $\Delta_g(X_0, \tilde{X})$ in the output of the underlying model g.

Fig. 3: Illustration of our revised metrics for measuring the stability and correctness of explanations

where $\Delta_g\left(X_0, \tilde{X}\right) = \left|g\left(X_0\right) - g(\tilde{X})\right|$ and η is a small constant used for numerical stability.

4 Experiments and results

In this section, we present our experimental setup and results, with the goal of showcasing the limitations of state-of-the-art metrics described above, but also of studying the consensus between metrics from multiple perspectives.

4.1 Setup

Since evaluating the plausibility of explanations requires additional ground-truth information representing the user's expectations of the model, we restricted ourselves to datasets providing such information. In our experiments, we used the Salicon dataset [13], which provides Gaze Fixation Density Maps (GFDMs) for all images. For our classifier, we use a ResNet50 [11], processing images of size 256×256, pre-trained on the ImageNet [8] dataset and fine-tuned to the Salicon dataset with 20,000 images split into 10,000 images for training, 5,000 for validation and 5,000 for test. All explanation methods were evaluated on 50 images from the Salicon test dataset, that we further call Salicon50.

Explanation methods We evaluated several popular explanation methods: Grad-CAM [22], back-propagation [24] (denoted *Grads* in this work), Integrated Gradients [28] (black baseline, 10 samples), SmoothGrads [26] (10 samples, with noise

level 0.2), Guided Back-propagation [27] (GBP), FEM [9] and ML-FEM [6]. To showcase possible limitations of evaluation metrics, we also implemented the trivial method Fake-CAM [19] returning a saliency map obtained after upsampling a 7×7 map - equal to 0 on the top-left corner and 1 everywhere else - to the size of X_0. We also proposed *center-biased*-CAM (or CB-CAM), obtained after upsampling a 7×7 map - equal to 1 at its center and 0 everywhere else - to the size of X_0. Since images in popular classification datasets are usually centered on the object, the purpose of CB-CAM is to act as a baseline for plausibility evaluation metrics.

Evaluation metrics For the evaluation of *stability*, we compared LIP [3] to our proposed metric (LSS, see Eq. 4). For the evaluation of *correctness*, we compared the Deletion metric [17] (DEL), the Average Drop (AD)/Average Increase (AI)/Average Gain (AG) metrics [7,32], the Causal Local Explanation metric [18] (CLE) and our Local Relative Correctness (LRC, see Eq. 6). For the evaluation of *plausibility*, we computed the Pearson Correlation Coefficient (PCC) and the Similarity metric [30] (SIM) between the provided GFDMs and the saliency map generated by the explanation methods.

Sampling strategies As indicated in Sec. 3, LIP/LSS metrics (stability) and CLE/LRC metrics (correctness) evaluate the behaviour of explanation methods in a neighbourhood of X_0 defined by a radius ϵ *w.r.t.* the L2 distance. In this work, these metrics are estimated by sampling 50 perturbed inputs $\tilde{X}$ per image X_0 from the Salicon50 dataset, either using a *uniform* or *adversarial* strategy. When using the uniform strategy, each perturbed input $\tilde{X}$ is sampled uniformly in the n-dimensional sphere of radius ϵ around X_0, then rounded and clipped to values in $[0, 255]$. These last operations ensure that $\tilde{X}$ is a valid RGB image with integer pixel values.

The adversarial strategy helps identify regions in the neighbourhood of X_0 where the underlying model g might be unstable (high variation in the output) and to act as a sanity check for evaluation metrics using sampling. When using the adversarial strategy, each sample $\tilde{X}$ is generated with the goal of minimizing the value $g\left(\tilde{X}\right)$ as follows: we draw a target norm $d \in]0, \epsilon[$ from a uniform distribution; using back-propagation, starting from $A_0 = X_0 + \Phi$ - with Φ a small random perturbation (modification of 3 pixels in practice) - we generate the series $A_{i+1} = A_i - \nabla_X g(A_i)$ until $\|A_{i+1} - X_0\|_2 > d$, then set $\tilde{X} = A_i$. Similar to the uniform strategy, A_i is rounded in order to ensure that $\tilde{X}$ is a valid RGB image with integer pixel values. Since multiple perturbed samples $\tilde{X}$ must be generated from the same image X_0, the goal of the initial noise Φ is to ensure that the series will not systematically converge towards the same $\tilde{X}$. Finally, since evaluation metrics aim to measure *local* properties of explanations, we set $\epsilon = 250$ for both strategies in order to generate perturbed samples very close to original image (this roughly corresponds to switching one pixel from white to black). For additional results involving a Gaussian noise instead of a uniform noise, we refer the reader to [33].

Table 1: Evaluating properties on various explanation methods. Each value corresponds to the average score obtained by a given explanation method evaluated using a given metric over the Salicon50 dataset. For each metric, the best score is indicated in **bold**. ↑/↓ indicates that the lowest/highest value is better. For sample-based methods (LIP, LSS, CLE and LRC), we indicate the average score when using the uniform (uni.) or adversarial (adv.) strategy. For each correctness evaluation method, we indicate the average evaluation radius $\mathcal{R}$ (see Sec. 4.1) between the original images and evaluation samples.

Property	Evaluation metric	Explanation method								
		Grad-CAM [22]	FEM [9]	ML-FEM [6]	Grads [24]	Int.Grads [28]	SmoothGrads [26]	GBP [27]	Fake-CAM [19]	CB-CAM (ours)
Stability	LIP [3] ↓ (uni.)	0.01± 0.02	6.43± 7.82	1.86± 0.75	0.64± 1.33	0.90± 2.03	2.99± 3.78	9.05± 7.44	**0.00**± 0.00	**0.00**± 0.00
	LIP [3] ↓ (adv.)	1.79± 3.02	58.38± 64.21	8.20± 5.80	10.51± 20.75	8.99± 15.92	38.77± 47.87	26.35± 25.78	**0.00**± 0.00	**0.00**± 0.00
	LSS (ours) ↓ (uni)	0.00± 0.00	0.59± 0.16	0.09± 0.04	**0.00**± 0.00	0.00± 0.01	0.00± 0.00	0.21± 0.15	3.70± 2.57	0.61± 0.12
	LSS (ours) ↓ (adv)	0.59± 0.69	1.30± 0.73	0.63± 0.68	0.39± 0.57	**0.24**± 0.34	0.59± 0.69	1.29± 1.00	3.18± 1.33	1.42± 0.98
Correctness	CLE [18] ↓ (uni)	0.00± 0.00	0.18± 0.05	0.03± 0.01	**0.00**± 0.00	0.00± 0.00	0.00± 0.00	0.05± 0.03	1.55± 1.90	0.18± 0.04
	CLE [18] ↓ (adv)	0.15± 0.19	0.30± 0.24	0.16± 0.19	0.11± 0.16	**0.11**± 0.15	0.15± 0.19	0.18± 0.18	0.68± 0.52	0.31± 0.31
	LRC (ours) ↓ (uni)	0.00± 0.00	0.18± 0.05	0.03± 0.01	**0.00**± 0.00	0.00± 0.00	0.00± 0.00	0.05± 0.03	1.55± 1.90	0.18± 0.04
	LRC (ours) ↓ (adv)	0.10± 0.11	0.24± 0.16	0.11± 0.10	0.07± 0.08	**0.07**± 0.08	0.10± 0.11	0.13± 0.10	0.57± 0.41	0.25± 0.22
	DEL [17] ↓	13.94± 7.36	14.24± 7.02	13.07± 7.98	5.84± 5.55	5.23± 4.91	**3.81**± 3.80	6.09± 5.48	24.29± 1.60	15.88± 8.50
	AD [7] ↓	0.18± 0.26	0.39± 0.26	0.22± 0.24	0.02± 0.10	0.03± 0.12	0.04± 0.10	0.03± 0.09	**0.00**± 0.00	0.45± 0.23
	AG [32] ↑	0.00± 0.00	0.00± 0.00	0.00± 0.00	**0.00**± 0.00	0.00± 0.00	0.00± 0.00	0.00± 0.00	0.00± 0.00	0.00± 0.00
	AI [7] ↑	0.47± 0.50	0.14± 0.35	0.31± 0.47	**0.78**± 0.42	0.76± 0.43	0.45± 0.50	0.59± 0.50	0.53± 0.50	0.14± 0.35
Plausibility	PCC [33] ↑	0.28± 0.19	0.31± 0.20	**0.57**± 0.22	0.20± 0.13	0.19± 0.14	0.39± 0.12	0.24± 0.13	0.06± 0.02	0.35± 0.25
	SIM [30] ↑	0.37± 0.10	0.39± 0.11	**0.54**± 0.10	0.39± 0.10	0.38± 0.11	0.47± 0.09	0.40± 0.07	0.37± 0.12	0.33± 0.14
Average radius $\mathcal{R}$	CLE [18]/LRC	306 (uni.) / 233 (adv.)							306 (uni.) / 233 (adv.)	
	DEL [17]	38,767	38,343	39,428	39,194	39,216	39,336	43,283	1,882	24,062
	AD/AG/AI [7,32]	44,259	52,083	49,308	28,378	30,600	27,657	27,139	2,409	53,487

(a) Original (b) Deletion (c) Exp. map (d) Uni. sampling (e) Adv. sampling

Fig. 4: Evaluation radius of correctness metrics on a GRAD CAM explanation: (a) Original image; (b) After deleting 50% of the pixels when computing the AUDC (deletion metric) ($\mathcal{R} = 57,994$); (c) Explanation map used in AD/AI/AG ($\mathcal{R} = 752,614$); (d) Perturbed samples generated in the neighbourhood of original image for CLE/LRC using uniform sampling ($\mathcal{R} < \epsilon = 250$); (e) Perturbed samples generated in the neighbourhood of original image for CLE/LRC using adversarial sampling ($\mathcal{R} < \epsilon = 250$).

Locality of evaluation metrics The use of OoD samples for evaluating the correctness of an explanation describing the *local* behaviour of the model is debatable [10]. Thus, we restrict ourselves to replacing up to 50% of the pixels with black pixels when computing the AUDC for the DEL metric. Similarly, the use of explanation maps in AD/AI/AG metrics also creates images with large black regions (see Fig. 4). Thus, for a given correctness evaluation metric, we measure the average *evaluation radius* $\mathcal{R}$ between the original images from Salicon50 and the perturbed samples used for the evaluation: for the deletion metric (DEL), this corresponds to the L2 distance between the original image X_0 and the image where 50% of the most important pixels have been set to black; for AD/AI/AG, this corresponds to the L2 distance between X_0 and the explanation map $s(X_0) \odot X_0$; for CLE and our LRC metrics, this corresponds to the maximal L2 distance between X_0 and a perturbed sample $\tilde{X}$. Note that in this last case, the evaluation radius $\mathcal{R} < \epsilon$ by construction.

Consensus between metrics The consensus between evaluation metrics is measured as the Spearman Rank Correlation Coefficient (SR) between the mean scores of the metrics for all non-trivial explanation methods (*i.e.,* ignoring Fake-CAM and CB-CAM). Before computing the SR between the vectors of mean metric scores for AG/AI/PCC/SIM, we multiply them by -1 so that a strong consensus for two metrics is always represented by a SR close to 1.

4.2 Results and discussion

Table 1 indicates the mean scores of explanation methods over the Salicon50 dataset. As each metric measures a different quantity, comparisons can only be performed between explanation methods for a given metric (lines). Regarding the *stability* of explanation methods, our experiments show that trivial methods such as Fake-CAM and CB-CAM defeat the LIP metric but not our LSS metric,

Table 2: Spearman Rank Correlation Score (SR) between the vectors of average scores of all non-trivial explanation methods. The p-value is given in parenthesis for statistical significance. Results for sample-based evaluation metrics (LIP/LSS/CLE/LRC) are given when using a uniform sampling. A SR greater than 0.7 indicates a strong positive correlation (in **bold**).

Evaluation method	Stability			Correctness							Plausibility		
	LIP	LSS (ours)	Avg.	DEL	AD	AG	AI	CLE	LRC (ours)	Avg.	PCC	SIM	Avg.
LIP	-	**0.82** (0.02)	**0.82**	0.04 (0.94)	0.29 (0.53)	0.57 (0.18)	0.43 (0.34)	**0.82** (0.02)	**0.82** (0.02)	0.49	-0.29 (0.53)	-0.50 (0.25)	-0.39
LSS (ours)	**0.82** (0.02)	-	**0.82**	0.43 (0.34)	0.57 (0.18)	**0.75** (0.05)	0.61 (0.15)	**1.00** (0.00)	**1.00** (0.00)	**0.73**	-0.29 (0.53)	-0.21 (0.64)	-0.25
DEL	0.04 (0.94)	0.43 (0.34)	0.23	-	**0.71** (0.07)	**0.75** (0.05)	0.54 (0.22)	0.43 (0.34)	0.43 (0.34)	0.57	-0.29 (0.53)	0.29 (0.53)	0.00
AD	0.29 (0.53)	0.57 (0.18)	0.43	**0.71** (0.07)	-	**0.75** (0.05)	**0.96** (0.00)	0.57 (0.18)	0.57 (0.18)	**0.71**	-0.79 (0.04)	-0.11 (0.82)	-0.45
AG	0.57 (0.18)	**0.75** (0.05)	0.66	**0.75** (0.05)	**0.75** (0.05)	-	0.68 (0.09)	**0.75** (0.05)	**0.75** (0.05)	**0.74**	-0.39 (0.38)	0.07 (0.88)	-0.16
AI	0.43 (0.34)	0.61 (0.15)	0.52	0.54 (0.22)	**0.96** (0.00)	0.68 (0.09)	-	0.61 (0.15)	0.61 (0.15)	0.68	-0.86 (0.01)	-0.29 (0.53)	-0.57
CLE	**0.82** (0.02)	**1.00** (0.00)	**0.91**	0.43 (0.34)	0.57 (0.18)	**0.75** (0.05)	0.61 (0.15)	-	**1.00** (0.00)	0.67	-0.29 (0.53)	-0.21 (0.64)	-0.25
LRC (ours)	**0.82** (0.02)	**1.00** (0.00)	**0.91**	0.43 (0.34)	0.57 (0.18)	**0.75** (0.05)	0.61 (0.15)	**1.00** (0.00)	-	0.67	-0.29 (0.53)	-0.21 (0.64)	-0.25
PCC	-0.29 (0.53)	-0.29 (0.53)	-0.29	-0.29 (0.53)	-0.79 (0.04)	-0.39 (0.38)	-0.86 (0.01)	-0.29 (0.53)	-0.29 (0.53)	-0.48	-	0.61 (0.15)	0.61
SIM	-0.50 (0.25)	-0.21 (0.64)	-0.36	0.29 (0.53)	-0.11 (0.82)	0.07 (0.88)	-0.29 (0.53)	-0.21 (0.64)	-0.21 (0.64)	-0.08	0.61 (0.15)	-	0.61

which actually indicates that Grads (uniform sampling) or Integrated Gradients (adversarial sampling) are the most stable among "non-trivial" methods. Regarding the *correctness* of explanation methods, most evaluation metrics also indicate that Grads and Integrated Gradients are the best performing methods. It is also interesting to note that Grad-CAM and SmoothGrads have similar scores across all evaluation metrics except DEL and AD. As expected, Fake-CAM represents the best performing explanation method when using the AD metric, confirming the results of [32]. Moreover, as indicated by the average evaluation radius $\mathcal{R}$ (in parenthesis in Table 1), CLE and LRC evaluate explanation methods using samples that are significantly closer to the original image *w.r.t.* to the L2 distance and that are more likely to reflect the local adequacy between the original classifier and the surrogate model built from the explanation maps. Finally, regarding the *plausibility* of explanation methods, our experiments show that saliency maps produced by ML-FEM and FEM are most correlated to the ground truth GFDMs. However, we also note that our trivial explanation CB-CAM correlates more strongly with the GFDMs than most non-trivial methods when using PCC.

Consensus between evaluation metrics From Table 2, it can be stated that evaluation metrics do not always agree on the best performing explanation methods *w.r.t.* the three properties under study.

In particular, the LIP metric does not correlate strongly with correctness methods (0.49 on average), while *our LSS stability metric seems to bridge the gap between LIP and correctness evaluation metrics*, strongly correlating with both sets of methods. Since our proposed LRC metric for evaluating correctness is similar to CLE (when using uniform sampling), the correlation between the two methods is very strong, as expected. Finally, we notice that both plausibility metrics seem very *uncorrelated* to all other metrics. In summary, although Grads seems to produce the most stable and correct explanations *on average* (see Table 1), the lack of conclusive consensus between various metrics evaluating the same property (or across multiple properties) raises several questions: in the absence of a formal definition of what a correct and/or stable explanation method should be, how to decide which metric best evaluates the desired property? And if no explanation method can simultaneously satisfy all the desired properties, how to achieve a compromise between diverging requirements?

Consistency w.r.t. *perturbed samples* To study the sensitivity of sample-based evaluation metrics to the choice of perturbed samples $\tilde{X}$ in the neighbourhood of X_0, we measure the PCC between vectors of mean scores of non-trivial explanation methods when using a uniform and adversarial sampling strategy. This score measures the *consistency* of the evaluation metric, when perturbed samples are potentially located in a region of instability for the model. Table 3 indicates that while our LRC metric offers a marginal improvement in consistency over CLE, our LSS metric is significantly more robust to adversarial samples than LIP.

Table 3: Consistency of sample-based evaluation metrics, measured as the PCC between mean scores over non-trivial explanation methods when using uniform or adversarial sampling. The p-value is given in parenthesis. Higher is best.

Property	Evaluation method	Consistency ↑
Stability	LIP	0.678 (0.09)
	LSS (ours)	**0.832** (0.02)
Correctness	CLE	0.954 (0.00)
	LRC (ours)	**0.972** (0.00)

5 Conclusion and perspectives

In this work, we showed that the use of evaluation metrics for measuring the quality of explanations based on feature importance *w.r.t.* a set of properties is not a panacea to electing the best explanation method. Indeed, metrics evaluating the same property - either stability, correctness or plausibility in this paper - do not necessarily agree, a limitation that also prevents a more global consensus across properties. In this regards, it would be interesting to check whether explanations generated from a classifier trained using GFDMs as auxiliary annotations [6] would reconcile these three properties. More generally, there exists a mutual dependency between the model, explanation methods and evaluation metrics, that is difficult to disentangle in the absence of any form of ground truth or formal definition of the desired properties.
This work also introduces improved metrics for evaluating the stability and correctness of explanations that are less sensitive to the model behaviour than current state-of-the-art metrics and that use samples in a close neighbourhood around the original image. In a future work, we would like to extend our study to more recent proposals - such as the metrics proposed by [2] for stability - and to other properties of explanations such as completeness - the extent to which an explanation covers the model behaviour - or consistency (relevant for sampling based methods).

Acknowledgments Experiments were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several universities and organizations. This work has been partially supported by MIAI@Grenoble Alpes (ANR-19-P3IA-0003), TAILOR and TRUMPET, projects funded respectively by EU H2020 research and innovation program GA No 952215 and Horizon Europe GA No 101070038. This work is part of the scientific network GIS Albatros (ALliance Between Universities in nouvelle Aquitaine and Thales in Research on aviOnics Systems).

References

1. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. p. 9525–9536. NIPS'18, Curran Associates Inc., Red Hook, NY, USA (2018)
2. Agarwal, C., Johnson, N., Pawelczyk, M., Krishna, S., Saxena, E., Zitnik, M., Lakkaraju, H.: Rethinking stability for attribution-based explanations (2022)
3. Alvarez Melis, D., Jaakkola, T.: Towards robust interpretability with self-explaining neural networks. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 31. Curran Associates, Inc. (2018), https://proceedings.neurips.cc/paper_files/paper/2018/file/3e9f0fc9b2f89e043bc6233994dfcf76-Paper.pdf
4. Ayyar, M.P., Benois-Pineau, J., Zemmari, A.: Review of white box methods for explanations of convolutional neural networks in image classification tasks. J. Electronic Imaging **30**(5) (2021). https://doi.org/10.1117/1.jei.30.5.050901, https://doi.org/10.1117/1.jei.30.5.050901
5. Benois-Pineau, J., Bourqui, R., Petkovic, D., Quenot, G.: Explainable Deep Learning AI: Methods and Challenges. Elsevier Science (2023), https://books.google.fr/books?id=WHt5EAAAQBAJ
6. Bourroux, L., Benois-Pineau, J., Bourqui, R., Giot, R.: Multi Layered Feature Explanation Method for Convolutional Neural Networks ⋆. In: International Conference on Pattern Recognition and Artificial Intelligence (ICPRAI). Paris, France (Jun 2022). https://doi.org/10.1007/978-3-031-09037-0_49, https://hal.science/hal-03689004
7. Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE (mar 2018). https://doi.org/10.1109/wacv.2018.00097, https://doi.org/10.1109%2Fwacv.2018.00097
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009). https://doi.org/10.1109/CVPR.2009.5206848
9. Fuad, K.A.A., Martin, P.E., Giot, R., Bourqui, R., Benois-Pineau, J., Zemmari, A.: Features understanding in 3d cnns for actions recognition in video. 2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA) pp. 1–6 (2020)
10. Gomez, T., Fr'eour, T., Mouchère, H.: Metrics for saliency map evaluation of deep learning explanation methods. In: International Conferences on Pattern Recognition and Artificial Intelligence (2022)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
12. Islam, S.R., Eberle, W., Ghafoor, S.K., Ahmed, M.: Explainable artificial intelligence approaches: A survey. ArXiv **abs/2101.09429** (2021)
13. Jiang, M., Huang, S., Duan, J., Zhao, Q.: Salicon: Saliency in context. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1072–1080 (2015). https://doi.org/10.1109/CVPR.2015.7298710
14. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence **267**, 1–38 (2019)

15. Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., Seifert, C.: From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. CoRR **abs/2201.08164** (2022)
16. Nie, W., Zhang, Y., Patel, A.B.: A theoretical explanation for perplexing behaviors of backpropagation-based visualizations. In: International Conference on Machine Learning (2018)
17. Petsiuk, V., Das, A., Saenko, K.: Rise: Randomized input sampling for explanation of black-box models. In: BMVC (2018)
18. Plumb, G., Molitor, D., Talwalkar, A.S.: Model agnostic supervised local explanations. In: Neural Information Processing Systems (2018)
19. Poppi, S., Cornia, M., Baraldi, L., Cucchiara, R.: Revisiting the evaluation of class activation mapping for explainability: A novel metric and experimental analysis. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) pp. 2299–2304 (2021)
20. Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., Zhong, C.: Interpretable machine learning: Fundamental principles and 10 grand challenges. Statistics Surveys **16**(none), 1 – 85 (2022). https://doi.org/10.1214/21-SS133, `https://doi.org/10.1214/21-SS133`
21. Saleem, R., Yuan, B., Kurugollu, F., Anjum, A., Liu, L.: Explaining deep neural networks: A survey on the global interpretation methods. Neurocomputing **513**, 165–180 (2022)
22. Selvaraju, R.R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. International Journal of Computer Vision **128**, 336–359 (2016)
23. Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. In: International Conference on Machine Learning (2017)
24. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. In: Workshop at International Conference on Learning Representations (2014)
25. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (2015)
26. Smilkov, D., Thorat, N., Kim, B., Viégas, F.B., Wattenberg, M.: Smoothgrad: removing noise by adding noise. Proceedings of the ICML Workshop on Visualization for Deep Learning (2017)
27. Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.A.: Striving for simplicity: The all convolutional net. CoRR **abs/1412.6806** (2014)
28. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 70, pp. 3319–3328. PMLR (06–11 Aug 2017), `https://proceedings.mlr.press/v70/sundararajan17a.html`
29. Tomsett, R.J., Harborne, D., Chakraborty, S., Gurram, P.K., Preece, A.D.: Sanity checks for saliency metrics. Proceedings of the AAAI Conference on Artificial Intelligence pp. 6021–6029 (2020)
30. Wang, W., Shen, J.: Deep visual attention prediction. IEEE Transactions on Image Processing **27**, 2368–2378 (2017)
31. Yeh, C.K., Hsieh, C.Y., Suggala, A.S., Inouye, D.I., Ravikumar, P.: On the (in)Fidelity and Sensitivity of Explanations. Curran Associates Inc., Red Hook, NY, USA (2019)

32. Zhang, H., Torres, F., Sicre, R., Avrithis, Y., Ayache, S.: Opti-cam: Optimizing saliency maps for interpretability (2023). https://doi.org/10.48550/ARXIV.2301.07002, https://arxiv.org/abs/2301.07002
33. Zhukov, A., Benois-Pineau, J., Giot, R.: Evaluation of explanation methods of ai – cnns in image classification tasks with reference-based and no-reference metrics. Advances in Artificial Intelligence and Machine Learning (3), 620 – 646 (2023). https://doi.org/10.54364/AAIML.2023.1143