Representativity and Consistency Measures for Deep Neural Network Explanations

Thomas Fel, David Vigouroux

Representativity & Consistency Measures for Deep Neural Network Explanations

Thomas FEL
IRT Saint-Exupéry
thomas.fel@irt-saintexupery.com

David VIGOUROUX
IRT Saint-Exupéry
david.vigouroux@irt-saintexupery.com

Abstract

The adoption of machine learning in critical contexts requires a reliable explanation of why the algorithm makes certain predictions. To address this issue, many methods have been proposed to explain the predictions of these black box models. Despite the choice of those many methods, little effort has been made to ensure that the explanations produced are objectively relevant. While it is possible to establish a number of desirable properties of a good explanation, it is more difficult to evaluate them. As a result, no measures are actually associated with the properties of consistency and generalization of explanations. We are introducing a new procedure to compute two new measures, Relative Consistency $ReCo$ and Mean Generalization $MeGe$, respectively for consistency and generalization of explanations. Our results on several image classification datasets using progressively degraded models allow us to validate empirically the reliability of those measures. We compare the results obtained with those of existing measures. Finally we demonstrate the potential of the measures by applying them to different families of models, revealing an interesting link between gradient-based explanations methods and 1-Lipschitz networks.

1. Introduction

Machine learning techniques such as deep neural networks have become essential in multiple domains such as image classification, language processing and speech recognition. These techniques have achieved excellent predictive capability, allowing them to match human performances in many cases [18, 33].

However, their advantages come with major drawbacks, especially concerning the difficulty of interpreting their decisions, since they are black box models [20]. This problem is a serious obstacle to the adoption of these systems in safety-critical contexts such as aeronautics or medicine.

Recently, many strategies have been proposed to help users understand the underlying logics that led those models to a particular decision. While some methods offer explanations that are satisfactory to users, they do not, however, reflect the real behaviour of the model and some works has shown the potential pitfalls associated with current explanations methods [1, 11]. The explanation that was intended to provide confidence, is itself questionable.

Those observations have given rise to the need for an objective assessment of the explanations produced by these methods, thus enabling benchmarks and baselines to be established. To do this, one approach advocated by various works is to ensure that the explanations satisfy a certain number of properties (or axioms), such as Fidelity, Stability, Representativity or Consistency. Since some works [42, 22, 29, 12, 2] propose an exhaustive list of these properties, a good explanation could then be defined as quantitatively satisfactory according to a coherent set of measurements specific to each of these properties.

This work proposes a methodology applicable to a large family of models, based on distance between explanations coming from the same sample. We use this new methodology to introduces two new global measures, Relative Consistency ($ReCo$) and Mean Generalizability ($MeGe$). $ReCo$ is motivated by the idea that one explanation should not be used to justify two contradictory decisions. $MeGe$ is intended to measure the ability of a model to derive general rules from its explanations.

We experiment with this procedure to evaluate the two measures on different datasets. The results we obtain between normally trained models and degraded models allow us to assess that the measures reflect the loss of generalization and consistency of the explanations. We compare these results with those of existing metrics : **Fidelity** and **Stability**, allowing us to establish a ranking of the tested methods. Finally, we use those measures to highlight in a quantitative way the suggestions of different works on the generalization and consistency of 1-Lipschitz networks.

2. Related Works

Several works have recently proposed methods to explain the decisions of Deep Neural Network (DNN). They can nevertheless be classified in two categories: global and local [8], local methods focus on the reasons for a specific decision, while global explainability consists in understanding the general patterns that govern the model. We focus on local methods in this work.

Concerning local explanation methods, one of the main techniques relies on the model by using input perturbations to highlight important characteristics [48, 50, 19, 10, 52, 26, 27] or by using the gradient and decomposition of the model [49, 36, 32, 4, 35, 41, 37].

However, most of the research on explainability has focused on the development of new methods. Despite a wide range of estimators, there is a lack of research on the development of measures and approaches for assessing quality of explainability. One of the reasons why research has not focused on the evaluation (and therefore quality) of methods is the difficulty of obtaining objective ground truths [30]. In order to formally apprehend the problem, several works propose to define a system of general properties that must be satisfied by explanations [42, 22, 29, 12, 2]. Among those work, we can identify 5 major properties.

Definition 1 ***Fidelity***
The ability of the explanations to reflect the behaviour of the prediction model.

Definition 2 ***Stability***
The degree to which similar explanations are given for similar samples of the same class.

Definition 3 ***Comprehensibility***
The ability to describe internal elements of a system in a way that is understandable to humans.

Definition 4 ***Representativity***
The generalizability of the explanations, the extent to which the explanations are representative of the model.

Definition 5 ***Consistency***
The degree to which different models trained on the same task give the same explanations.

There are two approaches in recent work to evaluate explanations. The first subjective approach consists in putting the human at the heart of the process, either by explicitly asking for human feedback, or by indirectly measuring the performance of the human/model duo [17, 6, 23] such as the ITR transfer information rate [31]. Nevertheless, human intervention sometimes brings undesirable effects. The work of [1] illustrated an example of possible confirmation bias.

In the context of computer vision, a second approach has emerged and consists of using the model, allowing for objective quantitative measures. These measures essentially aim to measure two properties: **Fidelity** and **Stability**. Regarding **Fidelity**, [30] were the first to propose a measure based on the change of pixels, by comparing the drop in score when the pixels of interest are inverted. Several variants exist. ROAR [15] allowing to ensure that the drop in score does not come from a change in distribution by re-training the model. IROF [28] which requires low resources to be calculated. This correlation between the attributions for each pixel and the difference in the prediction score when they are modified can be measured according to Equation 1 from [5].

$$\mu F = \underset{S}{\text{corr}} \left(\sum_{i \in S} g(f, x)_i, f(x) - f(x_{[x_i = \bar{x}_i, i \in S]}) \right) \quad (1)$$

Where f is a predictor, g an explanation function, x a point of interest, S a subset of indices of x and $\bar{x}$ a baseline reference. The choice of a proper baseline is still an active area of research [40].

For **Stability**, there are two measures which consist in calculating the sensibility of the explanation around a point of interest [47, 3, 5]. In particular, we find S_{avg} the average sensitivity of the explanation. Formally, in a neighborhood of points of radius r around x, giving a distance between samples ρ, and D a distance between explanations:

$$S_{avg} = \int_{z:\ \rho(x,z) \leqslant r} D(g(f, x), g(f, z)) \mathbb{P}(z) dz \quad (2)$$

Up to our knowledge, several necessary properties do not have an associated measure, notably **Consistency** and **Representativity**. Indeed, a model that overfitted can give a faithful and stable explanation, but specific to a given input, and therefore not general. In the same way, an explanation, can be faithful, stable and inconsistent.

We propose an approach involving cross-validation on explanations to evaluate these properties through two new global measures : the Relative Consistency $ReCo$ and the Mean Generalization $MeGe$.

3. Method

In the first instance, we give a simple motivation behind our measures. In a second step, we propose a training procedure applicable on a large family of models, which will allow the application of both measures. Finally, the measures are introduced.

3.1. Notations

Restricting the scope of this work to supervised classification settings, where $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} \subseteq \mathbb{R}^c$ designate the input and label spaces, respectively. We denote

$\mathcal{D} = \{(x_n, y_n)\}_{n=0}^{N}$ a dataset, $\mathcal{D}_i$ a subset of $\mathcal{D}$, and $f_{\mathcal{D}_i} : \mathcal{X} \to \mathcal{Y}$ a black box predictor from a family of model $\mathcal{F}$ trained on the dataset $\mathcal{D}_i$. Let g an explanation function, that, given a black box predictor $f \in \mathcal{F}$, and an input of interest $x \in \mathcal{X}$ provide relevance scores of each input features of the predicted class such as $g(f_{\mathcal{D}_i}, x) = \phi_x^{\mathcal{D}_i} \in \mathbb{R}^d$. We assume $d : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^+$ a measure of similarity over explanations. Finally, The following Boolean connectives are used: $\neg$ denotes negation, $\wedge$ denotes conjunction, and $\oplus$ denotes exclusive or (XOR).

3.2. Motivation

From the definitions previously given for **Representativity** (Definition 4) and **Consistency** (Definition 5), we can give the motivations for their associated measures.

Consistency. The same explanation associated with two predictions y and $\neg\, y$ is said to be inconsistent. An explanation can then be defined as inconsistent if two models $f^{\mathcal{D}_i}, f^{\mathcal{D}_j}$, trained on the same task, give the same explanation but different predictions.

$$f^{\mathcal{D}_i}(x) \neq f^{\mathcal{D}_j}(x) \implies \phi_x^{\mathcal{D}_i} \neq \phi_x^{\mathcal{D}_j} \tag{3}$$

Representativity. If a sample of interest is removed from the training dataset, the explanation of the model for this sample should remain similar. One wishes to avoid that the explanations of a model depend too much on a single sample. Following this definition, for two models $f^{\mathcal{D}_i}, f^{\mathcal{D}_j}$, and x a training sample from one of the two models : $x \in \mathcal{D}_i, x \notin \mathcal{D}_j$. The explanations should remain similar conditioned by whether or not this sample belongs to the training dataset.

$$f^{\mathcal{D}_i}(x) = f^{\mathcal{D}_j}(x) \implies \phi_x^{\mathcal{D}_i} = \phi_x^{\mathcal{D}_j} \tag{4}$$

3.3. k-Fold Cross-Training

The goal of this training procedure illustrated in the Figure 1 is to set up a context allowing to measure the previously explained motivations. We will build two multisets : $\boldsymbol{\mathcal{S}}^{=}$ grouping the distances between explanations associated with the same prediction and $\boldsymbol{\mathcal{S}}^{\neq}$ the distances between explanations associated with different predictions.

We start by partitioning the original dataset into k independant blocks of equal size $\mathcal{D} = \{\mathcal{B}_i\}_{i=0}^{k}$. Several models $f^{\mathcal{D}_1}, \ldots, f^{\mathcal{D}_k}$ of the same architecture are trained on different coalitions of $k-1$ blocks such as $\mathcal{D}_i = \mathcal{D} \setminus \{\mathcal{B}_i\}$. We assume that the models have comparable performances. In our experiments we ensure a similar accuracy on the test set of each model.

We will now measure the distances between two explanations coming from different models. More specifically, measuring the distance between explanations coming from two models that have been trained on x doesn't have much interest (both models may have overfit and thus give the same explanation). We are only interested in the distances between two explanations $\phi_x^{\mathcal{D}_i}, \phi_x^{\mathcal{D}_j}$ coming from two models $f^{\mathcal{D}_i}, f^{\mathcal{D}_j}$, with only one that has been trained on the point of interest $x \in \mathcal{D}_i, x \notin \mathcal{D}_j$.

In the case where both models gave a good prediction, a small distance beetween two explanations means that the model build its explanations from several samples. Hence, the fact of having or removing a particular sample does not make these explanations vary, which is a sign of good **Representativity**. In the case where one of the two models give contrary predictions, we want to avoid that they give the same explanation. Indeed the **Consistency** of the explanations means that we cannot justify with the same explanation two different outcomes.

Hence, the distances are separated into two multisets, $\boldsymbol{\mathcal{S}}^{=}$ when the models have made good predictions where we want a small distance between the explanations, $\boldsymbol{\mathcal{S}}^{\neq}$ when one of the models gives a wrong prediction where we want a high distance between the two explanations. The case where both models give a bad prediction is ignored (for details, see the Algorithm 1 in the appendix).

$$\forall (x, y) \in \mathcal{D},\ \forall \mathcal{D}_i : x \in \mathcal{D}_i,\ \forall \mathcal{D}_j : x \notin \mathcal{D}_j$$

$$\boldsymbol{\mathcal{S}}^{=} = \{d(\phi_x^{\mathcal{D}_i}, \phi_x^{\mathcal{D}_j}) \mid f_{\mathcal{D}_i}(x) = y \wedge f_{\mathcal{D}_j}(x) = y\} \tag{5}$$

$$\boldsymbol{\mathcal{S}}^{\neq} = \{d(\phi_x^{\mathcal{D}_i}, \phi_x^{\mathcal{D}_j}) \mid f_{\mathcal{D}_i}(x) = y \oplus f_{\mathcal{D}_j}(x) = y\} \tag{6}$$

3.4. Relative Consistency : $ReCo$

From the Definition 5, and the Equation 3, explanations coming from differents models are said to be consistent if they are closer for the same prediction than for opposite predictions. As a reminder, the distances between explanations of the same predictions are represented by $\boldsymbol{\mathcal{S}}^{=}$, and the one associated to opposite predictions by $\boldsymbol{\mathcal{S}}^{\neq}$. Visually, we seek to maximize the shift between the histograms of the sets $\boldsymbol{\mathcal{S}}^{=}$ and $\boldsymbol{\mathcal{S}}^{\neq}$. Formally, we are looking for a distance value that separates $\boldsymbol{\mathcal{S}}^{=}$ and $\boldsymbol{\mathcal{S}}^{\neq}$, e.g. such that all the lower distances belong to $\boldsymbol{\mathcal{S}}^{=}$ and the higher ones to $\boldsymbol{\mathcal{S}}^{\neq}$. The clearer the separation, the more consistent the explanations. In order to find this separation, we introduce $ReCo$, a statistical measure based on maximizing balanced accuracy.

Where $\boldsymbol{\mathcal{S}} = \boldsymbol{\mathcal{S}}^{=} \cup \boldsymbol{\mathcal{S}}^{\neq}$ and $\gamma \in \boldsymbol{\mathcal{S}}$ a fixed threshold value, we can define the true positive rate (TPR), the true negative rate (TNR) and $ReCo$ as follows:

$$TPR(\gamma) = \frac{|\{d \in \boldsymbol{\mathcal{S}}^{=} \mid d < \gamma\}|}{|\{d \in \boldsymbol{\mathcal{S}} \mid d < \gamma\}|}$$

$$TNR(\gamma) = \frac{|\{d \in \boldsymbol{\mathcal{S}}^{\neq} \mid d > \gamma\}|}{|\{d \in \boldsymbol{\mathcal{S}} \mid d > \gamma\}|}$$

$$ReCo = \max_{\gamma}\ TPR(\gamma) + TNR(\gamma) - 1 \tag{7}$$

The score 1 indicating consistency of the models explanations, 0 indicating a total inconsistency.

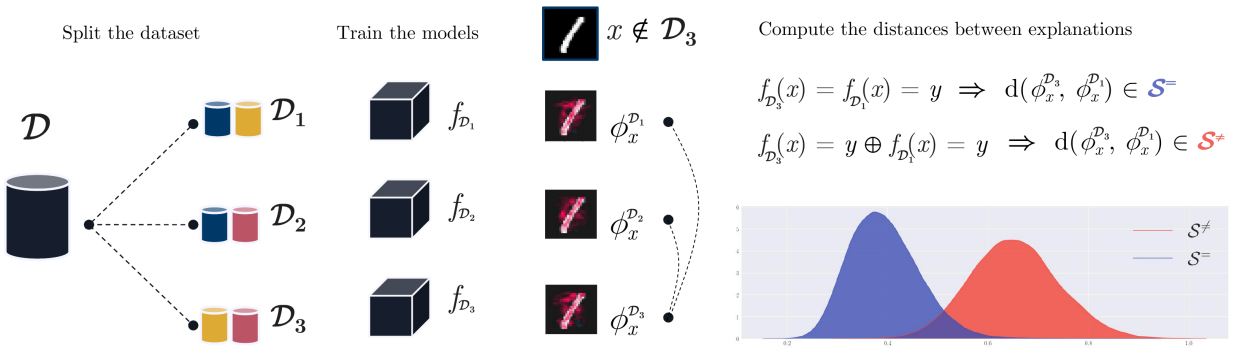

Figure 1: Example of application of the proposed procedure, with $\mathcal{D}$ a dataset and the number of splits $k = 3$, each of the models is trained on two of the three splits. For a given sample x such that $x \notin \mathcal{D}_3$, the explanations for each model are calculated ($\phi_x^{\mathcal{D}_1}, \phi_x^{\mathcal{D}_2}, \phi_x^{\mathcal{D}_3}$). The distances between the explanation of the model not trained on x : $\phi_x^{\mathcal{D}_3}$ and the two others : $\phi_x^{\mathcal{D}_1}, \phi_x^{\mathcal{D}_2}$ are computed. For each distance, if the predictions of both models are correct, the distances is added to $\boldsymbol{\mathcal{S}}^{=}$, if one of the two models makes a false prediction, the distance is added to $\boldsymbol{\mathcal{S}}^{\neq}$.

3.5. Mean Generalizability : $MeGe$

From the Definition 4, and the Equation 4, the explanations for samples are representative if they remain similar conditioned by whether or not those samples belongs to the training dataset. As a results, the distances between explanations coming from models where one has been trained on the point of interest and the other has not should be small. As those distances are contained in $\boldsymbol{\mathcal{S}}^{=}$, one way to measure the **Representativity** property is to compute the average of $\boldsymbol{\mathcal{S}}^{=}$. We want a high generalization score when distances are small, so we define the $MeGe$ measure as similarity.

$$MeGe = \frac{1}{1 + \sum_{d \in \boldsymbol{\mathcal{S}}^{=}} \frac{d}{|\boldsymbol{\mathcal{S}}^{=}|}} \tag{8}$$

Models with good **Representativity** capacity will therefore have high similarity between explanations and a score close to 1.

4. Experiments

In order to assess the measures, we have compared the $ReCo$ and $MeGe$ scores between a set of correctly trained models and degraded models. We compare the results obtained on Cifar-10 with those of **Fidelity** and **Stability** measures. Then we extended the experiments by conducting an application of the measures on 1-Lipsichtz networks.

For all experiments, we used 5 splits ($k = 5$), i.e. 5 models, with comparable accuracy ($\pm 3\%$). The models are based on a ResNet-18 architecture [13], adapted according to each dataset, see appendix A.3 for details on each model.

4.1. Explanation methods

In order to produce the necessary explanations for the experiment, we used several methods of explanation that we will briefly describe. However, as the aim of this experiment was not to exhaustively test all the available explanatory methods, we have limited the list to five regularly mentioned methods. This list was extended for the Cifar10 dataset by adding SHAP [21] (which requires significant computational power).

- **Saliency Map (SM)** [36] is a visualization techniques based on the gradient of a class score relative to the input, indicating in an infinitesimal neighborhood, which pixels must be modified to most affect the score of the class of interest.
- **Gradient $\odot$ Input (GI)** [3] is based on the gradient of a class score relative to the input, element-wise with the input, it was introduced to improve the sharpness of the attribution maps.
- **Integrated Gradients (IG)** [41] consists of summing the gradient values along the path from a baseline state to the current value. The baseline is defined by the user and often chosen to be zero. This integral can be approximated with a set of m points at regular intervals between the baseline and the point of interest. In order to approximate from a finite number of steps, the implementation here use the Trapezoidal rule and not a left-Riemann summation, which allows for more accurate results and improved performance (see [38] for a comparison). The final result depends on both the choice of the baseline and the number of points to estimate the integral. In the context of these experiments, we use zero as the baseline and $m = 60$.
- **SmoothGrad (SG)** [37] is also a gradient-based explanation method, which, as the name suggests, averages the gradient at several points corresponding to small perturbations (drawn i.i.d from a normal distribution of standard

deviation σ) around the point of interest. The smoothing effect induced by the average help reducing the visual noise, and hence improve the explanations. In practice, Smoothgrad is obtained by averaging after sampling m points. In the context of these experiments, we took $m = 60$ and $\sigma = 0.2$ as suggested in the original paper.

- **Grad-CAM (GC)** [32] can be used on Convolutional Neural Network (CNN), it uses the gradient and the feature maps of the last convolution layer. More precisely, the gradient is used to calculate a specific weight for each feature map, and by taking the positive part after averaging the ponderated feature maps.

4.2. Datasets

We applied the procedure and evaluated the measures for each of the degradations on four image classification datasets:

Fashion MNIST [45]: a dataset containing 70,000 low-resolution (28x28) grayscale images labeled in 10 categories.

EuroSAT [14]: a labeled dataset with 10 classes consisting of 27,000 colour images (64x64) from the Sentinel-2 satellite.

Cifar10 & Cifar100 [16] : two low-resolution labeled datasets with 10 and 100 classes respectively, consisting of 60,000 (32x32) colour images.

4.3. Distance over Explanations

The procedure introduced in 3.3 requires to define a distance between two explanations from the same sample. Since the proper interpretation method is to rank the features most sensitive to the model's decision, it seems natural to consider the Spearman rank correlation [39] to compare the similarity between explanations. Several works provide theoretical and experimental arguments in line with this choice [11, 1, 43]. However, it is important to note that the visual similarity problem is still on open problem. We conduct two sanity check on several candidates distances to ensure they could respond to the problem. The distances tested are : 1-Wasserstein distance (the Earth mover distance from [9]) , Sørensen–Dice [7] coefficient, Spearman rank correlation, SSIM [51], ℓ_1 and ℓ_2 norms.

4.3.1 Spatial correlation

The first test concerns the spatial distance between two areas of interest for an explanation. It is desired that the spatial distance between areas of interest be expressed by the distance used. As a results, two different but spatially close explanations should have a low distance. The test consists in generating several masks representing a point of interest, starting from a left corner of an image of size (32 x 32) and moving towards the right corner by interpolating 100 different masks. The distance between the first image and each interpolation is then measured (see Figure 2).

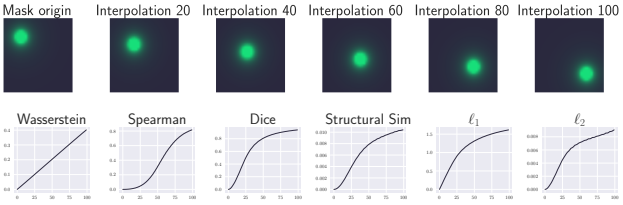

Figure 2: Distances with moving interest point. The first line shows the successive interpolations between the baseline image (left), and the target image (right). The second line shows the evolution of the distance between each interpolation and the baseline image.

The different distances evaluated pass this sanity check, i.e. a monotonous growth of the distance, image of the spatial distance of the two points of interest.

4.3.2 Noise test

The second test concerns the progressive addition of noise. It is desired that the progressive addition of noise to an original image will affect the distance between the original noise-free image and the noisy image. Formally, with x the original image, and $\varepsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ an isotropic Gaussian noise, we wish the distance d to show a monotonic positive correlation $\mathrm{corr}(dist(x, x + \varepsilon), \varepsilon)$.

In order to validate this prerogative, a Gaussian noise with a progressive intensity σ is added to an original image, and the distance between each of the noisy images and the original image is measured. For each value of σ the operation is repeated 50 times.

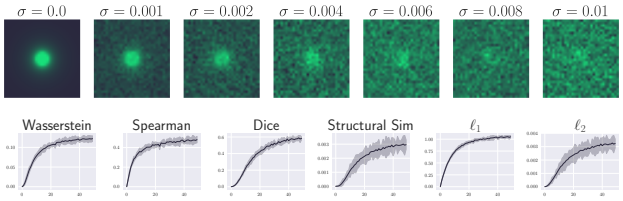

Figure 3: Distances with noisy images. The firs The first line shows original noise-free image (left) and noisy copies computed by increasing σ. The second line shows the distances between each noisy image and the baseline image.

Over the different distances tested, they all pass the sanity test : there is a monotonous positive correlation (as seen in Figure 3). Although SSIM and ℓ_2 have a higher variance.

One will nevertheless note the instability of the Dice score in cases where the areas of interest have a low surface area,

as well as a significant computation cost for the Wasserstein distance. For all these reasons, we chose to stay in line with previous work using the absolute value of Spearman rank correlation.

4.4. Validation Setup

For measures assessments, for each datasets, we applied three different types of degradations on the models : randomization of weights, label inversion, and training data limitation, with several degrees for each of them.

- Randomizing the weights, inspired by [1]. We gradually randomize 5%, 10% and 30% of the model layers by adding a Gaussian noise. By destroying the weights learned by the network, we expect to find degradation of explanations.
- Inversion of labels, inspired by [24, 1] the models are trained on a dataset with 5%, 10% and 30% of bad labels. By artificially breaking the relationship between the labels, we expect the explanations to lose their consistency.
- Limited data, where the models are trained on 75%, 50% and 25% of the available dataset. By decreasing the training data, we expect the model to over-fit at certain points, and lose generalization.

4.5. Application Setup

For measures application, we extend the experience on the Cifar10 dataset by adding a family of 1-Lipschitz models. Indeed different works mention the Lipschitz constrained networks as particularly robust [44, 25] and have good generalizability. As a reminder, a f function is called k-Lipschitz, with $k \in \mathbb{R}^+$ if

$$\|f(x_1) - f(x_2)\| \leq k\|x_1 - x_2\| \; \forall x_1, x_2 \in \mathbb{R}^n$$

The smallest of these k is called the Lipschitz constant of f. This constant certifies that the gradients of the function represented by the deep neural network are bounded (given a norm) and that this bound is known. This robustness certificate proves to be a good way to avoid gradient explosion, which is a problem with gradient-based explicability methods. The models were trained using the Deel-Lip library [34]. To our knowledge, no previous work has made the link between Lipschitz networks and the chosen interpretability methods.

5. Results

The purpose of this section is to report and synthesize the main results (for a complete detail, refer to the appendices A.4). They are essentially three observations, the first concerns the variation in the $MeGe$ and $ReCo$ scores according to the different explanation methods used. The second observation concerns the differences in scores obtained between the normally trained and degraded models. Finally the importance of the model-method couple, where we observe that certain methods are better suited to families of models, with a tendency emerging for the Grad-CAM method.

5.1. Methods ranking

5.1.1 Consistency and Representativity

Table 1 reports the $ReCo$ scores obtained for the ResNet-18 models trained without degradations on the different studied datasets. We can observe a clear difference between the methods. Grad-CAM appears robust to the datasets tested obtaining the best consistency score on each of them, followed by SmoothGrad on the Cifar datasets, and by Integrated Gradients on EuroSAT and FashionMNIST. This corroborates the observations of previous work [28, 32].

Dataset	IG	SG	SA	GI	GC
Cifar10	0.107	0.154	0.151	0.088	**0.637**
Cifar100	0.018	0.132	0.131	0.004	**0.800**
EuroSAT	0.309	0.182	0.177	0.241	**0.591**
FashionMNIST	0.369	0.125	0.1	0.369	**0.517**

Table 1: $ReCo$ score for ResNet-18 models normally trained. Higher is better.

Table 2 reports the $MeGe$ score results obtained for normally trained ResNet-18 models. Grad-CAM obtains the best **Representativity** score, except on Fashion-MNIST. Indeed, two methods tested here involve the element-wise multiplication of the explanation with the input: Integrated Gradients (IG) and Gradient Input (GI). On samples from Fashion-MNIST, multiplication of the explanation with the input eliminates the attribution score on a part of the image, thus reducing the distance between the two explanations.

Dataset	IG	SG	SA	GI	GC
Cifar10	0.584	0.457	0.449	0.552	**0.723**
Cifar100	0.595	0.499	0.495	0.571	**0.777**
EuroSAT	0.404	0.415	0.41	0.412	**0.667**
FashionMNIST	0.904	0.362	0.304	**0.906**	0.765

Table 2: $MeGe$ score for ResNet-18 models normally trained. Higher is better.

5.1.2 Comparisons with Fidelity and Stability

Tables 3 report the scores for the **Fidelity** (μF, Equation 1) and **Stability** (S_{avg} Equation 2) measures. The score obtained is averaged over $10,000$ test samples, with 0 for baseline. For μF, the size of the $|S|$ subset is 15% of the

image. For S_{avg}, the radius r is 0.1 according to the distance ℓ_1 and the Spearman rank correlation is used as the distance between explanations D.

This empirical results reveal that methods based only on the gradient appear to be unfaithful. Grad-CAM is the method with the highest **Fidelity** score, hence the one that best reflects the evidences for the predictions. This results are in lign with the two metrics introduced, and confirm from previous quantitative results [32, 28, 46]. As one would expect, the local sampling used by SmoothGrad allow the method to have a good **Stability**.

Metrics	IG	SG	SA	GI	GC
μF	0.107	0.305	0.229	0.101	**0.907**
S_{avg}	2.112	0.024	2.389	2.579	**0.012**
$ReCo$	0.107	0.154	0.151	0.088	**0.637**
$MeGe$	0.584	0.457	0.449	0.552	**0.723**

Table 3: **Fidelity**, **Stability**, **Consistency** and **Representativity** score for ResNet-18 models on Cifar10. Higher μF, $MeGe$ and $ReCo$ is better, lower S_{avg} is better.

5.2. Models ranking

By using the $MeGe$ measure, we can compare for the same dataset the family of models giving the most general explanations. We notice Figure 4 a correlation between the $MeGe$ score and the degradation applied to the models, which supports the results of several works, notably [24]. This tends to show the link between the degradations of the models and their generalization capacity. Unsurprisingly, regardless of the method, the models normally formed on the Cifar10 dataset obtain the best results. Nevertheless, we notice that the variation of the score depends on the methods used, and that some of them seem more sensitive to model modifications, such as Grad-CAM, in agreement with previous works [1].

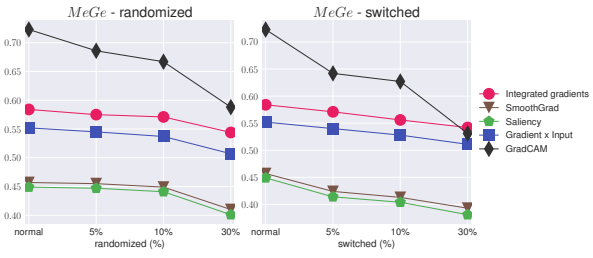

Figure 4: **Cifar10** $MeGe$ **scores** for normally trained models (first point from the left), as well as for progressively degraded models. On the left chart the weights are progressively randomized, on the right chart the models are trained with inverted labels. Higher is better.

5.3. Right Method for the Right Model

Table 4 show the $ReCo$ score obtained for ResNet-18 models and 1-Lipschitz models on Cifar10. All the models have comparable accuracy ($78 \pm 4\%$). We observe a large difference in the $ReCo$ scores, which is due to the difference in the $\mathcal{S}^{=}$ and $\mathcal{S}^{\neq}$ histograms obtained, as can be seen in Figure 5. On the left column, the results come from ResNet-18 models normally trained on Cifar10. We observe that Grad-CAM is the method that best allows to distinguish the two histograms (sign of **Consistency**), as well as the one with the lowest expectation of $\mathcal{S}^{=}$ (sign of **Representativity**), thus obtaining the best $ReCo$and $MeGe$ score among the different methods tested on the ResNet-18 models. On the right column, the results come from a Lipschitz model. We observe a clear improvement of the separation between the histograms, especially for the methods based exclusively on gradients. Indeed, SmoothGrad is the method obtaining the most consistent explanations as reported in Table 4, in front of Saliency and Grad-CAM. In addition to obtaining more consistent explanations, we observe that the explanations obtained are much less saturated than with the ResNet models, see Figure 6.

	IG	SG	SA	GI	GC	Shap
ResNet-18	0.107	0.154	0.151	0.088	**0.637**	0.387
1-Lipschitz	0.598	**0.898**	0.81	0.5	0.668	0.38

Table 4: $ReCo$ score obtained by 1-Lipschitz models and ResNet-18 models on Cifar10. Higher is better.

Concerning $MeGe$, the results reported in Table 5 show an improvement in the generalizability of the 1-Lipschitz models. Indeed, the **Representativity** score has increased compared to the ResNet models for all tested methods.

	IG	SG	SA	GI	GC	Shap
ResNet-18	0.584	0.457	0.449	0.552	**0.723**	0.459
1-Lipschitz	0.719	0.606	0.575	0.67	**0.749**	0.621

Table 5: $MeGe$ score obtained by 1-Lipschitz models and ResNet-18 models on Cifar10. Higher is better.

We note the importance of the association model-method, although some methods such as Grad-CAM seem robust to model changes and perform well on the different datasets. Lipschitz constrained networks seem to be tailored to be explained by methods based on gradients. These encouraging results show that there is a close link between the methods used and model architectures, as well as the usefulness of Lipschitz networks for explainability.

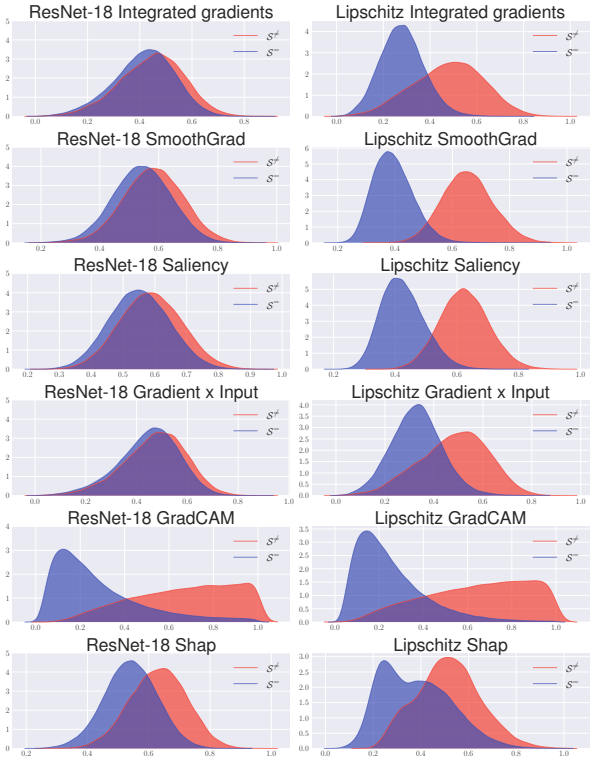

Figure 5: $\mathcal{S}^{=}$ and $\mathcal{S}^{\neq}$ for ResNet (left column) and 1-Lipschitz models (right column) on Cifar10. As explained in this paper, a clear separation between the $\mathcal{S}^{=}$ and $\mathcal{S}^{\neq}$ histograms is a sign of consistent explanations.

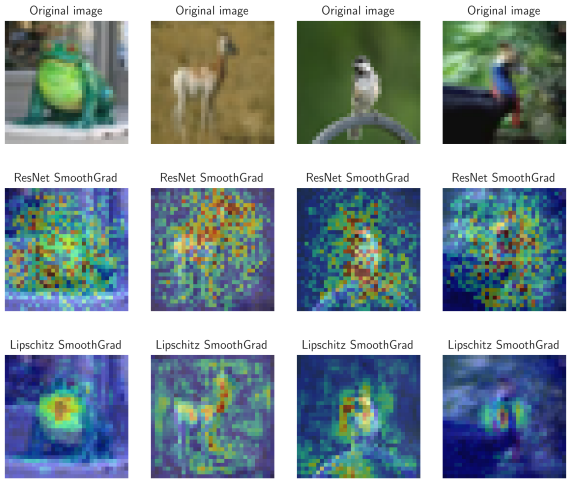

Figure 6: Example of results obtained using the Smoothgrad method, both models had a correct prediction, the middle column represents the heatmaps obtained after application of SmoothGrad on one of the 1-Lipschitz models, with the same parameters ($\sigma = 0.1, n = 200$) as for one of the ResNet-18 models (right column).

6. Conclusion

Research in the field of interpretable AI has been traditionally focused on the creation of explainability method, aimed at highlighting the major elements of decision-making. They are however only a first step towards adopting neural networks in safety-critical context. The second step, of crucial importance, is to objectively measure the explanations in order to validate the quality of the model.

This paper introduces two new measures measuring the **Consistency** ($ReCo$) and **Representativity** ($MeGe$) of explanations for a family of models. We used $MeGe$ to illustrate the link between the degradation of a model and the loss of generalizability, confirming previous works. We highlighted differences between explanations methods, and offer a new way of ordering them. Finally, the important consistency potential of 1-Lipschitz networks was quantified using $ReCo$, and we showed the usefulness of gradient-based methods coupled with 1-Lipschitz networks.

Although this work focuses on convolutional neural networks and explanability methods specific to their architecture, the procedure introduced here is voluntarily general and can be applied to large classes of models (such as Decision Tree, Rule-based, GAN...) in that it only requires to define a notion of distances between explanations.

We hope that this work will guide efforts in the search for measures of explainability towards crucial properties still understudied, in order to build better and more reliable models.

Acknowledgement

This work has been realised in the frame of the DEEL project[1] It received funding from the French Investing for the Future PIA3 program within the Artificial and Natural Intelligence Toulouse Institute (ANITI). We thank Mélanie Ducoffe and Mikaël Capelle of the DEEL team for critical feedback that helped improved the work.

References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems (NIPS)*, 2018.

[2] David Alvarez-Melis and Tommi S. Jaakkola. Towards robust interpretability with self-explaining neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2018.

[3] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. In *Proceedings of*

[1] https://www.deel.ai/

the International Conference on Learning Representations (ICLR), 2018.

[4] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *Public Library of Science (PloS One)*, 2015.

[5] Umang Bhatt, Adrian Weller, and José M. F. Moura. Evaluating and aggregating feature-based model explanations. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2020.

[6] Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L. Boyd-graber, and David M. Blei. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems (NIPS)*, 2009.

[7] Lee R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 1945.

[8] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *ArXiv e-print*, 2017.

[9] Rémi Flamary and Nicolas Courty. Pot python optimal transport library, 2017.

[10] Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[11] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2017.

[12] Leilani H. Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *Proceedings of the IEEE International Conference on data science and advanced analytics (DSAA)*, 2018.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[14] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. In *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (JSTARS)*, 2019.

[15] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2019.

[16] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images, 2009.

[17] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, and Finale Doshi-Velez. An evaluation of the human-interpretability of explanation. In *Workshop on Correcting and Critiquing Trends in Machine Learning, Advances in Neural Information Processing Systems (NIPS)*, 2019.

[18] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 2015.

[19] Jiwei Li, Will Monroe, and Dan Jurafsky. Understanding neural networks through representation erasure, 2016.

[20] Zachary C. Lipton. The mythos of model interpretability. In *Workshop on Human Interpretability in Machine Learning, Proceedings of the International Conference on Machine Learning (ICML)*, 2016.

[21] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[22] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, February 2019.

[23] Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation, 2018.

[24] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nathan Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[25] Patricia Pauli, Anne Koch, Julian Berberich, and Frank Allgöwer. Training robust neural networks using lipschitz bounds, 2020.

[26] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Knowledge Discovery and Data Mining (KDD)*, 2016.

[27] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[28] Laura Rieger and Lars Kai Hansen. Irof: a low resource evaluation metric for explanation methods. In *Workshop, Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

[29] Marko Robnik-Šikonja and Marko Bohanec. Perturbation-based explanations of prediction models. In *Human and machine learningSpringer International Publishing*, 2018.

[30] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Bach, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. In *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, 2015.

[31] Philipp Schmidt and Felix Biessmann. Quantifying interpretability and trust in machine learning systems. In *Workshop on Network Interpretability for Deep Learning, Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

[32] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[33] Thomas Serre. Deep learning: The good, the bad, and the ugly. *Annual review of vision science*, 2019.

[34] Mathieu Serrurier, Franck Mamalet, Alberto González-Sanz, Thibaut Boissin, Jean-Michel Loubes, and Eustasio del Barrio. Achieving robustness in classification using optimal transport with hinge regularization, 2020.

[35] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2017.

[36] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Workshop, Proceedings of the International Conference on Learning Representations (ICLR)*, 2013.

[37] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. In *Workshop on Visualization for Deep Learning, Proceedings of the International Conference on Machine Learning (ICML)*, 2017.

[38] Matthew Sotoudeh and Aditya V. Thakur. Computing linear restrictions of neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2019.

[39] Charles Spearman. The proof and measurement of association between two things. *American Journal of Psychology*, 1904.

[40] Pascal Sturmfels, Scott Lundberg, and Su-In Lee. Visualizing the impact of feature attribution baselines. *Distill*, 2020.

[41] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2017.

[42] Nava Tintarev and Judith Masthoff. A survey of explanations in recommender systems. *Workshop on Recommender Systems and Intelligent User InterfacesIEEE International Conference Data Engineering (ICDE)*, 2007.

[43] Richard Tomsett, Dan Harborne, Supriyo Chakraborty, Prudhvi Gurram, and Alun Preece. Sanity checks for saliency metrics. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

[44] Muhammad Usama and Dong Eui Chang. Towards robust neural networks with lipschitz continuity. In *Digital Forensics and Watermarking, Springer International Publishing*, 2018.

[45] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

[46] Mengjiao Yang and Been Kim. Benchmarking attribution methods with relative feature importance, 2019.

[47] Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Sai Suggala, David I. Inouye, and Pradeep Ravikumar. On the (in)fidelity and sensitivity for explanations. In *Advances in Neural Information Processing Systems (NIPS)*, 2019.

[48] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, 2014.

[49] M. D. Zeiler, G. W. Taylor, and R. Fergus. Adaptive deconvolutional networks for mid and high level feature learning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2011.

[50] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene cnns. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016.

[51] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 2004.

[52] Luisa M Zintgraf, Taco S Cohen, Tameem Adel, and Max Welling. Visualizing deep neural network decisions: Prediction difference analysis. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

A. Appendix

A.1. Method Details

Algorithm 1 Training procedure to compute $\boldsymbol{\mathcal{S}}^{=}$ and $\boldsymbol{\mathcal{S}}^{\neq}$

Require: $k \in \mathbb{N}_{\geqslant 2}$, $\mathcal{D} = \{\mathcal{B}_i\}_{i=0}^{k}$
  $\boldsymbol{\mathcal{S}}^{=} \leftarrow \{\}, \boldsymbol{\mathcal{S}}^{\neq} \leftarrow \{\}$
  for all $i \in \{1, \ \ldots, \ k\}$ **do**
    $\mathcal{D}_i = \mathcal{D} \setminus \{\mathcal{B}_i\}$
    Train $f_{\mathcal{D}_i}$ on $\mathcal{D}_i$
    for all $(x, y) \in \mathcal{D}$ **do**
      // generate explanations on all dataset
      $\phi_x^{\mathcal{D}_i} \leftarrow g(f_{\mathcal{D}_i}, x)$
    end for
  end for
  for all $(x, y) \in \mathcal{D}$ **do**
    for all $i \in \{1, \ \ldots, \ k \mid x \in \mathcal{D}_i\}$ **do**
      for all $j \in \{1, \ \ldots, \ k \mid x \notin \mathcal{D}_j\}$ **do**
        // $f_{\mathcal{D}_i}$ was trained on x $f_{\mathcal{D}_j}$ was not
        $\Delta_x^{ij} \leftarrow d(\phi_x^{\mathcal{D}_i}, \phi_x^{\mathcal{D}_j})$
        if $f_{\mathcal{D}_i}(x) = y$ **and** $f_{\mathcal{D}_j}(x) = y$ **then**
          // both model are correct
          $\boldsymbol{\mathcal{S}}^{=} \leftarrow \boldsymbol{\mathcal{S}}^{=} \cup \{\Delta_x^{ij}\}$
        else if $f_{\mathcal{D}_i}(x) = y$ **or** $f_{\mathcal{D}_j}(x) = y$ **then**
          // only one model is correct
          $\boldsymbol{\mathcal{S}}^{\neq} \leftarrow \boldsymbol{\mathcal{S}}^{\neq} \cup \{\Delta_x^{ij}\}$
        else
          // pass
        end if
      end for
    end for
  end for
  return $\boldsymbol{\mathcal{S}}^{=}, \boldsymbol{\mathcal{S}}^{\neq}$

A.2. Considered measures for ReCo

As mentioned in when introducing $ReCo$, one would be tempted to use directly a distance between distributions, we briefly explain why we did not make this choice. In addition, we detail an alternative measure, also based on balanced accuracy, which gives consistent results.

A first intuition to measure the shift between the $\boldsymbol{\mathcal{S}}^{=}$ and $\boldsymbol{\mathcal{S}}^{\neq}$ histograms would be to consider the usual measures, such as Kullback-Leibler (KL) divergence :

$$KL(\mathcal{S}^{=} \parallel \mathcal{S}^{\neq}) = \sum_{x \in \mathcal{S}} \mathcal{S}^{=}(x) \log \left(\frac{\mathcal{S}^{=}(x)}{\mathcal{S}^{\neq}(x)} \right)$$

or the 1-Wasserstein measure (W_1) :

$$W_1(\mathcal{S}^{=}, \mathcal{S}^{\neq}) = \inf_{\gamma \in \Gamma(\mathcal{S}^{=}, \mathcal{S}^{\neq})} \mathbb{E}_{(x,y) \sim \gamma} \big[d(x, y) \big]$$

However, these distances are problematic in that the order of the distributions actually matters more than the distance between them, and these two measures can give a good score even when the explanations are inconsistent : where $\mathbb{E}[\mathcal{S}^{=}] > \mathbb{E}[\mathcal{S}^{\neq}]$.

To expose this problem, let us consider the case where the explanations given by the family of models are consistent, that is $(\mu_1, \mu_2) \in [0, 1]^2$, $\mu_1 < \mu_2$, $\mathcal{S}_1^{=} \sim \mathcal{N}(\mu_1, \sigma_1)$, $\mathcal{S}_1^{\neq} \sim \mathcal{N}(\mu_2, \sigma_2)$ we measure the following KL distance:

$$KL(\mathcal{S}_1^{=} \parallel \mathcal{S}_1^{\neq}) = \log \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2}$$

We can then construct an inconsistent example ($\mathbb{E}[\mathcal{S}^{=}] > \mathbb{E}[\mathcal{S}^{\neq}]$) with the same KL score: with $\mathcal{S}_2^{=} \sim \mathcal{N}(1 - \mu_1, \sigma_1)$ and $\mathcal{S}_1^{\neq} \sim \mathcal{N}(1 - \mu_2, \sigma_2)$ then we get:

$$\begin{aligned} KL(\mathcal{S}_2^{=} \parallel \mathcal{S}_2^{\neq}) &= \log \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (1 - \mu_1 - (1 - \mu_2))^2}{2\sigma_2^2} - \frac{1}{2} \\ &= \log \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (-\mu_1 + \mu_2)^2}{2\sigma_2^2} - \frac{1}{2} \\ &= KL(\mathcal{S}_1^{=}, \mathcal{S}_1^{\neq}) \end{aligned}$$

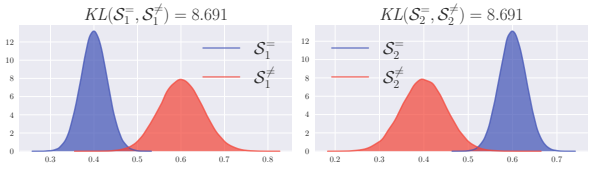

Figure 7: Toy example with on the left, a model family giving consistent explanations, and on the right a model family with inconsistent explanations and yet the same KL distance.

Similarly, considering the 1-Wasserstein measure, with δ the Dirac delta function, we begin in the case of consistent explanations, let $\mathcal{S}_1^{=}, \mathcal{S}_1^{\neq}$, two degenerate distributions $\mathcal{S}_1^{=} = \delta(\mu_1)$ and $\mathcal{S}_1^{\neq} = \delta(\mu_2)$ with $\mu_1 < \mu_2$, we then have :

$$W_1(\mathcal{S}_1^{=}, \mathcal{S}_1^{\neq}) = \|\mu_1 - \mu_2\|$$

we can then construct an inconsistent case by exploiting the invariance to the direction of transport, with $\mathcal{S}_2^{=} = \delta(\mu_2)$ and $\mathcal{S}_2^{\neq} = \delta(\mu_1)$.

$$\begin{aligned} W_1(\mathcal{S}_2^{=}, \mathcal{S}_2^{\neq}) &= \|\mu_2 - \mu_1\| \\ W_1(\mathcal{S}_2^{=}, \mathcal{S}_2^{\neq}) &= W_1(\mathcal{S}_1^{=}, \mathcal{S}_1^{\neq}) \end{aligned}$$

For these reasons, we have therefore chosen a classification measure, based on maximizing balanced accuracy.

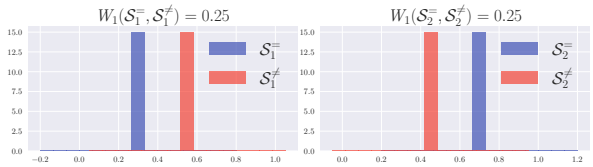

Figure 8: Toy example with on the left, a model family giving consistent explanations, and on the right a model family with inconsistent explanations and yet the same W_1 distance.

Nevertheless, one could also (observing similar results) use the area under the curve (AUC) of the balanced accuracy, such as :

$$ReCo_{AUC} = \int_{\gamma \in \mathcal{O}} TPR(\gamma) + TNR(\gamma) - 1 \; d\gamma$$

A.3. Models

As mentioned in the paper, the models used are all (with the exception of 1-Lipschitz networks) ResNet-18, with variations in size and number of filters used. Preserving the increase of filters at each depth by the original factor (x2), we took care to define for each dataset, a base filters value, as the number of filters for the first convolution layer. Another difference concerns the dropout rates used, indeed we have adapted the dropout to improve the performance of the tested models. Moreover, it should be remembered that there is no difference in architecture between the normally trained models and the degraded models.

We report here the architecture of the models for each of the datasets:

- **Fashion-MNIST** base filters 26, Dropout 0.4
- **EuroSAT** base filters 46, Dropout 0.25
- **Cifar100** base filters 32, Dropout 0.25
- **Cifar100** base filters 32, Dropout 0.25

A.3.1 Lipschitz models

The 1-Lipschitz models use spectral regularization on the Dense and Convolutions layers. The architecture is as described in Table 6.

A.4. Additional results

Table 6: 1-Lipschitz model architecture for Cifar10.

Conv2D(48)
PReLU
AvgPooling2D((2, 2))
Dropout(0.2)
Conv2D(96)
PReLU
AvgPooling2D((2, 2))
Dropout(0.2)
Conv2D(96)
AvgPooling2D((2, 2))
Flatten
Dense(10)

Table 7: $ReCo$ score for ResNet-18 models on Fashion-MNIST.

	Integrated gradients	SmoothGrad	Saliency	Gradient x Input	GradCAM
Normal	0.369	0.125	0.1	0.369	0.517
Limited 75%	0.258	0.006	0.001	0.253	0.508
Limited 50%	0.259	0.006	0.003	0.257	0.454
Limited 25%	0.084	0.0	0.0	0.078	0.435
Randomized 5%	0.26	0.019	0.001	0.259	0.365
Randomized 10%	0.245	0.0	0.0	0.244	0.212
Randomized 30%	0.029	0.0	0.0	0.028	0.0
Switched 5%	0.203	0.031	0.02	0.202	0.226
Switched 10%	0.199	0.056	0.031	0.197	0.144
Switched 30%	0.17	0.08	0.063	0.17	0.003

Table 8: $MeGe$ score for ResNet-18 models on Fashion-MNIST.

	Integrated gradients	SmoothGrad	Saliency	Gradient x Input	GradCAM
Normal	0.904	0.362	0.304	0.906	0.765
Limited 75%	0.891	0.354	0.285	0.893	0.752
Limited 50%	0.89	0.304	0.247	0.892	0.739
Limited 25%	0.871	0.292	0.221	0.872	0.727
Randomized 5%	0.889	0.331	0.271	0.89	0.647
Randomized 10%	0.88	0.296	0.234	0.879	0.507
Randomized 30%	0.887	0.289	0.218	0.887	0.32
Switched 5%	0.888	0.301	0.247	0.89	0.548
Switched 10%	0.888	0.294	0.226	0.889	0.485
Switched 30%	0.89	0.347	0.278	0.892	0.258

Table 9: $ReCo$ score for ResNet-18 models on Cifar10.

	Integrated gradients	SmoothGrad	Saliency	Gradient x Input	GradCAM	Shapley
Normal	0.107	0.154	0.151	0.088	0.637	0.387
Limited 75%	0.073	0.12	0.117	0.052	0.588	0.338
Limited 50%	0.063	0.117	0.115	0.056	0.57	0.301
Limited 25%	0.056	0.106	0.106	0.048	0.482	0.266
Randomized 5%	0.075	0.115	0.11	0.056	0.568	0.322
Randomized 10%	0.086	0.109	0.106	0.067	0.548	0.309
Randomized 30%	0.038	0.04	0.037	0.016	0.422	0.208
Switched 5%	0.03	0.071	0.068	0.024	0.538	0.234
Switched 10%	0.024	0.051	0.049	0.018	0.531	0.182
Switched 30%	0.0	0.004	0.003	0.0	0.369	0.072

Table 10: $MeGe$ score for ResNet-18 models on Cifar10.

	Integrated gradients	SmoothGrad	Saliency	Gradient x Input	GradCAM	Shap
Normal	0.584	0.457	0.449	0.552	0.723	0.459
Limited 75%	0.574	0.445	0.435	0.548	0.698	0.437
Limited 50%	0.576	0.461	0.451	0.555	0.697	0.449
Limited 25%	0.556	0.445	0.435	0.542	0.641	0.423
Randomized 5%	0.575	0.455	0.447	0.545	0.686	0.451
Randomized 10%	0.571	0.449	0.441	0.537	0.667	0.443
Randomized 30%	0.544	0.41	0.401	0.507	0.588	0.397
Switched 5%	0.571	0.424	0.414	0.54	0.642	0.394
Switched 10%	0.556	0.413	0.404	0.528	0.627	0.378
Switched 30%	0.542	0.393	0.381	0.511	0.531	0.343

Table 11: $ReCo$ score for ResNet-18 models on Cifar100.

	Integrated gradients	SmoothGrad	Saliency	Gradient x Input	GradCAM
Normal	0.018	0.132	0.131	0.004	0.8
Limited 75%	0.0	0.126	0.124	0.0	0.799
Limited 50%	0.0	0.111	0.108	0.0	0.773
Limited 25%	0.0	0.112	0.11	0.0	0.722
Randomized 5%	0.002	0.111	0.107	0.002	0.694
Randomized 10%	0.004	0.133	0.128	0.005	0.672
Randomized 30%	0.015	0.038	0.043	0.033	0.33
Switched 5%	0.0	0.103	0.1	0.0	0.762
Switched 10%	0.0	0.095	0.093	0.0	0.744
Switched 30%	0.0	0.114	0.114	0.0	0.744

Table 12: $MeGe$ score for ResNet-18 models on Cifar100.

	Integrated gradients	SmoothGrad	Saliency	Gradient x Input	GradCAM
Normal	0.595	0.499	0.495	0.571	0.777
Limited 75%	0.597	0.509	0.505	0.577	0.762
Limited 50%	0.599	0.513	0.508	0.58	0.737
Limited 25%	0.58	0.504	0.497	0.569	0.706
Randomized 5%	0.573	0.483	0.478	0.552	0.688
Randomized 10%	0.56	0.466	0.461	0.539	0.472
Randomized 30%	0.499	0.347	0.342	0.493	0.039
Switched 5%	0.564	0.464	0.459	0.545	0.715
Switched 10%	0.558	0.455	0.45	0.536	0.685
Switched 30%	0.54	0.437	0.433	0.521	0.58

Table 13: $ReCo$ score for ResNet-18 models on EuroSAT.

	Integrated gradients	SmoothGrad	Saliency	Gradient x Input	GradCAM
Normal	0.309	0.182	0.177	0.241	0.591
Limited 75%	0.25	0.1	0.097	0.142	0.56
Limited 50%	0.192	0.105	0.11	0.111	0.538
Limited 25%	0.133	0.077	0.075	0.075	0.535
Randomized 5%	0.284	0.213	0.216	0.273	0.497
Randomized 10%	0.242	0.209	0.21	0.255	0.294
Randomized 30%	0.21	0.07	0.07	0.216	0.207
Switched 5%	0.058	0.048	0.049	0.058	0.238
Switched 10%	0.042	0.011	0.01	0.013	0.218
Switched 30%	0.008	0.002	0.002	0.0	0.141

Table 14: $MeGe$ score for ResNet-18 models on EuroSAT.

	Integrated gradients	SmoothGrad	Saliency	Gradient x Input	GradCAM
Normal	0.404	0.415	0.41	0.412	0.667
Limited 75%	0.411	0.415	0.411	0.411	0.65
Limited 50%	0.438	0.433	0.427	0.431	0.648
Limited 25%	0.419	0.412	0.41	0.413	0.64
Randomized 5%	0.396	0.385	0.381	0.385	0.587
Randomized 10%	0.366	0.36	0.356	0.356	0.457
Randomized 30%	0.351	0.295	0.293	0.319	0.357
Switched 5%	0.326	0.347	0.345	0.344	0.382
Switched 10%	0.323	0.333	0.331	0.332	0.356
Switched 30%	0.315	0.31	0.307	0.31	0.3