How good is your explanation? Algorithmic stability measures to assess the qualityof explanations for deep neural networks

Thomas Fel, David Vigouroux, Rémi Cadène, Thomas Serre

How good is your explanation? Algorithmic stability measures to assess the quality of explanations for deep neural networks

Thomas Fel[1,2,3], David Vigouroux[3], Rémi Cadène[2], and Thomas Serre[1,2]

[1]Artificial and Natural Intelligence Toulouse Institute, Universite de Toulouse, France
[2]Carney Institute for Brain Science, Department of Cognitive Linguistic & Psychological Sciences, Brown University, Providence, RI 02912
[3]IRT Saint-Exupery

Abstract

A plethora of methods have been proposed to explain how deep neural networks reach a decision but comparatively little effort has been made to ensure that the explanations produced by these methods are objectively relevant. While desirable properties for a good explanation are easy to come, objective measures have been harder to derive. Here, we propose two new measures to evaluate explanations borrowed from the field of algorithmic stability: relative consistency ReCo and mean generalizability MeGe. We conduct several experiments on multiple image datasets and network architectures to demonstrate the benefits of the proposed measures over representative methods. We show that popular fidelity measures are not sufficient to guarantee good explanations. Finally, we show empirically that 1-Lipschitz networks provide general and consistent explanations, regardless of the explanation method used, making them a relevant direction for explainability.

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 classification accuracy – approaching human performance in specific domains [24, 42]. However, one significant drawback associated with these deep networks is that it is difficult to interpret their decisions [27]. This problem constitutes a serious obstacle for the wide adoption of these systems for safety-critical applications such as aeronautics.

Recently, several explainability methods have been proposed to help understand how these predictors make particular decisions [51, 47, 33, 41]. Unfortunately, these methods have strong limitation. One particularly problematic limitation is the so-called confirmation bias: while some methods appear to offer useful explanations to a human experimenter, these methods turn out not to reflect the actual behaviour of the predictor [1, 16]. In other words, the explanations produced by these methods, which are supposed to provide

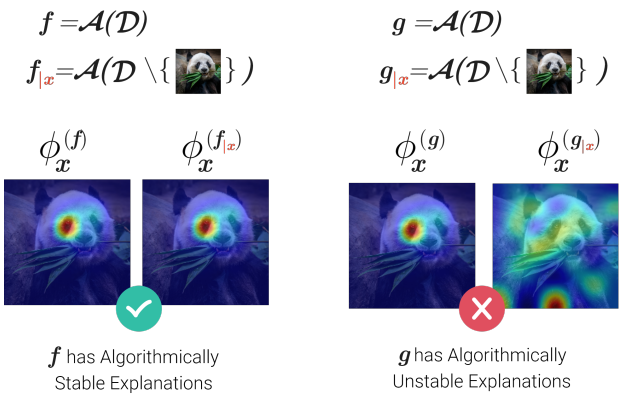

Figure 1. A predictor $\boldsymbol{f}$ can be algorithmically stable in the sense that $\boldsymbol{f}$ and $\boldsymbol{f}_{|\boldsymbol{x}}$ trained with the same algorithm $\mathcal{A}$ on a dataset $\mathcal{D}$ with or without the sample $\boldsymbol{x}$ produces similar predictions. Yet, even for a stable predictor, the associated explanations $\phi_x^{(\boldsymbol{f})}$ and $\phi_x^{(\boldsymbol{f}_{|\boldsymbol{x}})}$ may or may not be stable (i.e., the explanations may be similar or dissimilar for these two scenarios). Any variation in the explanations associated with the two scenarios informs us about the *Representativity* of the underlying explanations: if the explanations are stable over perturbations of the training set, these explanation are intuitively more representative of an underlying strategy used by the predictor to arrive at its decision. Hence, we propose to use the distance between the explanations $\phi_x^{(\boldsymbol{f})}$ and $\phi_x^{(\boldsymbol{f}_{|\boldsymbol{x}})}$ as a measure of the *Representativity* of an explanation or how representative it is of a predictor's underlying strategy: the more stable the explanations, the more structured the associated predictions are, and the more support they receive from other samples.

confidence in a system's decisions are themselves potentially erroneous.

What is needed are methods to objectively assess the quality of explanations produced to allow for systematic benchmarks and baselines to be established. The main approaches aim to ensure that the underlying explanations satisfy a certain number of properties (or axioms) such as *Fidelity*, *Stability*, *Representativity*or *Consistency* [8, 52, 29, 37, 17, 2]. Among those, the most studied property is *Fidelity*, which allows to choose the best method to explain a given predictor. However, as we will show there this measure cannot capture all the relevant desiderata of an explanation.

In this work, we propose two novel criteria and associated measures to characterize the *Representativity* and the *Consistency* of explanations. This measures from the fields of algorithmic stability and generalisation [7]. Informally, a learning algorithm is guarantee to generalize if its decision does not change too much when a single training example is removed from its training set. Here, the main idea is to apply a similar notion to the explanations. That is, an explanation is said to be representative if similar explanations are obtained for a predictor trained on the slightly perturbed versions of the data set. Intuitively, a good explanation is one such that given consistent predictor decisions when trained on two perturbed data sets, will give consistent explanations. In practice, we use a k-*fold* cross-training approach in order to derive these measures. We then estimate the average generalizability (MeGe) and relative consistency (ReCo): MeGe is intended to measure the ability of a predictor to derive general rules from its explanations while ReCo is motivated by the idea that one explanation should not be used to justify two contradictory decisions.

We provide an extensive experimental validation of the approach using different neural network predictors and multiple image datasets. We compare the proposed *Representativity* and *Consistency* measures against the leading *Fidelity*measure [6, 57] and show that a faithful explanation is not necessarily representative or consistent. Finally, we show quantitatively that 1-Lipschitz networks give more general and consistent explanations, offering an interesting research track in the field of explainable AI towards more explainable predictors.

To summarize, **our main contributions** include:

- Novel measures borrowed from the field of algorithmic stability for evaluating the quality of explanations provided by a predictor: *Representativity* (MeGe) and *Consistency* (ReCo).
- Extensive experimental validation of the approach on various images datasets (including ImageNet).
- Demonstration that current *Fidelity* measures are severly limited in their ability to characterize the quality of explanations.
- Empirical demonstration that 1-Lipschitz networks deliver general and consistent explanations irrespective of the method of explanation used.

2. Related Works

In this work, we focus on evaluating explainability methods that better understand how a given neural network architecture reaches a particular decision [13]. 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 [59, 45, 4, 15, 44, 51, 47, 41, 18] or with respect to a perturbation of the input [58, 60, 34, 25, 62, 35].

Despite a wide range of explainability methods, there is a lack of research on the development of measures and approaches for assessing the quality of these explanations. It is in part due to the difficulty of obtaining objective ground truths [38, 26]. Several criteria have been proposed to evaluate the quality of explanations [52, 29, 37, 17, 2, 8]. According to [8], the five major properties include: *Fidelity*, *Stability*, *Comprehensibility*, *Representativity* and *Consistency*.

There are two main approaches currently used 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 [41, 34, 28], or by indirectly measuring the performance of the human/classifier duo [23, 9, 30, 40]. Nevertheless, human intervention sometimes brings undesirable effects, including a possible confirmation bias [1].

A second approach has also started to emerge specifically for the domain of computer vision. 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 [38] 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, a method called ROAR was proposed [21] based on re-training a classifier model between each deletion step. An alternative approach which requires low resources to be calculated is IROF [36] . 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 [57, 6]. Nevertheless, it should be noted that the different fidelity metrics proposed requires to define a proper baseline state which is not always available [50].

Those *Fidelity* metrics are a first step toward a good explanations: by making sure that we have faithful explanations,

we can then look at other criteria to quantitatively measure these explanations.

Stability measures, on the other hand, consist in calculating the sensibility of an explanation around [57, 3, 6]. Intuitively, a good explanation should be valid not just for a particular sample but also for local neighborhood. Several necessary properties lack an associated measure, notably *Consistency* and *Representativity*. Indeed, a classifier that overfit 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.

This work describes a new approach involving cross-validation on explanations to evaluate the *Consistency* and *Representativity* through two new measures: the relative consistency MeGe and the mean generalization ReCo.

3. Method

Below, we briefly provide some motivation for the proposed MeGe and ReCo measures before describe 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 datapoint is added or removed from the dataset.

3.1. Notations

We consider a standard supervised learning setting where a datapoint is denoted $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{y})$ s.t. $\boldsymbol{x} \in \mathcal{X}$ is an observation (e.g., $\mathcal{X} = \mathbb{R}^d$) and $\boldsymbol{y} \in \mathcal{Y}$ is a class label (e.g., $\mathcal{Y} = \mathbb{R}^p$). The data set is denoted as $\mathcal{D} = \{\boldsymbol{z}_1, ..., \boldsymbol{z}_m\}$, we designate $\mathcal{V} = \{\mathcal{V}_1, ..., \mathcal{V}_k\}$ the set of k disjoints subsets (*folds*) of size m/k at random where each $\mathcal{V}_i \subset \mathcal{D}$. Throughout this work, we will assume k divides m for convenience. Let $\mathcal{A}$ be a deterministic learning algorithm which maps any number of data points onto a function $\boldsymbol{f}$ from $\mathcal{X}$ to $\mathcal{Y}$. In particular, we consider the *fold* $\mathcal{V}_i$ and the associated predictor $\boldsymbol{f}_i = \mathcal{A}(\mathcal{V} \setminus \mathcal{V}_i)$.

An explanation method is a functional, denoted $\boldsymbol{\Phi}$, which, given a predictor $\boldsymbol{f}_i$ and a datapoint $\boldsymbol{x}$, assigns an importance score for each input dimension $\boldsymbol{\phi}_{\boldsymbol{x}}^{(i)} = \boldsymbol{\Phi}(\boldsymbol{f}_i, \boldsymbol{x})$. 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 *Representativity*: 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 *Representativity*
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 can be associated with 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 characteristic is that it can be evaluated based solely on a distance between explanations from the same samples. 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:

$$\delta_{\boldsymbol{x}}^{(i,j)} = d(\boldsymbol{\phi}_{\boldsymbol{x}}^i, \boldsymbol{\phi}_{\boldsymbol{x}}^j) \; s.t. \; \boldsymbol{x} \in \mathcal{V}_i, i \neq j. \tag{1}$$

By making sure that $\boldsymbol{x}$ belongs to the *fold* $\mathcal{V}_i$, we measure the distance between two explanations, one of which comes from a predictor that was not fitted to the sample $\boldsymbol{x}$. By computing these distances, we hope to characterize the *Representativity* 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 $\boldsymbol{x} \in \mathcal{V}_i$, two predictors $\boldsymbol{f}_i, \boldsymbol{f}_j$ (where $i \neq j$), trained on the same task, give the same explanation but different predictions:

$$\boldsymbol{f}_i(\boldsymbol{x}) \neq \boldsymbol{f}_j(\boldsymbol{x}) \implies \boldsymbol{\phi}_{\boldsymbol{x}}^{(i)} \neq \boldsymbol{\phi}_{\boldsymbol{x}}^{(j)} \tag{2}$$

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 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 $\mathcal{D} = \{\mathcal{V}_i\}_{i=0}^{k}$, and that each predictor is trained through a learning algorithm $\boldsymbol{f}_i = \mathcal{A}(\mathcal{V} \setminus \mathcal{V}_i)$. 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 associate with these different predictors. To be more precise, we are really only interested in computing $\delta_{\boldsymbol{x}}^{(i,j)}$ (see Eq. 1):, the distance between two explanations whereby one of the two predictors was not fitted on $\boldsymbol{x}$. Otherwise, it may be trivial for two predictors that were trained on that sample to yield the same explanation – especially if overfitting occurs..

In the case where both predictors gave a correct prediction, a small distance between the two explanations suggest 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 suggest good *Representativity*. 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, $\boldsymbol{\mathcal{S}}^{=}$ when the predictors have made correct predictions s.t. it is desirable to have a small distance between explanations, $\boldsymbol{\mathcal{S}}^{\neq}$ 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 Algorithm 1 in the appendix).

$$\boldsymbol{\mathcal{S}}^{=} = \{\delta_{\boldsymbol{x}}^{(i,j)} : \boldsymbol{f}_i(\boldsymbol{x}) = \boldsymbol{y} \wedge \boldsymbol{f}_j(\boldsymbol{x}) = \boldsymbol{y}\} \tag{3}$$

$$\boldsymbol{\mathcal{S}}^{\neq} = \{\delta_{\boldsymbol{x}}^{(i,j)} : \boldsymbol{f}_i(\boldsymbol{x}) = \boldsymbol{y} \oplus \boldsymbol{f}_j(\boldsymbol{x}) = \boldsymbol{y}\} \tag{4}$$

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

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 $\boldsymbol{\mathcal{S}}^{=}$, one way to measure the *Representativity* of explanations is to compute the average distance over $\boldsymbol{\mathcal{S}}^{=}$.

As a reminder, the average of $\boldsymbol{\mathcal{S}}^{=}$ corresponds to the average change of explanation when the sample is removed from the training set. This change is related to the *Representativity* 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:

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

Explanations with good *Representativity* 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 $\boldsymbol{\mathcal{S}}^{=}$, and those associated with inconsistent predictions by $\boldsymbol{\mathcal{S}}^{\neq}$. Visually, we seek to maximize the shift between the corresponding distributions for 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 are. In order to find this separation, we introduce ReCo, a statistical measure based on maximizing the 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 as the rate for which distances below a threshold from predictors with a consistent prediction among all distances below the threshold $TPR(\gamma) = \frac{|\{\delta \in \boldsymbol{\mathcal{S}}^{=} : \delta < \gamma\}|}{|\{\delta \in \boldsymbol{\mathcal{S}} \; : \; \delta < \gamma\}|}$. In a similar way, TNR denotes the rate for which distances above a threshold from predictors with opposite predictions among all the distances above the threshold $TNR(\gamma) = \frac{|\{\delta \in \boldsymbol{\mathcal{S}}^{\neq} : \delta > \gamma\}|}{|\{\delta \in \boldsymbol{\mathcal{S}} \; : \; \delta > \gamma\}|}$. Basing our measure on these rates allows us to assess the qaulity of these

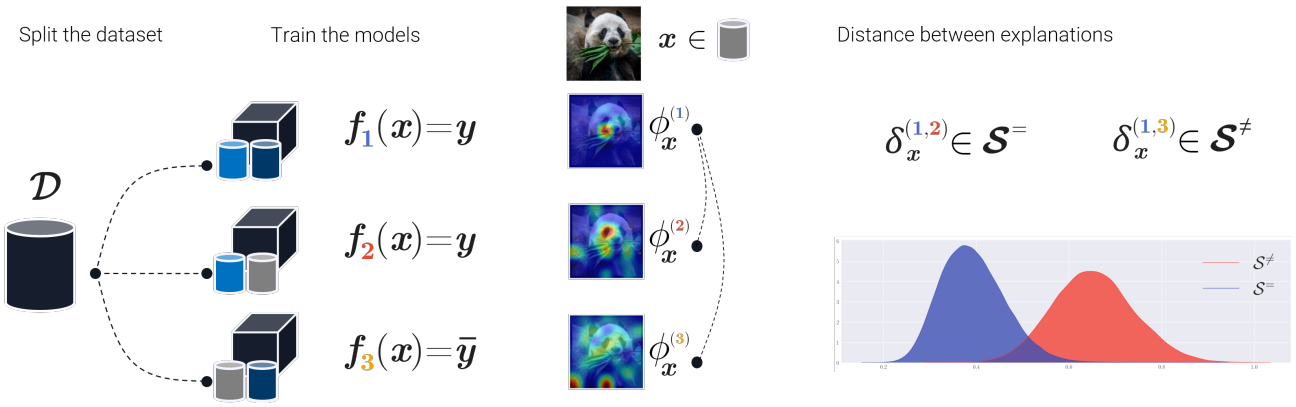

Figure 2. Application of the proposed procedure for 3 *folds*. Each predictor is trained on two of the 3 *folds*, e.g, f_1 is trained on $\mathcal{D} \setminus \mathcal{V}_1$. For a given sample x such that $x \in \mathcal{V}_1$, the explanations for each predictors are calculated ($\phi_x^{(1)}, \phi_x^{(2)}, \phi_x^{(3)}$). The distance between $\phi_x^{(1)}$ and the other two explanations $\phi_x^{(2)}, \phi_x^{(3)}$ are computed. All distances for which predictions do not contradict each other are added to $\mathcal{S}^=$ while the others are added to $\mathcal{S}^\neq$ (note that this is the case for $\delta_x^{(1,3)}$ since $f_1(x) \neq f_3(x)$).

explanations independently of the accuracy of the predictor, we define ReCo as the maximal balanced accuracy:

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

with a score of 1 indicating 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 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 ($\pm 3\%$). For ILSVRC 2012, our predictors are based on a ResNet-50 architecture [19], 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 explanations methods chosen are as follow: Saliency **(SA)** [45], Gradient $\odot$ Input **(GI)** [3], Integrated Gradients **(IG)** [51], SmoothGrad **(SG)** [47], Grad-CAM **(GC)** [41], Grad-CAM++ **(G+)** [10] and RISE **(RI)** [33]. 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 [11]: a subset of the ImageNet dataset from which we randomly selected 50 classes. The size of the images considered was 224×224.

CIFAR10 [22]: a low-resolution labeled datasets with 10 classes respectively, consisting of $60,000$ (32×32) color images.

EuroSAT [20]: a labeled dataset with 10 classes consisting of $27,000$ color images (64×64) from the Sentinel-2 satellite.

Fashion MNIST [56]: 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 [49] to compare the similarity between explanations. Several authors have provided theoretical and experimental arguments in line with this choice [16, 1, 53]. 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 re-

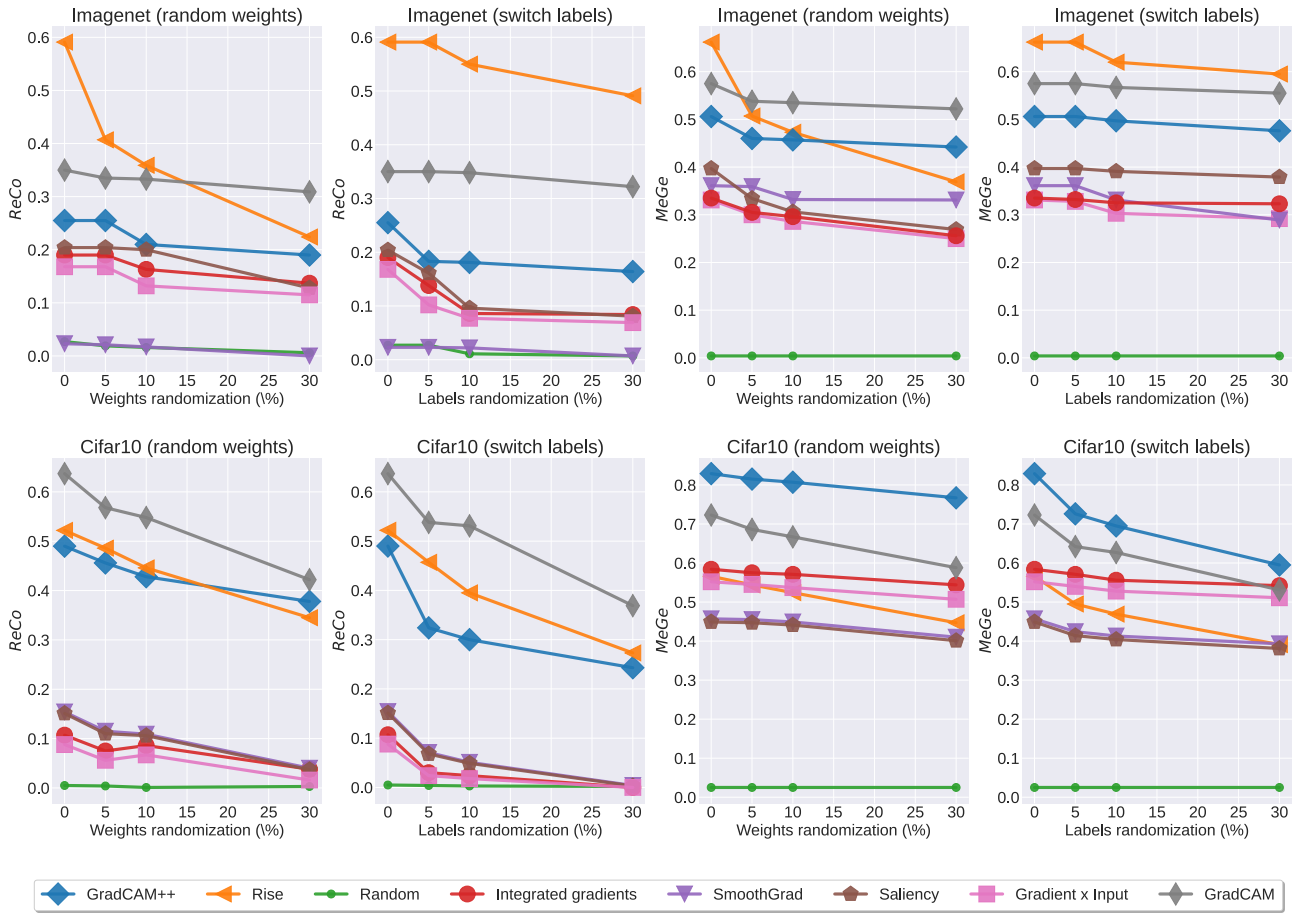

Figure 3. **MeGe and ReCo scores** for predictors trained with no degradations (first point from the left), as well as for progressively randomized predictors and predictors trained with switched labels. For all the methods tested, the more the predictor is degraded, the more the *Consistency* and *Representativity* scores drop, which means that the associated metrics pass the sanity check. **Top** ImageNet. **Bottom** Cifar-10.

spond to the problem. The distances tested were built from: 1-Wasserstein distance (the Earth mover distance from [14]), Sørensen–Dice [12] coefficient, Spearman rank correlation, SSIM [61], and ℓ_1 and ℓ_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

The first stage of our experiments consists in ensuring the reliability of the measures by performing a sanity check: on average, as the learning is degraded, we expect to see an overall increase in the number of specific and inconsistent explanations. To ensure that the metric captures these notions, we applied two different types of degradation on the predictors for each data set : randomization of weights and label inversion, with several degrees.

- Randomizing the weights, inspired by [1]. We gradually randomize 5%, 10% and 30% of the predictor 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 [31, 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 *Representativity* of the explanations, which is related to the ability of the predictor to derive general strategies. Thus, the destruction of the parameters of a predictor directly degrades these strategies. The figure 3 allows us to reveal this impact which is materialized by the correlation of the measures with the degradation intensity: MeGe and ReCo capture the degradation of the explanation and pass the sanity check.

We notice 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, 46]. It was subsequently observed that this sensitivity seems to translate into a better *Fidelity* score for the methods.

4.3. The Implications of the Fidelity Metric

To mark the difference between the proposed measures and the *Fidelity*, we applied the μF 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 μF, 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.

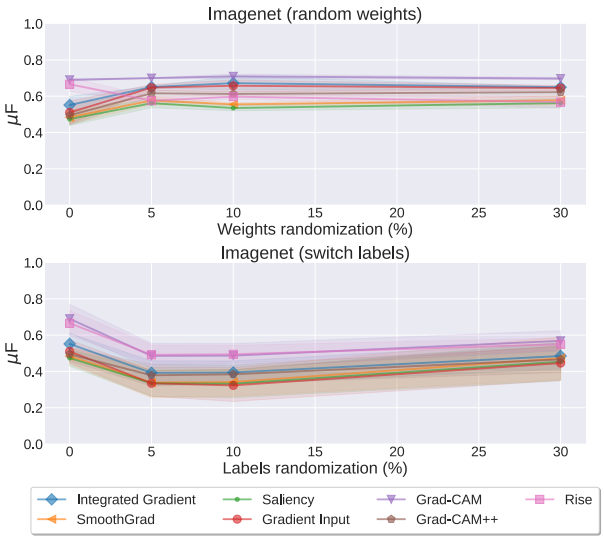

Figure 4. *Fidelity* scores (Equation 8) on ImageNet for normally ResNet-50 predictors (first point on the left) as well as for progressively randomized predictors and predictors trained with switched labels. Even a strong degradation of the predictor does not impact the *Fidelity* of the tested methods. Hence, 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 Figure 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* measure 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 an 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.

ImageNet	SA	GI	IG	SG	GC	G+	RI
μF	0.47	0.51	0.55	0.48	**0.69**	0.49	<u>0.67</u>
MeGe	0.40	0.50	<u>0.58</u>	0.36	0.34	0.33	**0.66**
ReCo	0.20	0.17	0.16	0.02	<u>0.35</u>	0.26	**0.59**

Table 1. *Consistency*, *Representativity* and *Fidelity* score for ResNet-50 models on ImageNet. Higher is better. The first and second best results are respectively in **bold** and <u>underlined</u>.

Table 1 reports the *Fidelity* (μF), *Consistency* (ReCo) and *Representativity* (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 *Representativity* and the *Consistency* of the explanations. This score can then be used as a criterion to separate RIS from Grad-CAM. In view of the differences between the MeGe and ReCo scores, RISE method seems preferable.

Concerning the *Representativity* 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 eliminates 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 [54, 39, 32] and have good generalizability. As a reminder, a $\boldsymbol{f}$ function is called L-Lipschitz, with $L \in \mathbb{R}^+$

if

$$|\boldsymbol{f}(\boldsymbol{x}_1) - \boldsymbol{f}(\boldsymbol{x}_2)| \leq L|\boldsymbol{x}_1 - \boldsymbol{x}_2| \tag{7}$$

For every pair $(\boldsymbol{x}_1, \boldsymbol{x}_2) \in \mathcal{X}^2$. The smallest of these L is called the Lipschitz constant of $\boldsymbol{f}$. 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 generalisation bounds that critically rely on the Lipschitz constant of the neural network [55, 31, 5].

The predictors were trained using the Deel-Lip library [43]. All the predictors have comparable accuracy ($78 \pm 4\%$). To our knowledge, no previous work has made the link between Lipschitz networks and the chosen explainability methods.

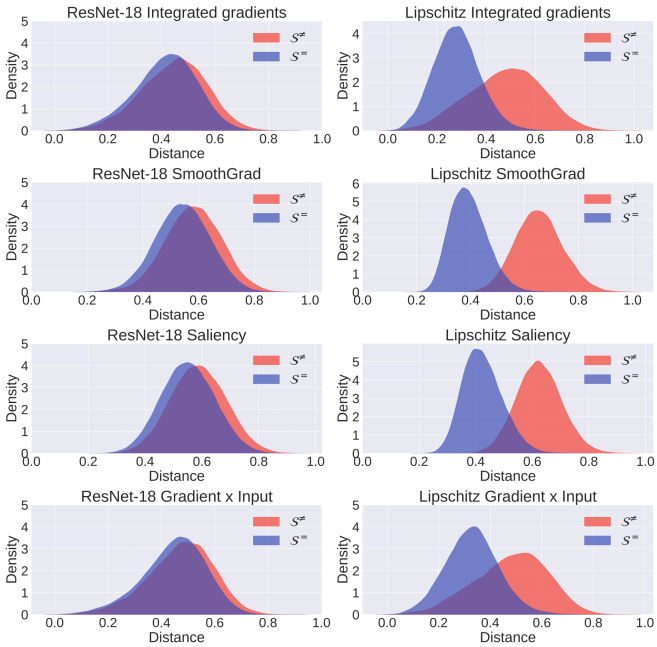

Figure 5. Lipschitz predictors (right column) on Cifar10. As explained in this paper, a clear separation between the $\boldsymbol{\mathcal{S}}^{=}$ and $\boldsymbol{\mathcal{S}}^{\neq}$ histograms is a sign of consistent explanations.

The Figure 5 shows the difference in $\boldsymbol{\mathcal{S}}^{\neq}$ and $\boldsymbol{\mathcal{S}}^{=}$ between ResNet and 1-Lipschitz predictors. In the left column, the results come from ResNet-18 predictors normally 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 a better separation of the histograms and a smaller expectation of $\boldsymbol{\mathcal{S}}^{=}$. SmoothGrad is the method that obtains the most consistent explanations as indicated in the table 3, in front of Saliency and Grad-CAM (more results in the supplementary material G Figure 10).

Concerning MeGe, the results reported in Table 2 show an improvement in the *Representativity* of the explanations for the 1-Lipschitz predictors. Indeed, the *Representativity* score has increased compared to the ResNet predictors for all tested methods, except Grad-CAM++.

MeGe	IG	SG	SA	GI	GC	G+	RI
ResNet-18	0.58	0.46	0.45	0.55	0.72	**0.83**	0.57
1-Lipschitz	**0.72**	**0.60**	**0.58**	**0.67**	**0.75**	0.54	**0.85**

Table 2. MeGe scores obtained by 1-Lipschitz models and ResNet-18 models on Cifar10. Higher is better. For almost all methods, the *Representativity* of explanations increases significantly on 1-Lipschitz models.

ReCo	IG	SG	SA	GI	GC	G+	RI
ResNet-18	0.11	0.15	0.15	0.09	0.64	**0.49**	0.52
1-Lipschitz	**0.60**	**0.90**	**0.81**	**0.50**	**0.67**	0.24	**0.84**

Table 3. ReCo scores obtained by 1-Lipschitz models and ResNet-18 models on Cifar10. Higher is better. For almost all methods, the *Consistency* of explanations increases significantly on 1-Lipschitz models.

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 $\boldsymbol{\mathcal{S}}^{=}$ and $\boldsymbol{\mathcal{S}}^{\neq}$ in Figure 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: *Representativity* and *Consistency*. The procedure requires access to the training algorithm and training data, which covers a wide range of use cases, especially in industrial applications. 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. Conversely, we conducted several experimental sanity checks to ensure the proposed measures capture the notion of *Representativity* and *Consistency*. In addition, we showed that it is possible to use these measures as criteria for selecting a explanation method in conjunction with the fidelity metric. 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 the class of networks

give much more coherent and representative explanations compared to alternative models.

Although our analyses have focused on convolutional neural networks, the approach and measures we described are general enough and are broadly applicable to any machine learning models (including as Decision Trees, GANs, etc). 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 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

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

[9] 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.

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

[10] 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.

[11] 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.

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

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

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

[15] 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.

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

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

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

[23] 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.

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

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

[26] Drew Linsley, Dan Shiebler, Sven Eberhardt, and Thomas Serre. Learning what and where to attend. 2019.

[27] 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.

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

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

[30] 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.

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

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

[33] 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.

[34] 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.

[35] 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.

[36] 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.

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

[38] 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.

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

[40] 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.

[41] 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.

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

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

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

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

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

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

[52] 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.

[53] 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.

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

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

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

[57] 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.

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

[59] 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.

[60] 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.

[61] 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.

[62] 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 $\mathcal{S}^{=}$ and $\mathcal{S}^{\neq}$

Require: $k \in \mathbb{N}_{\geq 2}$, $\mathcal{D} = \{\mathcal{V}_i\}_{i=1}^{k}$
  $\mathcal{S}^{=} \leftarrow \{\}, \mathcal{S}^{\neq} \leftarrow \{\}$
  for all $i \in \{1, \ldots, k\}$ **do**
    Train $\boldsymbol{f}_i$ on $\mathcal{D} \setminus \mathcal{V}_i$
    for all $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{D}$ **do**
      // generate explanations on all dataset
      $\boldsymbol{\phi}_{\boldsymbol{x}}^{(i)} \leftarrow \boldsymbol{\Phi}(\boldsymbol{f}_i, \boldsymbol{x})$
    end for
  end for
  for all $i \in \{1, \ldots, k\}$ **do**
    for all $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{V}_i$ **do**
      for all $j \in \{1, \ldots, k \mid i \neq j\}$ **do**
        // $\boldsymbol{f}_j$ was trained on $\boldsymbol{x}$, $\boldsymbol{f}_i$ was not
        $\delta_{\boldsymbol{x}}^{(i,j)} \leftarrow d(\boldsymbol{\phi}_{\boldsymbol{x}}^{(i)}, \boldsymbol{\phi}_{\boldsymbol{x}}^{(j)})$
        if $\boldsymbol{f}_i(\boldsymbol{x}) = \boldsymbol{y}$ **and** $\boldsymbol{f}_j(\boldsymbol{x}) = \boldsymbol{y}$ **then**
          // both model are correct
          $\mathcal{S}^{=} \leftarrow \mathcal{S}^{=} \cup \{\delta_{\boldsymbol{x}}^{(i,j)}\}$
        else if $\boldsymbol{f}_i(\boldsymbol{x}) = \boldsymbol{y}$ **or** $\boldsymbol{f}_j(\boldsymbol{x}) = \boldsymbol{y}$ **then**
          // only one model is correct
          $\mathcal{S}^{\neq} \leftarrow \mathcal{S}^{\neq} \cup \{\delta_{\boldsymbol{x}}^{(i,j)}\}$
        end if
      end for
    end for
  end for
  Return $\mathcal{S}^{=}, \mathcal{S}^{\neq}$

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 $\boldsymbol{f} : \mathbb{R}^d \rightarrow \mathbb{R}^C$ where C is the number of classes. We assume $\boldsymbol{f}_c(\boldsymbol{x})$ the logit score (before softmax) for class c. An explanation method provides an attribution $\boldsymbol{\phi} \in \mathbb{R}^d$ 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.

$$\boldsymbol{\Phi}^{SA}(\boldsymbol{x}) = \Big|\frac{\partial \boldsymbol{f}_c(\boldsymbol{x})}{\partial \boldsymbol{x}}\Big|$$

Gradient $\odot$ 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 $\odot$ Input is equivalent to ϵ-LRP and DeepLIFT methods under certain conditions: using a baseline of zero, and with all biases to zero.

$$\boldsymbol{\Phi}^{GI}(\boldsymbol{x}) = \boldsymbol{x} \odot \Big|\frac{\partial \boldsymbol{f}_c(\boldsymbol{x})}{\partial \boldsymbol{x}}\Big|$$

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 [48] for a comparison). The final result depends on both the choice of the baseline $\boldsymbol{x}_0$ and the number of points to estimate the integral. In the context of these experiments, we use zero as the baseline and $m = 60$.

$$\boldsymbol{\Phi}^{IG}(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{x}_0) \int_0^1 \frac{\partial \boldsymbol{f}_c(\boldsymbol{x}_0 + \alpha(\boldsymbol{x} - \boldsymbol{x}_0))}{\partial \boldsymbol{x}} d\alpha$$

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 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.

$$\boldsymbol{\Phi}^{SG}(\boldsymbol{x}) = \mathop{\mathbb{E}}_{\varepsilon \sim \mathcal{N}(0, \boldsymbol{I}\sigma^2)} \Big[\frac{\partial \boldsymbol{f}_c(\boldsymbol{x} + \varepsilon)}{\partial \boldsymbol{x}}\Big]$$

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

$$\boldsymbol{\Phi}^{GC} = max(0, \sum_k \alpha_c^{(k)} \boldsymbol{A}^{(k)})$$

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. The

weights $\alpha_c^{(k)}$ associated to each feature map is computed as follow :

$$\alpha_k^c = \sum_i \sum_j [\frac{\frac{\partial^2 \boldsymbol{f}_c(\boldsymbol{x})}{(\partial A_{ij}^{(k)})^2}}{2\frac{\partial^2 \boldsymbol{f}_c(\boldsymbol{x})}{(\partial A_{ij}^{(k)})^2} + \sum_i \sum_j A_{ij}^{(k)} \frac{\partial^3 \boldsymbol{f}_c(\boldsymbol{x})}{(\partial A_{ij}^{(k)})^3}}]$$

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 $\boldsymbol{m} \sim \mathcal{M}$ 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 $m = 1000$.

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

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 [38, 57, 36, 33]. In this work, we use μF from [6]:

$$\mu F = \underset{\substack{S \subseteq \{1,\dots,d\} \\ |S|=k}}{\text{Corr}} \left(\sum_{i \in S} \boldsymbol{\Phi}(\boldsymbol{f}, \boldsymbol{x})_i, \boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x}_{[\boldsymbol{x}_i = \bar{\boldsymbol{x}}_i, i \in S]}) \right) \tag{8}$$

Where $\boldsymbol{f}$ is a predictor, $\boldsymbol{\Phi}$ an explanation function, S a subset indices of $\boldsymbol{x}$ and $\bar{\boldsymbol{x}}$ a baseline reference. The choice of a proper baseline is still an active area of research [50].

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 $\boldsymbol{\mathcal{S}}^{=}$ and $\boldsymbol{\mathcal{S}}^{\neq}$ 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 :

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

Table 4. 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)

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 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 (92%, $\pm$1%)
EuroSAT base filters 46, Dropout 0.25 (95%, $\pm$1%)
Cifar10 base filters 32, Dropout 0.25 (78%, $\pm$4%)
ImageNet ResNet50 (88%, $\pm$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 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 6).

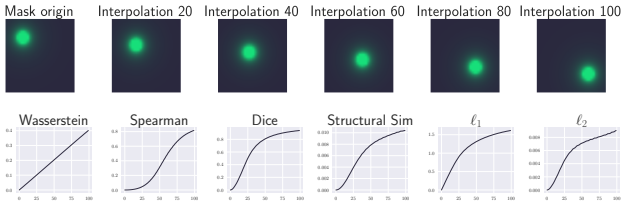

Figure 6. Distances with moving interest point. The first line shows the successive interpolations between the baseline image (left), and the target image (right). The second line shows the evolution of the distance between each interpolation and the baseline image.

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.

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

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.

Over the different distances tested, they all pass the sanity test : there is a monotonous positive correlation (as seen in Figure 7). 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.

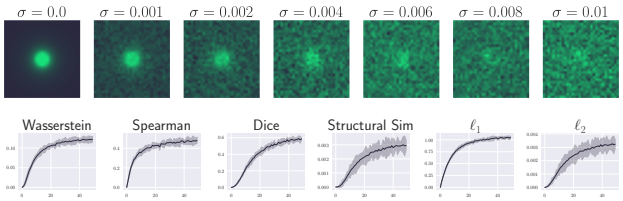

Figure 7. Distances with noisy images. The first line shows original noise-free image (left) and noisy copies computed by increasing σ. The second line shows the distances between each noisy image and the baseline image.

G. Additional results

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

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

Metrics	IG	SG	SA	GI	GC	G+	RI
MeGe	0.40	<u>0.42</u>	0.41	0.41	**0.67**	**0.67**	0.39
ReCo	0.31	0.18	0.18	0.23	<u>0.59</u>	**0.64**	0.34

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

Metrics	IG	SG	SA	GI	GC	G+	RI
MeGe	**0.90**	0.36	0.30	**0.90**	0.77	<u>0.84</u>	0.52
ReCo	<u>0.37</u>	0.13	0.10	<u>0.37</u>	**0.52**	0.32	<u>0.37</u>

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

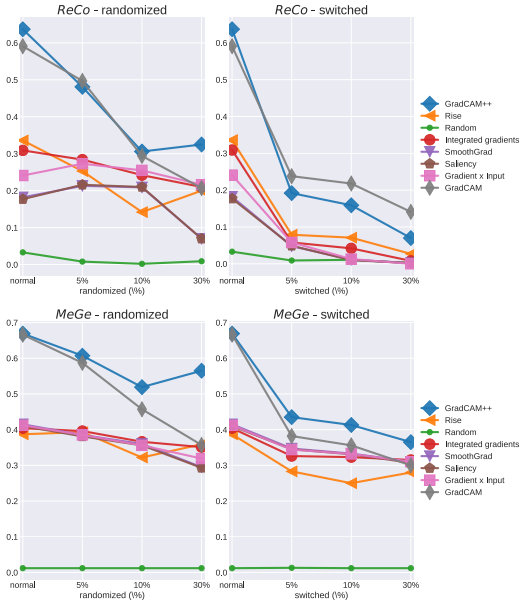

Figure 8. **Eurosat** MeGe and ReCo scores for normally trained models (first point from the left), as well as for progressively randomized models and models trained with switched labels.

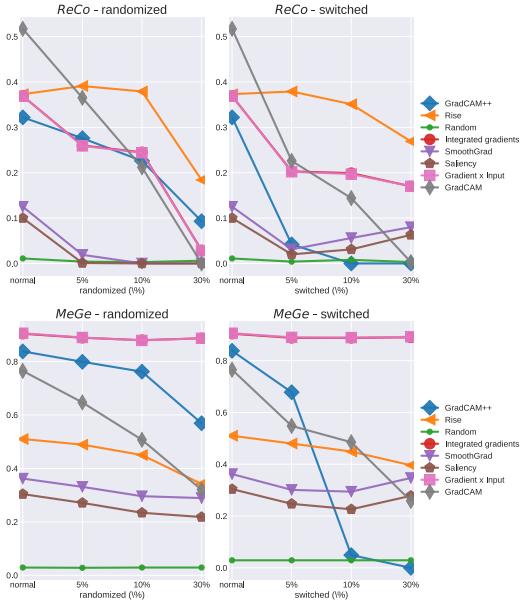

Figure 9. **Fashion-MNIST** MeGe and ReCo scores for normally trained models (first point from the left), as well as for progressively randomized models and models trained with switched labels.

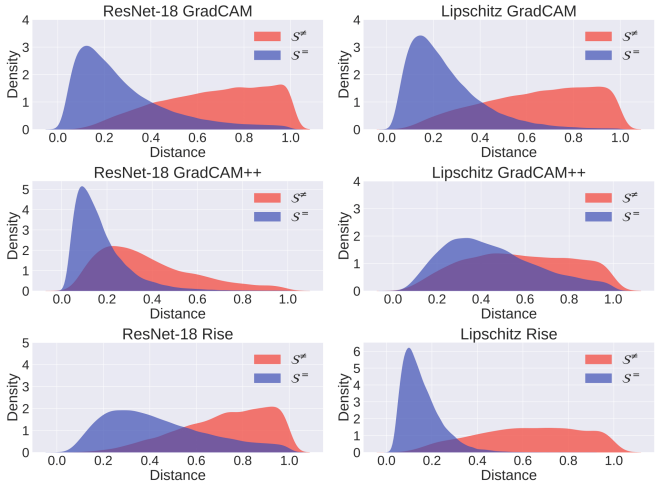

Figure 10. $\mathcal{S}^{=}$ and $\mathcal{S}^{\neq}$ for ResNet (left column) and 1-Lipschitz models (right column) on Cifar10. As explained in this paper, a clear separation between the $\mathcal{S}^{=}$ and $\mathcal{S}^{\neq}$ histograms is a sign of consistent explanations.