How Good is your Explanation? Algorithmic Stability Measures to Assess the Quality of Explanations for Deep Neural Networks

Thomas Fel\textsuperscript{1,2,3}, David Vigouroux\textsuperscript{3}, Rémi Cadène\textsuperscript{1}, and Thomas Serre\textsuperscript{1,2}

\textsuperscript{1}Carney Institute for Brain Science, Brown University
\textsuperscript{2}Artificial and Natural Intelligence Toulouse Institute, Université de Toulouse, France
\textsuperscript{3}IRT Saint-Exupery

Abstract

A plethora of methods have been proposed to explain how deep neural networks reach their decisions but comparatively, little effort has been made to ensure that the explanations produced by these methods are objectively relevant. While several desirable properties for trustworthy explanations have been formulated, objective measures have been harder to derive. Here, we propose two new measures to evaluate explanations borrowed from the field of algorithmic stability: mean generalizability $\text{MeGe}$ and relative consistency $\text{ReCo}$. We conduct extensive experiments on different network architectures, common explainability methods, and several image datasets to demonstrate the benefits of the proposed measures. In comparison to ours, popular fidelity measures are not sufficient to guarantee trustworthy explanations. Finally, we found that 1-Lipschitz networks produce explanations with higher MeGe and ReCo than common neural networks while reaching similar accuracy. This suggests that 1-Lipschitz networks are a relevant direction towards predictors that are more explainable and trustworthy.

1. Introduction

Machine learning approaches such as deep neural networks have become essential in multiple domains such as image classification, language processing and speech recognition. These approaches have achieved excellent classification accuracy – approaching human performance in specific domains \cite{26, 44}. However, one significant drawback associated with these deep networks is that it is difficult to interpret their decisions \cite{29}. This problem constitutes a serious obstacle for the wide adoption of these systems for safety-critical applications.

Recently, several explainability methods have been proposed to help understand how predictors make a particular decision \cite{53, 49, 35, 43}. These methods are also meant to improve predictors and more importantly to build trust with people affected by their predictions. Unfortunately, these methods have strong limitations. In particular, they are subject to confirmation bias: while some methods appear to offer useful explanations to a human experimenter, they turn out not to reflect the actual behaviour of the predictor \cite{1, 18}. Instead of providing confidence in a system’s decisions, these explanations can themselves be potentially erroneous and cannot be trusted.

In this work, we focus on evaluation methods that objectively assess the quality and trustworthiness of explanations. Numerous methods have been proposed \cite{9, 55, 31, 39, 19, 2}. They are often derived from axioms and properties that good
explanations should possess. The most studied property is **Fidelity**. It is associated with a metric that evaluates an explanation method on its ability to reflect the internal decision process of a given neural network predictor. An important limitation of **Fidelity** is that it does not take into account the prediction capability of the predictor. For instance, an explainability method could have a high **Fidelity** given a random predictor. Thus, **Fidelity** is only a first step towards predictors that are explainable and can be trusted.

Instead, we propose to additionally take the prediction capability of the predictor into account through the **Generalizability** and **Consistency** of its explanations. A shown in Figure 1, we borrow these two concepts from the fields of generalization and algorithmic stability [7], and adapt them to the field of explainability. Intuitively, when some images of a certain class are well classified by a given predictor, trustworthy explanations should point out to the same evidence across all images of the same category. For instance, the black region near the eyes of a panda. That is captured by the notion of **Generalizability**. Conversely, when images of the same category are misclassified, trustworthy explanations should ideally point out to image evidence which differs from that used for correctly classified images. That is captured well by the notion of **Consistency**. In practice, we estimate with a k-fold cross-training approach the metrics associated with different predictors, i.e., their average generalizability (MeGe) and their relative consistency (ReCo).

We provide an extensive experimental validation of our approach using different neural networks, common explainability methods and several image datasets. We compare the average generalizability (MeGe) and relative consistency (ReCo) measures against the leading **Fidelity** measure [6, 60]. Importantly, we experimentally show that the **Fidelity** is not enough to ensure that a given pretrained model produces trustworthy explanations. Finally, we show that 1-Lipschitz networks obtain a higher average generalizability (MeGe) and relative consistency (ReCo) compared to common neural networks while reaching similar accuracies. This finding offers an interesting research track towards more explainable and trustworthy predictors. To summarize, our contributions include:

- **Generalizability** (MeGe) and **Consistency** (ReCo) practical measures to assess the quality and trustworthiness of explanations,
- Extensive experimental validation of our approach on several neural networks, explainability methods, and image datasets (including ImageNet),
- Empirical demonstration that 1-Lipschitz networks deliver explanations with higher MeGe and ReCo.

2. Related work

In this work, we focus on evaluating explanations provided by explainability methods, which give insight into how a given neural network architecture reaches a particular decision [14]. These explainability methods produce an influence score for each input dimension. In the case of image classification, these methods will produce heatmaps indicating the diagnosticity of individual image regions. Most of these explainability methods rely on backpropagating the gradient with respect to a given input image [62, 47, 4, 17, 46, 53, 49, 43, 20] or with respect to a perturbation of the input [61, 63, 36, 27, 65, 37, 54].

Despite a wide range of explainability methods, assessing the quality and trustworthiness of these explanations is still an open problem. It is in part due to the difficulty of obtaining objective ground truths [40, 28]. Several criteria have been proposed to evaluate the quality of explanations [55, 31, 39, 19, 2, 9, 15]. According to [9], the five major properties include: **Fidelity**, **Stability**, **Comprehensibility**, **Generalizability** and **Consistency**. Yet, properties such as **Generalizability** and **Consistency** do not come with a practical definition.

In order to measure these different properties, there are two main approaches currently used. The first subjective approach consists in putting the human at the heart of the process, either by explicitly asking for human feedback [43, 36, 30], or by indirectly measuring the performance of the human/classifier duo [25, 10, 32, 42]. Nevertheless, human intervention sometimes brings undesirable effects, including a possible confirmation bias [1].

A second type of approaches has also started to emerge specifically for computer vision applications. The main idea is to build objective proxy tasks that a good explanation must be able to solve. These measures aim to evaluate explanations based on two properties: **Fidelity** and **Stability**. The first method to measure **Fidelity** was first proposed in [40] based on estimating the drop in prediction score resulting from deleting pixels deemed important by an explanation method. To ensure that the drop in score does not come from a change in distribution, ROAR[23] was proposed which re-train a classifier model between each deletion step. This boils down to measuring the correlation between the attributions for each pixel and the difference in the prediction score when they are modified and has been clearly formalized [60, 6, 38]. Nevertheless, it should be noted that the different fidelity metrics proposed requires defining a baseline state which might favor explainability methods that internally relies on the same baseline [52].

Those **Fidelity** metrics are a first step toward trustworthy explanations: by making sure that we have faithful explanations, we can then look at other criteria to quantitatively measure these explanations. However, the different **Fidelity** metrics suffer from limitations demonstrated by [56] which proposes several sanity checks revealing their high variance, their sensitivity to the baseline and their low consistency.
3. Methods

Below, we briefly provide some motivation for the proposed MeGe and ReCo measures before describing a training procedure applicable to a large family of machine learning models in order to estimate these two values. One basic assumption for the proposed approach is borrowed from algorithmic stability and generalization: to be reliable, an explanation obtained for a specific predictor for a given image should be stable when the training dataset used to train the predictor is perturbed slightly as when, for instance, this particular image is added or removed from the dataset.

3.1. Notations

We consider a standard supervised learning setting where a datapoint is denoted \( z = (x, y) \) s.t. \( x \in X \) is an observation (e.g., \( X = \mathbb{R}^d \)) and \( y \in Y \) is a class label (e.g., \( Y = \mathbb{R}^p \)). The data set is denoted as \( D = \{z_1, \ldots, z_m\} \), we designate \( Y = \{V_1, \ldots, V_k\} \) the set of \( k \) disjoints subsets \( (folds) \) of size \( m/k \) at random where each \( V_i \subset D \). Throughout this work, we will assume \( k \) divides \( m \) for convenience. Let \( \mathcal{A} \) be a deterministic learning algorithm that maps any number of data points onto a function \( f \) from \( X \) to \( Y \). In particular, we consider the fold \( V_i \) and the associated predictor \( f_i = \mathcal{A}(V \setminus V_i) \).

An explanation method is a functional, denoted \( \Phi \), which, given a predictor \( f_i \) and a datapoint \( x \), assigns an importance score for each input dimension \( \phi_x = \Phi(f_i, x) \). We define a distance \( d(\cdot, \cdot) \) over the explanations. Finally, the following Boolean connectives are used: \( \neg \) denotes a negation, \( \wedge \) denotes a conjunction, and \( \oplus \) denotes an exclusive or (XOR).

3.2. Motivation

We first consider Generalizability: we provide a definition, discuss the inherent difficulties associated with its measurement, and describe a method for estimating it. We then motivate the need for assessing the Consistency of an explanation and propose a measure.

Definition 1 Generalizability
A measure of how generalizable an explanation is, and the extent to which it truly reflects the underlying process by which the predictor makes a decision.

Intuitively, a representative explanation would be an explanation that holds for a large number of samples. To assess the number of samples that can be covered by a given explanation, it might be tempting to compute a distance between the explanations associated with those samples. However, because of the large variations in the appearance of objects that arise because of translation, scale, and 3D rotation in natural images, two explanations can be similar (i.e., close in pixel space) without necessarily reflecting a similar visual strategy used by the predictor (for instance, decisions could be driven by the same pixel locations – yet driven by different visual features). Conversely, two spatially distant explanations could be based on the same features that appear at different locations because of translation.

Our proposed solution to this problem is to only use distance measured between explanations for the same sample. This constraint leads us to consider the notion of algorithmic stability as a proxy for generalization: intuitively, given a predictor and a training data set, a good explanation for a decision made for a given data point should be robust to the addition or removal of that data point from the training set. One benefit of such a characteristic is that it can be evaluated based solely on a distance between explanations from the same sample.

In what follows, we will propose a relaxed version of the algorithmic stability – computationally more manageable – applied to the explanations using several predictors trained on different folds. It is important to note that the term algorithmic stability [7] is not related to the Stability of an explanation as defined in [6].

Following this consideration, we will be looking at how well a predictor’s explanations generalize from seen to unseen data points:

\[
d_{E}^{(i,j)}(x) = d(\phi_x^{(i)}, \phi_x^{(j)}) \text{ s.t. } x \in V_i, i \neq j. \tag{1}
\]

By making sure that \( x \) belongs to the fold \( V_i \), we measure the distance between two explanations, one of which comes from a predictor that was not fitted to the sample \( x \). By computing these distances, we hope to characterize the Generalizability of the explanations.

Definition 2 Consistency
The extent to which different predictors trained on the same task do not exhibit logical contradictions.

A statement, or a set of statements, is said to be logically consistent when it has no logical contradictions. A logical contradiction occurs when both a statement and its negation are found to be true. In logic, a fundamental law – the law of non-contradiction – is that a statement and its negation cannot both be true simultaneously. Similarly, we measure the consistency between explanations by ensuring that contradictory predictions lead to different explanations.

Following this definition, if the same explanation gets associated with two contradictory predictions the explanation is said to be inconsistent. This means avoiding the case where for an observation \( x \in V_i \), two predictors \( f_i, f_j \) (where \( i \neq j \)), trained on the same task, give the same explanation but different predictions:

\[
f_i(x) \neq f_j(x) \implies \phi_x^{(i)} \neq \phi_x^{(j)} \tag{2}
\]
Nevertheless, we have to define what it means for two explanations to be different. For this, we use a measure of dissimilarity between explanations and a threshold to judge whether the explanations are consistent or not. This threshold will be relative to the distance between explanations when predictions are not contrary. By measuring the rate of inconsistent explanations, we hope to capture the notion of consistency for explanations.

### 3.3. k-Fold Cross-Training

We recall that our data set is divided into $k$-folds of the same size $D = \{ \mathcal{V}_i \}_{i=0}^{k}$, and that each predictor is trained through a learning algorithm $f_i = \mathcal{A}(\mathcal{Y} \setminus \mathcal{V}_i)$. We assume that the predictors exhibit comparable accuracies across folds. In our experiments, we ensure a similar accuracy on the test set.

We will now measure the distances between two explanations associated with these different predictors. To be more precise, we are only really interested in computing $\delta^{(i,j)}_x$ (see Eq. 1), the distance between two explanations whereby one of the two predictors was not fitted on $x$. Otherwise, it may be trivial for two predictors that were trained on that sample to yield the same explanation – especially if overfitting occurs (see Fig. 2).

In the case where both predictors gave a correct prediction, a small distance between the two explanations suggests that the explanations receive support from several samples. In other words, the fact that explanations do not vary widely when adding or removing a particular sample or set of samples suggests good generalizability. Alternatively, if the two predictors give contrary predictions, the corresponding explanations should be different. Indeed, the very notion of consistency between explanations implies that the same explanation cannot account for two different outcomes.

We separate distances into two sets, $\mathcal{S}^=$ when the predictors have made correct predictions s.t. it is desirable to have a small distance between explanations, $\mathcal{S}^<$ when one of the predictors have given a wrong prediction s.t. it is desirable to have higher distances between the pairs of explanations. The case where both predictors give a bad prediction is ignored (for details, see the Alg. 1 in the appendix).

$\mathcal{S}^= = \{ \delta^{(i,j)}_x : f_i(x) = y \land f_j(x) = y \} \quad (3)$

$\mathcal{S}^< = \{ \delta^{(i,j)}_x : f_i(x) = y \land f_j(x) = \bar{y} \} \quad (4)$

$\forall (i, j) \in \{ 1, ..., k \}^2 \text{ s.t. } i \neq j, \forall (x, y) \in \mathcal{V}_i$

### 3.4. Mean generalizability : MeGe

From Def. 1, the distance between explanations arising from predictors trained on a dataset that contained vs. did not contain a given sample should be small. As those distances are contained in $\mathcal{S}^=$, one way to measure the generalizability of explanations is to compute the average distance over $\mathcal{S}^=$.

As a reminder, the average of $\mathcal{S}^<$ corresponds to the average change of explanation when the sample is removed from the training set. This change is related to the generalizability of the explanation: the more representative an explanation is, the more it persists when we remove a point.

To ensure a high value for low distances, we define the MeGe measure as a similarity measure:

$$ MeGe = \left(1 + \frac{1}{|\mathcal{S}^=|} \sum_{\delta \in \mathcal{S}^=} \delta \right)^{-1} \quad (5) $$

Explanations with good generalizability will therefore be associated with higher similarity scores between explanations (close to 1).

### 3.5. Relative consistency : ReCo

From Def. 2 and Eq. 2, explanations arising from different predictors are said to be consistent if they are close when the predictions agree with one another. As a reminder, the distance between explanations for the consistent predictions are represented by $\mathcal{S}^=$, and those associated with inconsistent predictions by $\mathcal{S}^<$. Visually, we seek to maximize the shift between the corresponding distributions for the sets $\mathcal{S}^=$ and $\mathcal{S}^<$. Formally, we are looking for a distance value that separates $\mathcal{S}^=$ and $\mathcal{S}^<$, e.g., such that all the lower distances belong to $\mathcal{S}^=$ and the higher ones to $\mathcal{S}^<$. The clearer the separation, the more consistent the explanations are. In order to find this separation, we introduce ReCo, a statistical measure based on maximizing the balanced accuracy.

Where $\mathcal{S} = \mathcal{S}^= \cup \mathcal{S}^<$ and $\gamma \in \mathcal{S}$ a fixed threshold value, we can define the true positive rate $TPR$ as the rate for which distances below a threshold come from consistent predictions among all distances below the threshold $TPR(\gamma) = \frac{|\{ \delta \in \mathcal{S}^= : \delta \leq \gamma \}|}{|\{ \delta \in \mathcal{S}^= : \delta \leq \gamma \}|}$.

A similar way, $TNR$ denotes the rate for which distances above a threshold come from opposite predictions among all the distances above the threshold $TNR(\gamma) = \frac{|\{ \delta \in \mathcal{S}^< : \delta > \gamma \}|}{|\{ \delta \in \mathcal{S}^< : \delta > \gamma \}|}$.

Basing our measure on these rates allows us to assess the quality of these explanations independently of the accuracy of the predictor, we define ReCo as the maximal balanced accuracy:

$$ ReCo = \max_{\gamma \in \mathcal{S}} TPR(\gamma) + TNR(\gamma) - 1, \quad (6) $$

with a score of 1 indicating perfect consistency of the predictors’ explanations, and a score of 0 indicating a complete inconsistency.

### 4. Experiments

We carried out three sets of experiments using a variety of neural network architectures and explanation methods. The first one consisted in ensuring the functioning and the reliability of the measures ReCo and MeGe via a simple sanity check done over a large number of predictors (175
in total). The second set of experiments consisted in highlighting a limitation of the fidelity measure – namely its independence with respect to the quality of the explanations. This underlines the need for new measures that are dedicated to explanations and not to methods. We developed these considerations in a dedicated section where we demonstrate an application to the selection of a method using the two new criteria MeGe and ReCo. Finally, in a third set of experiments, we showed quantitatively that some predictors are more interpretable: our analyses revealed that 1-Lipschitz neural networks yield explanations that are more coherent and representative.

4.1. Setup

For all experiments, we used 5 splits ($k = 5$), i.e., 5 predictors with comparable accuracy ($\pm 3\%$), which allows us to study the explanations in common training conditions (80% of the data are used for training and 20% for testing).

For ILSVRC 2012, our predictors are based on a ResNet-50 architecture [21], and a ResNet-18 for the other datasets (see appendix E for details on each predictor).

Explanation methods In order to produce the necessary explanations for the experiment, we used 7 methods of explanation. The methods selected are those commonly found in the literature in addition to one control method (Random).

The explanation methods chosen are as follow: Saliency (SA) [47], Gradient ⊙ Input (GI) [3], Integrated Gradients (IG) [53], SmoothGrad (SG) [49], Grad-CAM (GC) [43], Grad-CAM++ (G+) [11] and RISE (RI) [35]. Further information on these methods can be found in the appendix B.

Datasets We applied the procedure described above and evaluated the proposed measures for each of the degradations on 4 image classification datasets:

**ILSVRC 2012** [12]: a subset of the ImageNet dataset from which we randomly selected 50 classes. The size of the images considered was 224 × 224. The reduced number of classes being sufficient to show that the metrics pass the test performed in 4.2 even in the case of high dimensional images.

**CIFAR10** [24]: a low-resolution labeled datasets with 10 classes respectively, consisting of 60,000 (32 × 32) color images.

**EuroSAT** [22]: a labeled dataset with 10 classes consisting of 27,000 color images (64 × 64) from the Sentinel-2 satellite.

**Fashion MNIST** [59]: a dataset containing 70,000 low-resolution (28 × 28) grayscale images labeled in 10 categories.

Distance over explanations The procedure introduced in section 3.3 requires to define a distance between two explanations derived for the same sample. Since a feature attribution consists of ranking the features most sensitive to the predictor’s decision, it seems natural to consider the Spearman rank correlation [51] to compare the similarity between explanations. Several authors have provided theoretical and experimental arguments in line with this choice [18, 1, 56]. However, it is important to note that the problem of measuring similarity between explanations is still an open problem. We conduct two sanity checks: spatial correlation, and noise test on several candidates distances to ensure they could respond to the problem. The distances tested were built from: 1-Wasserstein distance (the Earth mover distance from [16]), Sørensen–Dice [13] coefficient, Spearman rank correlation, SSIM [64], and $\ell_1$ and $\ell_2$ norms. In line with prior work, we chose to use one minus the absolute value of the Spearman rank correlation (see F for more details).

4.2. Sanity check for explanation measures

Our first set of experiments aims to ensure that the propose metrics approximate the desired quantities by perform-
Figure 3. **MeGe and ReCo scores** for predictors trained with no degradations (first point from the left), as well as for progressively randomized predictors and predictors trained with switched labels. For all the methods tested, the more the predictor is degraded, the more the Consistency and Generalizability scores drop, which means that the associated metrics pass the sanity check. **Top** ImageNet. **Bottom** Cifar-10.

To ensure that the metric captures these notions, we applied two different types of degradation on the predictors for each data set: weight randomizations and label shuffling.

- Randomizing the weights, inspired by [1]. We gradually randomize 5%, 10% and 30% of the predictor layers by adding Gaussian noise. By degrading the weights learned by the network, we expect to find degraded explanations.

- Shuffling of labels, inspired by [33, 1] the predictors are trained on a data set with 5%, 10% and 30% of bad labels. By artificially breaking the relationship between the labels, we expect the explanations to lose their consistency.

The MeGe measure encodes the Generalizability of the explanations, which is related to the ability of the predictor to derive general strategies. Thus, the degradation of the parameters of a predictor directly affects these strategies. Fig. 3 shows the correlation of the measures with the intensity of the degradation applied: MeGe and ReCo capture the degradation of the explanation and pass the sanity check.

We note that all the tested methods perform better than the random baseline (random). However, the drop in score, is not the same and some methods are more sensitive to predictor changes, such as Grad-CAM or RISE, in accordance with previous work [1, 48]. It was subsequently observed that this sensitivity seems to translate into a better Fidelity score for the methods. Nevertheless, it should be noted that this sanity test is a necessary but not sufficient condition for a Generalizability and Consistency metric.

### 4.3. The implications of the fidelity metric

To mark the difference between the proposed measures and the Fidelity, we applied the $\mu F$ measure from [6] (see supplementary document C) to the normally trained predictors and those progressively degraded. We seek to verify that this property does not pass the sanity check: the fidelity
measure is invariant to the performancee of the predictor as well as to the quality of its explanations. For $\mu F$, the score obtained is averaged over 10,000 test samples, with 0 for baseline. The size of the $|S|$ subset is 15% of the image.

As shown in Fig. 4, predictor degradation does not impact the Fidelity metric on the methods tested. The Fidelity property is essential in a good explanation since it allows us to make sure that we are studying the strategies of the predictor. However, it is not sufficient: if the explanation reflects well the strategies of the predictor, the latter may use specific and inconsistent strategies. In that, the Fidelity is intended to ensure that the explanations correctly reflect the underlying strategies of the model, regardless of whether these strategies are general or consistent.

As shown in Fig. 4, predictor degradation does not impact the Fidelity metric on the methods tested. The Fidelity property is essential in a good explanation since it allows us to make sure that we are studying the strategies of the predictor. However, it is not sufficient: if the explanation reflects well the strategies of the predictor, the latter may use specific and inconsistent strategies. In that, the Fidelity is only a first step towards a good explanation.

### 4.4. Method selection criterion

The MeGe and ReCo measures can be used as additional criteria for choosing an explainability method. As a reminder, a good method should provide explanations that are as faithful as possible and, if possible, consistent and representative. Thus, the tested methods can be compared using the scores obtained for these measures. We note that these measures are complementary in that the fidelity score can be interpreted as a confidence bound on the other measures performed on the explanations.

Table 1 reports the Fidelity ($\mu F$), Consistency (ReCo) and Generalizability (MeGe) scores obtained for the ResNet-50 predictors trained without degradation on ImageNet. We can exploit a selection criterion from the differences in scores.

First of all, we notice that the two methods obtaining a good fidelity score are RISE and Grad-CAM, they reflect well the predictor functioning. Their high fidelity score acts as a confidence bound on the MeGe and ReCo metrics: by correctly transcribing the functioning of the predictor, we obtain at the same time the Generalizability and the Consistency of the explanations. This score can then be used as a criterion to separate RISE from Grad-CAM. In view of the differences between the MeGe and ReCo scores, RISE method seems preferable.

Concerning the Generalizability score, it is important to note that two methods tested here involve the element-wise product of the explanation with the input: Integrated Gradients and Gradient Input. This operation could eliminate the attribution score on a part of the image, thus reducing the distance between the two explanations. The result is a better MeGe score which is in fact due to the dominance of input in the element-wise product.

It can be observed that the change of predictor has an effect on this ranking, and that a good method of explainability must be chosen according to a context: predictor and data set. However, even considering these effects, the experiments carried out suggest 3 methods that give faithful, representative and consistent explanations: Grad-CAM, Grad-CAM++ and RISE (for more results on Cifar-10, EuroSAT and Fashion MNIST, see appendix G).

### 4.5. Towards predictors with better explanations

In an attempt to find predictors that give better explanations, we extend the experience on the Cifar-10 dataset by adding a family of 1-Lipschitz networks. Indeed different works mention the Lipschitz constrained networks as particularly robust [57, 41, 34, 8] and have good generalizability. As a reminder, a $f$ function is called $L$-Lipschitz, with $L \in \mathbb{R}^+$ if $|f(x_1) - f(x_2)| \leq L|x_1 - x_2|$ for every pair $(x_1, x_2) \in \mathbb{X}^2$. The smallest of these $L$ 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 also comes with new generalization bounds that critically rely on the Lipschitz constant of the neural network [58, 33, 5].

The predictors were trained using the Deel-Lip library [45]. All the predictors, including the 1-Lipschitz,
have comparable accuracy (78 ± 4%). To our knowledge, no previous work has made the link between Lipschitz networks and the chosen explainability methods.

Figure 5. Lipschitz predictors (right column) on Cifar10. As explained in this paper, a clear separation between the $S^\neq$ and $S^\#$ histograms is a sign of consistent explanations.

The Fig. 5 shows the difference in $S^\neq$ and $S^\#$ between ResNet and 1-Lipschitz predictors. In the left column, the results come from ResNet-18 predictors trained on Cifar-10 while the right column is dedicated to 1-Lipschitz predictors. We observe a clear improvement of the consistency and generalization of the explanations respectively as a result of better separation of the histograms and a smaller expectation of $S^\neq$. SmoothGrad is the method that obtains the most consistent explanations as indicated in the table 3, in front of RISE and Saliency (more results in the supplementary material G Fig. 10).

Like MeGe, the results in Table 3 show an improvement for the 1-Lipschitz predictors in the Consistency of the explanations for all the methods tested except for Grad-CAM++, reflecting the more marked separation between the two histograms of $S^\neq$ and $S^\#$ in Fig. 5.

In general, the experiments carried out allow us to observe a clear improvement in the quality of explanations from the 1-Lipschitz predictor. These encouraging results show that there is a close link between the methods used and predictor architectures, as well as the usefulness of Lipschitz networks for explainability. Furthermore, it underlines the fact that the search for new methods is not the only path to explainability: the search for predictors with better explanations is another under-exploited avenue.

5. Conclusion

We introduced a procedure to derive two new measures to characterize important properties of a good explanation: Generalizability and Consistency. We highlight the fact that Fidelity is intended to ensure that the explanations correctly reflect the underlying strategies of the model, regardless of whether these strategies are general or consistent. Finally, as a case in point, we presented a novel analysis using 1-Lipschitz networks. We used our measures to quantify the consistency of their explanations and showed that this class of networks gives much more stable and trustworthy explanations compared to standard neural networks.

We see the present work as constituting a necessary next step in characterizing good explanations – towards the quest for more explainable ML models.

Acknowledgement

This work was conducted as part of the DEEL project*. Funding was provided by ANR-3IA Artificial and Natural Intelligence Toulouse Institute (ANR-19-P3IA-0004). Additional funding to TS was provided by ONR (N00014-19-1-2029), NSF (IIS-1912280 and EAR-1925481), DARPA (D19AC00015), NIH/NINDS (R21 NS 112743). The computing hardware was supported in part by NIH Office of the Director grant S10OD025181. We thank Melanie Ducoffe and Mikael Capelle of the DEEL team for insightful comments which have helped improved the manuscript.

---

*https://www.deel.ai/
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 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] Peter Bartlett, Dylan J. Foster, and Matus Telgarsky. Spectrally-normalized margin bounds for neural networks. In Advances in Neural Information Processing Systems (NIPS), 2017.

[6] 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.

[7] Olivier Bousquet and André Elisseeff. Stability and generalization. The Journal of Machine Learning Research, 2002.

[8] Louis Béthune, Alberto González-Sanz, Franck Mamalet, and Mathieu Serrurier. The many faces of 1-lipschitz neural networks, 2021.

[9] Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso. Machine learning interpretability: A survey on methods and metrics. Electronics, 2019.

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

[11] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2018.

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009.

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

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

[15] Gabriel Ferretti, Elodie Escriva, Julien Aïgon, Jean-Baptiste Excoffier, and Chantal Soulé-Dupuy. Coalitional strategies for efficient individual prediction explanation. Information Systems Frontiers, pages 1–27, 2021.

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

[17] 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.

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

[19] 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.

[20] Thomas Hartley, Kirill Sidorov, Christopher Willis, and David Marshall. Swag: Superpixels weighted by average gradients for explanations of cnns. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021.

[21] 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.

[22] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damien 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.

[23] 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 (NeurIPS), 2019.

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

[25] 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.

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

[27] Jiwei Li, Will Monroe, and Dan Jurafsky. Understanding machine learning interpretability: A survey on methods and metrics. Electronics, 2019.

[28] Drew Linsley, Dan Shiebler, Sven Eberhardt, and Thomas Hartley. Gradient-based visual explanations for deep convolutional networks, 2021.

[29] 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.

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

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

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

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

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

[35] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. In Proceedings of the British Machine Vision Conference (BMVC), 2018.

[36] 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.

[37] 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.

[38] 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.

[39] Marko Robnik-Šikonja and Marko Bohanec. Perturbation-based explanations of prediction models. In Human and machine learning Springer International Publishing, 2018.

[40] 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.

[41] Kevin Scaman and Aladin Virmaux. Lipschitz regularity of deep neural networks: analysis and efficient estimation. In Advances in Neural Information Processing Systems (NeurIPS), 2019.

[42] 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.

[43] 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.

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

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

[46] 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.

[47] 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.

[48] Leon Sixt, Maximilian Granz, and Tim Landgraf. When explanations lie: Why many modified bp attributions fail. In Proceedings of the International Conference on Machine Learning (ICML), 2020.

[49] 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.

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

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

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

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

[54] Fel Thomas, Cadene Remi, Chalvidal Mathieu, Cord Mathieu, Vigouroux David, and Serre Thomas. Look at the variance! efficient black-box explanations with sobol-based sensitivity analysis. In Advances in Neural Information Processing Systems (NeurIPS), 2021.

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

[56] 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.

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

[58] Ulrike von Luxburg and Olivier Bousquet. Distance-based classification with lipschitz functions. The Journal of Machine Learning Research, 2004.

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

[60] 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 (NeurIPS), 2019.

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

[62] 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.
[63] 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.

[64] 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.

[65] 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. Method Details

Algorithm 1 Training procedure to compute \( S^= \) and \( S^\neq \)

Require: \( k \in \mathbb{N}_{\geq 2}, \ D = \{ \mathcal{V}_i \}_{i=1}^k \)

\( S^= \leftarrow \{ \}, \ S^\neq \leftarrow \{ \} \)

for all \( i \in \{ 1, \ldots, k \} \) do

Train \( f_i \) on \( D \setminus \mathcal{V}_i \)

for all \( (x, y) \in D \) do

// generate explanations on all dataset
\( \Phi^{(i)}_x \leftarrow \Phi(f_i, x) \)

end for

end for

for all \( i \in \{ 1, \ldots, k \} \) do

for all \( (x, y) \in \mathcal{V}_i \) do

for all \( j \in \{ 1, \ldots, k \mid i \neq j \} \) do

// \( f_j \) was trained on \( x \), \( f_i \) was not
\( \delta^{(i,j)}_x \leftarrow d(\Phi^{(i)}_x, \Phi^{(j)}_x) \)

if \( f_i(x) = y \) and \( f_j(x) = y \) then

// both model are correct
\( S^= \leftarrow S^= \cup \{ \delta^{(i,j)}_x \} \)

else if \( f_i(x) = y \) or \( f_j(x) = y \) then

// only one model is correct
\( S^\neq \leftarrow S^\neq \cup \{ \delta^{(i,j)}_x \} \)

end if

end for

end for

end for

Return \( S^=, S^\neq \)

B. Explanation methods

In the following section, the formulation of the different methods used is given. As a reminder, we focus on a classification model \( f : \mathbb{R}^d \rightarrow \mathbb{R}^C \) where \( C \) is the number of classes. We assume \( f_c(x) \) the logit score (before softmax) for class \( c \). An explanation method provides an attribution \( \Phi \in \mathbb{R}^d \) for each input feature from a model and an input of interest. Each value then corresponds to the importance of this feature for the model results.

Saliency Map (SA) 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.

\[
\Phi^{SA}(x) = \left| \frac{\partial f_c(x)}{\partial x} \right|
\]

Gradient ⊙ Input (GI) 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. A theoretical analysis conducted by [3] showed that Gradient ⊙ Input is equivalent to e-LRP and DeepLIFT methods under certain conditions: using a baseline of zero, and with all biases to zero.

\[
\Phi^{GI}(x) = x \odot \left| \frac{\partial f_c(x)}{\partial x} \right|
\]

Integrated Gradients (IG) 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, we use a Trapezoidal rule and not a left-Riemann summation, which allows for more accurate results and improved performance (see [50] for a comparison). The final result depends on both the choice of the baseline \( x_0 \) and the number of points to estimate the integral. In the context of these experiments, we use zero as the baseline and \( m = 60 \).

\[
\Phi^{IG}(x) = (x - x_0) \int_0^1 \frac{\partial f_c(x_0 + \alpha(x - x_0))}{\partial x} \, d\alpha
\]

SmoothGrad (SG) 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 \( \sigma \)) 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.

\[
\Phi^{SG}(x) = \mathbb{E}_{\epsilon \sim \mathcal{N}(0, \sigma^2)} \left[ \frac{\partial f_c(x + \epsilon)}{\partial x} \right]
\]

Grad-CAM (GC) can be used on Convolutional Neural Network (CNN), it uses the gradient and the feature maps \( A^{(k)} \) of the last convolution layer. More precisely, to obtain the localization map for a class, we need to compute the weights \( \alpha^{(k)}_c \) associated to each of the feature map activation \( A^{(k)} \), with \( k \) the number of filters and \( Z \) the number of features in each feature map we define \( \alpha^{(k)}_c = \frac{1}{Z} \sum_j \frac{\partial f_c(x)}{\partial A^{(k)}_j} \) and

\[
\Phi^{GC} = \max(0, \sum_k \alpha^{(k)}_c A^{(k)})
\]

Notice that the size of the explanation depends on the size (height, width) of the last feature map, a bilinear interpolation is performed in order to find the same dimensions as the input.

Grad-CAM++ (G+) is an extension of Grad-CAM combining the positive partial derivatives of feature maps of a convolutional layer with a weighted special class score.
weights $\alpha^{(k)}_c$ associated to each feature map is computed as follow:

$$
\alpha^{(k)}_c = \sum_i \sum_j \left[ \frac{\partial^2 f_c(x)}{(\partial A^{(k)}_{ij})^2} + \sum_i \sum_j A_{ij}^{(k)} \frac{\partial^3 f_c(x)}{(\partial A^{(k)}_{ij})^3} \right]
$$

RISE (RI) is a black-box method that consist of probing the model with randomly masked versions of the input image to deduce the importance of each pixel using the corresponding outputs. The binary masks $m \sim \mathcal{M}$ are generated in a subspace of the input space, then upsampled with a bilinear interpolation (once upsampled the masks are no longer binary).

For ImageNet the number of masks was $m = 4000$, for all the other datasets $m = 1000$.

$$
\Phi^{RI}(x) = \frac{1}{E(\mathcal{M})N} \sum_{i=0}^{N} f_c(x \odot m_i) m_i
$$

C. Fidelity

Various fidelity metrics have been proposed that essentially measure the correlation between input variables and the drop in score when these variables are set to a baseline state [40, 60, 39, 35]. In this work, we use $\mu F$ from [6]:

$$
\mu F = \mathbb{C}orr_{S \subseteq \{1, \ldots, d\}} \left( \sum_{i \in S} \Phi(f, x)_i, f(x) - f(x|x_i = \bar{x}, i \in S) \right)
$$

(7)

Where $f$ is a predictor, $\Phi$ an explanation function, $S$ a subset indices of $x$ and $\bar{x}$ a baseline reference. The choice of a proper baseline is still an active area of research [52].

D. 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 $\mathcal{S}^\equiv$ and $\mathcal{S}^\neq$ histograms would be to consider the usual measures, such as Kullback-Leibler ($KL$) divergence.

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. Similarly, considering the 1-Wasserstein measure, we could construct an inconsistent case by exploiting the invariance to the direction of transport. For these reasons, we have therefore chosen a classification measure, based on maximizing balanced accuracy. Nevertheless, one could also (observing similar results) use the area under the curve (AUC) of the balanced accuracy, such as:

$$
ReCo_{AUC} = \frac{1}{|S|} \sum_{\gamma \in S} TPR(\gamma) + TNR(\gamma) - 1
$$

E. 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 had drop out 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:

| Dataset     | Model      | Base Filters | Dropout Rate | Accuracy ±1% |
|-------------|------------|--------------|--------------|---------------|
| Fashion-MNIST | Conv2D(26) | 26           | 0.4          | 92% ±1%       |
| EuroSAT     | Conv2D(46) | 46           | 0.25         | 95% ±1%       |
| Cifar10     | Conv2D(32) | 32           | 0.25         | 78% ±4%       |
| ImageNet    | ResNet50   | 256          |             | 88% ±3%       |

E.1. Lipschitz models

The 1-Lipschitz models use spectral regularization on the Dense and Convolutions layers. The architecture is as described in Table 4.

E.2. Randomization test

For the randomisation of the model weights, we added noise drawn from a normal distribution $\varepsilon \sim \mathcal{N}(0, 0.5)$ to each convolution layer, with the intensity of the degradation impacting on the number of parameters affected by this noise.
F. Distances tests

F.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 result, 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 Fig. 6).

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.

F.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 $\varepsilon \sim \mathcal{N}(0, I\sigma^2)$ an isotropic Gaussian noise, we wish the distance $d$ to show a monotonic positive correlation $\text{corr} (\text{dist}(x, x + \varepsilon), \varepsilon)$.

In order to validate this prerogative, a Gaussian noise with a progressive intensity $\sigma$ 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 $\sigma$ the operation is repeated 50 times.

Over the different distances tested, they all pass the sanity test: there is a monotonous positive correlation (as seen in Fig. 7). Although SSIM and $\ell_2$ 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.

---

G. Additional results

Table 5. Fidelity, Consistency and Generalizability score for ResNet-18 models on Cifar10. Higher is better. The first and second best results are respectively in **bold** and *underlined*.

| Metrics | IG    | SG    | SA    | GI    | GC    | G+   | RI   |
|---------|-------|-------|-------|-------|-------|------|------|
| µF      | 0.11  | 0.31  | 0.23  | 0.10  | **0.91** | 0.89 | 0.84 |
| MeGe    | 0.58  | 0.46  | 0.45  | 0.55  | **0.72** | **0.82** | 0.56 |
| ReCo    | 0.11  | 0.15  | 0.15  | 0.09  | 0.64  | 0.49 | 0.52 |

Table 6. Consistency and Generalizability score for ResNet-18 models on Eurosat. Higher is better. The first and second best results are respectively in **bold** and *underlined*.

| Metrics | IG    | SG    | SA    | GI    | GC    | G+   | RI   |
|---------|-------|-------|-------|-------|-------|------|------|
| MeGe    | 0.40  | 0.42  | 0.41  | 0.41  | **0.67** | **0.67** | 0.39 |
| ReCo    | 0.31  | 0.18  | 0.18  | 0.23  | 0.59  | **0.64** | 0.34 |

Table 7. Consistency and Generalizability score for ResNet-18 models on Fashion-MNIST. Higher is better. The first and second best results are respectively in **bold** and *underlined*.

| Metrics | IG    | SG    | SA    | GI    | GC    | G+   | RI   |
|---------|-------|-------|-------|-------|-------|------|------|
| MeGe    | **0.90** | 0.36  | 0.30  | **0.90** | 0.77  | 0.84 | 0.52 |
| ReCo    | **0.37** | 0.13  | 0.10  | **0.37** | **0.52** | 0.32 | 0.37 |
Figure 8. **Eurosat** MeGe and ReCo scores for normally trained models (first point from the left), as well as for progressively randomized models and models trained with switched labels.

Figure 9. **Fashion-MNIST** MeGe and ReCo scores for normally trained models (first point from the left), as well as for progressively randomized models and models trained with switched labels.

Figure 10. $S^{=}$ and $S^{\neq}$ for ResNet (left column) and 1-Lipschitz models (right column) on Cifar10. As explained in this paper, a clear separation between the $S^{=}$ and $S^{\neq}$ histograms is a sign of consistent explanations.