From Flexibility to Manipulation: The Slippery Slope of XAI Evaluation

Kristoffer Wickstrøm[1], Marina Höhne[2,3,6], and Anna Hedström[2,4,5]

[1] Department of Physics and Technology, UiT The Arctic University of Norway
kwi030@uit.no
[2] UMI Lab, Leibniz Institute of Agricultural Engineering and Bioeconomy e.V.
(ATB)
[3] Department of Computer Science, University of Potsdam
MHoehne@atm-potsdam.de
[4] Department of Electrical Engineering and Computer Science, TU Berlin
[5] Department of Artificial Intelligence, Fraunhofer HHI, Berlin, Germany
[6] BIFOLD - Berlin Institute for the Foundations of Learning and Data

Abstract. The lack of ground truth explanation labels is a fundamental challenge for quantitative evaluation in explainable artificial intelligence (XAI). This challenge becomes especially problematic when evaluation methods have numerous hyperparameters that must be specified by the user, as there is no ground truth to determine an optimal hyperparameter selection. It is typically not feasible to do an exhaustive search of hyperparameters so researchers typically make a normative choice based on similar studies in the literature, which provides great flexibility for the user. In this work, we illustrate how this flexibility can be exploited to manipulate the evaluation outcome. We frame this manipulation as an adversarial attack on the evaluation where seemingly innocent changes in hyperparameter setting significantly influence the evaluation outcome. We demonstrate the effectiveness of our manipulation across several datasets with large changes in evaluation outcomes across several explanation methods and models. Lastly, we propose a mitigation strategy based on ranking across hyperparameters that aims to provide robustness towards such manipulation. This work highlights the difficulty of conducting reliable XAI evaluation and emphasizes the importance of a holistic and transparent approach to evaluation in XAI. Code is available at https://github.com/Wickstrom/quantitative-xai-manipulation.

Keywords: Explainablity · Reproducibility · Reliability · Faithfulness

1 Introduction

Explainable artificial intelligence (XAI) is a crucial research area to ensure trustworthiness in computer vision [44], which contains a wide range of methods that provide explanations for the output of a predictive model [7,38,53]. To determine which XAI method is suitable for a given problem setting, quantitative evaluation analysis is necessary to provide an objective measurement for comparison. Such

XAI method	Faithfulness score ($\downarrow$)
LRP	25.19
Saliency	**20.23**
Kernel SHAP	23.94

XAI method	Faithfulness score ($\downarrow$)
LRP	**19.31**
Saliency	22.96
Kernel SHAP	24.87

Table 1: Faithfulness comparison of XAI methods on MNIST before (left table) and after manipulation (right). Here, the different between the left and right table is the perturbation methods used (uniform noise vs. blurring, respectively). Both perturbation methods are commonly used, but completely change the outcome of the evaluation.

quantitative analysis of XAI methods has made great leaps forward over the last couple of years [2, 26], and generally consists of evaluating several metrics that measure desirable properties that an XAI method should have i.e., *metric-based quality estimation* [24]. However, the progress in XAI and its evaluation has led to an overwhelming variety of methods and metrics, making it challenging for researchers to navigate their choices [12, 24, 25, 33].

A fundamental limitation in XAI evaluation is the lack of ground truth explanation labels [24]. Since such information is generally not available, we approximate explanation quality by measuring desirable properties like faithfulness [4, 9, 17, 40, 41, 43], complexity [9, 16, 39], or robustness [3, 17, 37, 55] and translate these properties into empirical tests . In this translation, a challenge appears in the parameterization of the empirical tests. For example, how do we mask out pixels and how large should the masks be? Preliminary works [24, 33] have shown that the evaluation outcomes are sensitive to choices like these. This sensitivity underscores the need to investigate the impact of hyperparameter choice, making it an important research area to ensure the reliability of XAI evaluations.

This challenge becomes particularly prominent for evaluation methods with many hyperparameters that must be set, since it is generally not possible to find an objective measure of the optimal set of hyperparameters. For instance, faithfulness evaluations view model behavior changes as signals of explanations quality, with substantial changes reflecting the explanations faithfulness [4, 7, 9, 17, 40, 43]. This type of evaluation often requires replacing pixel values with some baseline value, which can be highly data-dependent and difficult to tune [46]. Furthermore, it can often be computationally impractical to evaluate all possible choices for hyperparameters. Therefore, hyperparameters are usually selected normatively with the researcher's own subjective judgment, frequently drawing on prior studies. Since there is variation in what hyperparameters are being used in the community [24, 33], there is some flexibility in selecting hyperparameters from an acceptable of possible choices.

In this work, we demonstrate how this flexibility of XAI evaluation can be exploited to manipulate the evaluation outcome. By making seemingly small changes to hyperparameters that are widely used in the literature, the outcome of the faithfulness evaluation can change completely. Tab. 1 illustrates this, where

standard XAI methods are compared with only slight changes in hyperparameters but with significant changes in evaluation outcome. We propose to frame the finding of these small changes as an optimization problem that manipulates the evaluation, where either the evaluation of a single XAI method is manipulated or the evaluation of multiple XAI methods is manipulated jointly. Our contributions are:

C1 A method-specific manipulation method that can increase the evaluation score for a specific XAI method, which we entitle *intra-manipulation*.

C2 A holistic manipulation method that can manipulate the quantitative comparison of several XAI methods, which we entitle *inter-manipulation*.

C3 A comprehensive experimental analysis on manipulation of faithfulness evaluation that demonstrate how the evaluation outcome can completely change after manipulation.

C4 Towards improving the robustness of quantitative evaluation of XAI, we propose Mean Resilience Rank, a ranking-based procedure that reduces the sensitivity to hyperparameter manipulation.

Our findings have significant implications for the XAI community. Quantitative evaluation is crucial to provide objective measurements of explanation quality, which can be used to select an appropriate method for a particular task or for comparison in method development. If these measurements can be easily altered, it reduces the trustworthiness of both method selection and comparison. Therefore, the findings and solutions in this work are of critical importance for the community both by highlighting the issue of manipulation and by presenting strategies towards mitigating the issue.

2 Related Work

Metric-based Quality Estimation Quantitative analysis of XAI explanation has improved considerably in recent years, and researchers now have a vast amount of evaluation metrics at their disposal [2,26]. Due to the lack of ground truth explanations, researchers try to quantify the quality of an explanation by measuring desirable properties, which can be categorized into 6 families of properties [26]; faithfulness [10], robustness [3], localisation [51], complexity [16], randomisation [1], and axiomatic [30]. Within each family, a variety of metrics exists.

Prior Studies on Hyperparameter Sensitivity in XAI Increasing attention has been given to the influence and potential confounding effects of hyperparameters in XAI evaluations [24]. These studies vary in defining dependent versus independent variables and the hyperparameter space of intervention, be it model, explanation, or evaluation space. Studies have examined the sensitivity of attribution methods to explanation hyperparameters like random seed and number of samples [8], and the impact of baseline choices in methods like Integrated Gradients on explanation outcomes [46, 49]. Additionally, the sensitivity of explanation outcomes concerning model performance variables such as optimizer,

activation function, learning rate, and dataset split has been studied [28], along with the effects of model priors and random weight initialization on explanations and evaluations [22]. Disagreement among different explanation methods regarding top-K features and ranking has also been investigated [33], while analyzing the impact of baselines [31].

Recently, researchers have explored how evaluation parameters affect outcomes, including the sensitivity of randomisation metrics to hyperparameters like normalisation, randomisation order, and similarity measures [11,25,48]. Faithfulness metrics have been examined for hyperparameter influences such as baseline choice and perturbation order [12–14,20,36,42,43,52]. Unlike existing work, inspired by adversarial machine learning, we introduce a novel, general-purpose manipulation approach, applicable across a variety of evaluation approaches. Our findings reveal that faithfulness evaluation outcomes are highly susceptible to manipulation. This is a key issue for the XAI community to address. We put forward a preliminary mitigating solution for this in Sec. 6.

3   Preliminaries

For clarity, we present the core concepts and notation used in the work.

Local explanations Let the input to a black-box classifier f be denoted as $\mathbf{x} \in \mathbb{R}^d$ and the output of the classifier as $f(\mathbf{x}) = \hat{y}$. Local explanation methods [7,15,45,50] interpret the decision of f by attributing an importance score to each component of $\mathbf{x}$. We denote the explanation of f for a given class y as $\mathbf{e} \in \mathbb{R}^d$.

Evaluating Explanations Here, we present a generalized formulation of quantitative XAI evaluation to illustrate the static input parameters and adjustable hyperparameters. We assuming an evaluation function $F \to \mathbb{R}$ on the form:

$$F(f, \mathbf{x}, \mathbf{e}, a, b, c) = s. \tag{1}$$

Here, f, $\mathbf{x}$, and $\mathbf{e}$ are input parameters provided by the user, while a, b, and c are hyperparameters that must be determined by the user. The output of the evaluation is represented by s, which is a scalar indicating the performance of the particular explanation. Here, we keep the hyperparameters a, b, and c completely general for the sake of clarity. But note that there could be more or less hyperparameters and they can take many different forms (e.g. a number or a function), depending on the particular test and the data in questions.

4   Manipulating XAI Evaluation

Here, we introduce our manipulation strategies for changing the evaluation outcome of XAI evaluation with only small hyperparameter alterations. The motivation for this approach is that there often exists several agreed-upon hyperparameters for a given XAI evaluation method. For instance, when conducting a

faithfulness evaluation [4, 7, 9, 17, 40, 43] (see Sec. 5 for further details), an important component is perturbing input pixels. There exist numerous methods for conducting this perturbation, and it is known that selecting a suitable one can be challenging [12, 42, 46]. However, evaluating numerous such methods can be highly computationally demanding, and due to the lack of ground truth explanations we cannot decide which method is correct. Therefore, in practice, it is common to consider only a single perturbation method [3, 10, 37]. However, as we have shown in Tab. 1, even a slight change in the hyperparameter setting can have a big impact on the evaluation. Those who are aware of this sensitivity can potentially exploit it, which is the motivation for our manipulation strategy.

Intra-manipulation We propose two ways to manipulate XAI evaluation methods. First, we propose to focus on manipulating the evaluation outcome for a single XAI method, which we refer to as *intra-manipulation* and is defined as:

Definition 1 (Intra-Manipulation). *Given an evaluation function F, an input sample $\mathbf{x}$, an explanation $\mathbf{e}$, hyperparameters a, b, and c, and a feasible set of hyperparameters A_a^* for the hyperparameter a, the intra-manipulation method solves the following optimization problem to determine the hyperparameter a, which maximizes the evaluation score of F:*

$$\underset{a}{maximize} \quad F(f, \mathbf{x}, \mathbf{e}, a, b, c)$$
$$subject\ to \quad a \in A_a^*.$$

Definition 1 defines an optimization problem where the goal is to find hyperparameters that maximize the evaluation outcome, but are constrained to lie within a feasible set of values (A_a^* in this case) for the hyperparameters in questions. Determining this feasible set requires a researcher's judgment and a good understanding of the particular XAI evaluation method that the user wants to manipulate. But more deeply, it fundamentally depends on the model: i.e. the feasible set is and should be dependent on the learned functional response of the model. In Sec. 5, we further explain how to determine the feasible set. If the feasible set is large, Definition 1 can be solved through suitable optimization techniques. If the feasible set if small, an exhaustive search can be performed. Also note that Definition 1 can be extended to optimize across several hyperparameters, e.g. maximizing both a and b.

Inter-manipulation Definition 1 allows for improving the evaluation outcome of a single XAI method. But in many cases it could be desirable to alter the outcome of the evaluation of several XAI methods. Our second manipulation approach is to take a holistic view and manipulate the evaluation of several XAI methods jointly. We refer to this approach as *inter-manipulation* and define it as:

Definition 2 (Inter-Manipulation). *Given an evaluation function F, an input sample $\mathbf{x}$, a set of explanations $\{\mathbf{e}_1, \cdots, \mathbf{e}_M\}$ from M different XAI methods, hyperparameters a, b, and c, and a feasible set of hyperparameters A_a^* for the*

hyperparameter a, the inter-manipulation method solves the following optimization problem to determine the hyperparameter a, which maximizes the following objective:

$$\underset{a}{maximize} \quad F(f, \mathbf{x}, \mathbf{e}_m, a, b, c) - \sum_{m' \neq m} F(f, \mathbf{x}, \mathbf{e}_{m'}, a, b, c)$$
$$subject\ to \quad a \in A_a^*$$

Here, $\mathbf{e}_m$ is the explanation from the XAI method we wish to improve the performance of. We entitled this method *the focus method.* The explanation from a *non-focus method* is denoted as $\mathbf{e}'_m$, which we seek to worsen the performance of. The optimization problem presented in Definition 2 is more complex compared to Definition 1 due to the interplay between the different XAI methods. For example, the optimal solution could be found by a combination of increasing the performance of the focus-method while simultaneously decreasing the performance of the *non-focus methods.* Similarly, as Definition 1, the optimization problem can be solved in several ways (e.g. Bayesian optimization) and can be extended to include several hyperparameters.

5   Manipulating Faithfulness Evaluation

Some types of XAI evaluation methods are more susceptible to manipulation than others. For instance, localization metrics, which aims to measure if an explanation is within a region-of-interest, usually only have 1 or even 0 hyperparameters to select [5,51] and are therefore harder to manipulate. On the other hand, faithfulness metrics [3,10,39] have at least 3 hyperparameters that must be determined, and often more. This is one of the most popular evaluation methods in XAI [4, 7, 9, 17, 40, 43] and is therefore an important evaluation category to study. Therefore, we will focus on manipulating faithfulness metrics. The following section provides an overview of the fundamental components in faithfulness evaluation.

The fundamental components of faithfulness Faithfulness measures to what extent explanations follow the predictive behavior of the model by iteratively perturbing the input and monitoring the corresponding change in the output of the model. Our focus will be on the task of classification, since this is the most common setting in the context of explainability and vision. This section presents the mathematical formulation of the general components of most faithfulness metrics. Let S denote the set of indices $\{1, \cdots, d\}$ for each element in the input sample $\mathbf{x} \in \mathbb{R}^d$. Partition S into K sets $S_1, \cdots, S_K$ of equal cardinality C and arranged such that:

$$\sum_{i \in S_1} e_i \geq \cdots \geq \sum_{i \in S_K} e_i. \tag{2}$$

For convenient notation, we define the sum of attributions for one partition as:

$$\tilde{e}_{S_k} = \sum_{i \in S_k} e_i \tag{3}$$

Inequality (2) instructs us to rank the indices according to the input features with highest importance in a descending fashion, and are used to iteratively perturb the input. Note that some metrics sort the indices in an ascending fashion [4, 6, 39] and some perturb the input randomly [9], but the general approach in faithfulness metrics is to perturb the inputs according to Equation (2) [3, 40, 42, 43]. Let $\mathbf{x}_{S_1}$ denote a perturbed version of $\mathbf{x}$, where all x_i for $i \in S_1$ are replaced by some baseline perturbation function g_p. We denote the output of the classifier based on $\mathbf{x}_{S_1}$ as $\hat{y}_{S_1}$. For $\mathbf{x}_{S_2}$, all x_i for $i \in S_1 \cup S_2$ are perturbed. In general, $\mathbf{x}_{S_i}$ will have all have the indices in all sets up to set S_i replaced by the baseline perturbation function.
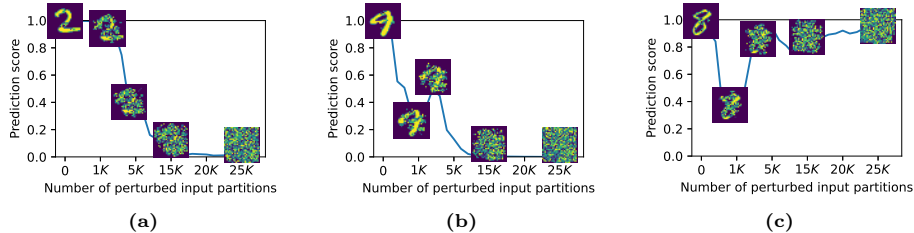

Fig. 1: Example of possible faithfulness curves for digit classification. The leftmost curve illustrates how an "intuitive" faithfulness curve might look, while the remaining curves show that there is a lot of variation in how these curves can appear.

Illustrating the faithfulness curve Based on the K partitions of S, a set of progressively more perturbed inputs can be created, i.e. $\{\mathbf{x}_{S_1}, \cdots, \mathbf{x}_{S_K}\}$. Each of the perturbed inputs are classified, which gives a set of model outputs $\{\hat{y}_{S_1}, \cdots, \hat{y}_{S_K}\}$. These model outputs are the fundamental components for faithfulness evaluation in XAI. The rationale is that a good explanation should remove the essential parts of an input first, which should lead to a steep drop in the classification score. A poor explanation will remove parts that are not important, which will allow the classification score to stay high. Figure 1a shows an example where the classifier behaves as expected, with a sharp drop in accuracy when the important parts of the input are removed. To compare two explanations, one can inspect a plot such as in Figure 1a and see which explanation has the sharpest drop in classification score. However, such a visual approach has many limitations. First, we generally would like to compare explanations across many samples to get a reliable estimate of how they perform. Inspecting numerous such plots is cumbersome, and the curves can look different for different visual objects in classification, which makes comparison challenging. Also, real-world data is not

always as well-behaved as the plot shown in Figure 1a, as illustrated in Figures 1b and 1c. Another important aspect is that ensuring that the curve is a genuine depiction of explanation quality and not out-of-distribution (OOD) response of the model can be highly challenging [27, 42].

5.1   Hyperparameters in Faithfulness Metrics

Here, we briefly describe the different hyperparameters that must be determined by the user to conduct a faithfulness evaluation. It is important to note that many of these hyperparameters are inherently data dependent, which means that the user must re-parameterize each metric for their use case, making the results non-comparable across different datasets and potentially models.

Size of partition The size of each partition determines how many features are removed and replaced in each step of the faithfulness curve. To determine this size, there are several considerations. First, if the size of the partition is very small the evaluation will quickly become computationally infeasible, since the number of forward passes for each sample increases. Furthermore, removing only a single or a few pixels at a time can lead to adversarial effects [47]. Second, a large partition size will lead to course faithfulness curves which makes comparison between curves challenging. Therefore, there is a trade-off between computational efficiency and resolution of the faithfulness curves. Some researchers use the height and width of the image (assuming square images) as the size of the partition [26], but other choices are also common [7, 53].

Perturbation Function When a set of features are removed from an image, they are replaced by some perturbation function. An example of such a perturbation function could be Gaussian noise or setting pixel values to zero [3, 37], but more advanced approaches are also available [41]. The type of perturbation function to apply is highly dependent on the type of images that are being considered. For example, replacing pixels with a value of zero can be possible for natural images [21] but would not be a suitable choice for images with a black background, since this could potentially not induce a change in the network's output. In general, the choice of perturbation function varies greatly between papers [3, 10, 37, 41].

Aggregation Function Examples in Figures 1b and 1c, demonstrate that it can be difficult to assess which explanation is superior. Therefore, it is desirable to aggregate the perturbed model outputs into a single score that can be easily used for comparison using an aggregation function g_a. There are two main approaches to aggregate the curves shown in Figure 1. The first approach is to calculate the AUC of the faithfulness curve [7, 42, 43]. A low AUC is considered desirable, since it indicates that the important components of an input are removed first. The second approach is to correlate the model outputs with the sum of attributions within each partition [4, 9]. The motivation for this approach is that when important parts of an object are removed the predictive performance should gradually decrease, which will be captured by the correlation functions. Both correlation and AUC are used regularly in the literature [3, 7, 10, 37, 40].

Normalization Function Attributions produced by different XAI methods can have a widely different range of values. Therefore, it can be necessary to normalize the attributions such that they are comparable across different methods. A simple choice could be to standardize using the mean and standard deviation of the attributions. But choices such as these can influence evaluations [24] and more sophisticated normalization schemes are also used [11].

6 Towards More Reliable Quantitative Evaluation with Mean Resilience Rank

Due to the lack of ground truth explanations, we cannot determine what setting of hyperparameters constitutes the "correct" choice. However, we do know that it is desirable to perform well across all hyperparameter settings. Therefore, if an XAI method consistently appears among the highest-ranked methods across numerous hyperparameters, it provides an indication of high quality with less sensitivity to hyperparameters. Thus, to provide robustness towards hyperparameter manipulation, we propose to rank each XAI method for each hyperparameter setting in the feasible set, and average the ranking across the entire set. We will refer to this ranking-approach as Mean Resilience Rank (MRR).

Here, we describe mathematically how to perform this ranking. First, assume we want to evaluate M explanation methods, and that we only have a single hyperparameter a with a feasible set of values A_a^* that can be altered. We denote one element of A_a^* as a_i, such that the evaluation outcome for all M XAI methods can be collected in the set:

$$S_F(a_i) = \{F(f, \mathbf{x}, \mathbf{e}_1, a_i, b, c), \cdots, F(f, \mathbf{x}, \mathbf{e}_M, a_i, b, c)\}. \tag{4}$$

Then, we define a function $R(\cdot)$ that takes in a set of scores and outputs a vector with integer elements, where 0 indicates the lowest score within the set and $M-1$ indicates the highest score within the set. Finally, we define the outpout of the MMR as the following ranking vector:

$$\mathbf{r} = \frac{1}{|A_a^*|} \sum_{a_i \in A_a^*} \frac{R(S_F(a_i))}{M}. \tag{5}$$

For clarity, we have focused on a single hyperparameter, but Eq. (5) can easily be extended to several hyperparameters. For evaluation methods where a high value is desirable, a high ranking indicates good performance, and vice versa for evaluation methods where a low value is desirable.

7 Experimental Setup

We evaluate our manipulation strategy across numerous datasets, models, and XAI methods, which are described below. We also define the feasible sets used in our manipulation methods.

Models and Datasets: We examine several widely used computer vision datasets; MNIST [19], FashionMNIST [54], PneumoniaMNIST [29], and ImageNet [18], and two common deep learning architectures: LeNet [34] and ResNet18 [23]. The LeNet is used for classifying MNIST, FashionMNIST, and PneumoniaMNIST, while the Resnet18 is used for classifying ImageNet. For ImageNet, we randomly sample 100 samples to conduct the faithfulness evaluation, for PneumoniaMNIST we use 500 samples, and for the remaining datasets we use 1000 samples. We choose 100 samples for ImageNet because the larger size of these images increases the computational complexity. We choose 500 for PneumoniaMNIST as it does not have 1000 samples in its test set.

XAI Methods: We investigate the following XAI methods; Layer-wise relevance propagation (LRP) [7], Saliency [38], and KernelSHAP [35] using the `captum` library [32]. We have picked these three methods as they represent common choices in the XAI field, and we have focused on only three methods to provide a clear experimental analysis without overloading the reader.

7.1   Defining the Feasible Set of Hyperparameters for Faithfulness

A critical aspect of the manipulation methods outlined in Sec. 4 is to determine the feasible set of hyperparameters. This requires in-depth knowledge of the family of quantitative metrics that we aim to manipulate. In this work, we focus on the faithfulness family of evaluation metrics and the critical hyperparamters outlines in Sec. 5.1. We focus on a subset of hyperparameters to provide a clear and understandable evaluation of our manipulation strategies. The feasible set of hyperparameters considered in this work are shown in Tab. 2. This selection is based on common choices in the literature for partition size [7, 15, 24, 26, 53], perturbation function [3, 40, 46], and normalization function [10, 11, 24]. We consider the aggregation function fixed as AUC aggregation, which means that a lower faithfulness score is better. Specifically, we compute the AUC of the faithfulness curve from the set of perturbed model outputs $\{\hat{y}_{S_1}, \cdots, \hat{y}_{S_K}\}$.

	MNIST	FashionMNIST	PneumMNIST	ImageNet
Partition size	$\{14, 28, 56\}$	$\{14, 28, 56\}$	$\{14, 28, 56\}$	$\{112, 224, 448\}$
Perturbation:	$\{\mathcal{N}(0,1), \mathcal{U}(0,1), \mathcal{G}(\cdot)\}$	$\{\mathcal{N}(0,1), \mathcal{U}(0,1), \mathcal{G}(\cdot)\}$	$\{\mathcal{N}(0,1), \mathcal{U}(0,1), \mathcal{G}(\cdot)\}$	$\{\mathcal{N}(0,1), \mathcal{U}(0,1), \mathcal{G}(\cdot)\}$
Normalization	$\{$True, False$\}$	$\{$True, False$\}$	$\{$True, False$\}$	$\{$True, False$\}$

Table 2: The feasible set of hyperparameter considered in this work for different datasets. $\mathcal{G}(\cdot)$ denotes Gaussian blurring.

8   Results

Here we present the results of performing our proposed inter-manipulation and intra-manipulation. In both cases, we survey the literature and create what

XAI method	MNIST		FashionMNIST		PneumMNIST		ImageNet	
	base	manip.	base	manip.	base	manip.	base	manip.
LRP	25.20	7.86	21.46	5.37	21.31	6.06	129.61	41.48
Saliency	20.23	6.80	15.65	4.72	23.28	4.23	124.93	37.53
KernelSHAP	23.94	8.01	18.28	4.81	22.06	4.29	128.72	40.14

Table 3: Intra-results across several datasets and methods. Lower is better.

we call the *base* set of hyperparameters. The *base* set of hyperparameters for MNIST, FashionMNIST, and PneumoniaMNIST is a partition size of 28, uniform noise as perturbations, and no normalization. For ImageNet, the *base* set of hyperparameters is a partition size of 224, uniform noise as perturbations, and no normalization. After manipulation using Definition 1 and Definition 2, we will obtain a new set of hyperparameters that we refer to as the *manipulated* set of hyperparameters. Our results are centered around comparing the performance of the *base* set and the *manipulated* set.

8.1   Intra-Manipulation Results

Tab. 3 shows the results of performing the intra-manipulation proposed in Definition 1, where *base* is the score obtained with the selected set of hyperparameters described above and *manipulated* is the score obtained after manipulation. These results demonstrate that there is much room for changing the evaluation outcome for a single XAI method, in some cases as much as a 130 % improvement from the *base* to the *manipulated* evaluation outcome. Note that the *manipulated* scores are not directly comparable, since the manipulation is performed method-wise and the hyperparameters can be different. Therefore, the inter-manipulation shown in the next section must be used to alter the outcome of an evaluation across methods.

8.2   Inter-Manipulation Results

Tab. 4, Tab. 5, and Tab. 6 show the results of performing the inter-manipulation proposed in Definition 2, where the scores are manipulated towards LRP, Saliency, and KernelSHAP, respectively. For some tasks, the evaluation outcome can be manipulated such that most of the three methods achieves the best performance. This is particularly apparent for PneumoniaMNIST, where all XAI methods can achieve the best performance after manipulation. For some datasets there is less room for manipulation. This is most clear from the ImageNet results. That said, the evaluation difference between explanation methods can still be reduced and thus make the XAI evaluation findings less conclusive (see e.g. Imagenet results in Tab. 6). In Appendix A, we provide a summary of the amount of times each hyperparameter occurs in the manipulated set.

XAI method	MNIST base	MNIST manip.	FashionMNIST base	FashionMNIST manip.	PneumMNIST base	PneumMNIST manip.	ImageNet base	ImageNet manip.
LRP	25.19	**37.79**	21.46	35.42	**21.31**	**43.53**	129.61	128.02
Saliency	**20.23**	46.23	**15.65**	**34.75**	23.28	47.42	**124.93**	**123.93**
KernelSHAP	23.94	50.77	21.45	41.42	22.06	45.30	128.72	131.97

Table 4: Inter-results with manipulation towards *LRP*. Lower is better.

XAI method	MNIST base	MNIST manip.	FashionMNIST base	FashionMNIST manip.	PneumMNIST base	PneumMNIST manip.	ImageNet base	ImageNet manip.
LRP	25.19	51.41	21.46	43.80	**21.31**	25.86	129.61	167.14
Saliency	**20.23**	**41.57**	**15.65**	**31.83**	23.28	**19.61**	**124.93**	**147.56**
KernelSHAP	23.94	49.25	21.45	37.36	22.06	19.99	128.72	167.74

Table 5: Inter-results with manipulation towards *Saliency*. Lower is better.

8.3   Towards More Robust Faithfulness Evaluation

The results in Tab. 3, Tab. 4, Tab. 5, and Tab. 6, demonstrate that the evaluation outcome can be manipulated and can not be trusted, which reduces the trustworthiness of the quantitative evaluation. Here, we display the results of using MRR described in Sec. 6 towards mitigating the potential for manipulation.

Tab. 7 displays the results of this ranking procedure, which shows that the top-performing XAI methods change between datasets. However, if we average the ranking across all datasets, LRP comes out as the top-performing method closely followed by KernelSHAP, while Saliency seems to be consistently ranked lower. But note that there is notable variation in the scores, which we further illuminate in Fig. 2. The benefit of this ranking approach is that there is little room for manipulation since the top-performing methods will have to perform well across numerous hyperparameters and datasets. The downside of this ranking approach is that it requires a significant amount of computation to calculate the scores for all methods across all hyperparameters and datasets. Also, while averaging across datasets can provide robustness, it can also obfuscate important insights from a particular dataset. Therefore, it is important to include the dataset-wise ranking such that readers can get an overview of the evaluation.

XAI method	MNIST base	MNIST manip.	FashionMNIST base	FashionMNIST manip.	PneumMNIST base	PneumMNIST manip.	ImageNet base	ImageNet manip.
LRP	25.19	12.07	21.46	43.80	**21.31**	26.42	129.61	74.93
Saliency	**20.23**	**9.72**	**15.65**	**31.83**	23.28	19.95	**124.93**	74.21
KernelSHAP	23.94	11.53	21.45	37.36	22.06	**19.55**	128.72	74.66

Table 6: Inter-results with manipulation towards *KernelSHAP*. Lower is better.

XAI method	MNIST	FashionMNIST	PneumMNIST	ImageNet	All
LRP	$\mathbf{0.22 \pm 0.15}$	0.33 ± 0.00	0.21 ± 0.00	$\mathbf{0.26 \pm 0.00}$	$\mathbf{0.29 \pm 0.14}$
Saliency	0.41 ± 0.26	0.44 ± 0.31	0.37 ± 0.31	0.41 ± 0.33	0.41 ± 0.30
KernelSHAP	0.37 ± 0.33	$\mathbf{0.22 \pm 0.31}$	$\mathbf{0.33 \pm 0.27}$	0.33 ± 0.06	0.31 ± 0.31

Table 7: MRR across feasible set for each dataset and across datasets (last column). Lower is better, a rank of 0 is best and 1 is worst. Results show that the top performing method can change significantly between datasets, but when averaging across datasets LRP and KernelSHAP are highlighted as consistently higher ranked than Saliency.

Fig. 2 shows the faithfulness score for each configuration in the feasible set for each dataset. This plot illustrates that the average faithfulness score across the feasible set can often be quite close. However, there is large spread in the scores, which is present for all datasets. This spread demonstrates the lack of robustness in the faithfulness evaluation and is part of the reason why manipulation is possible in this case. But, that alone would not be enough to allow for manipulation, since the different methods could have the same change in scores for different set of hyperparameters. However, the large standard deviation in Tab. 7 shows that is not the case, since the ranking change between sets of hyperparameters. In other words, the XAI methods react differently to different sets of hyperparameters. This, in combination with the variation shown in Fig. 2, is what allows for manipulation in this study.
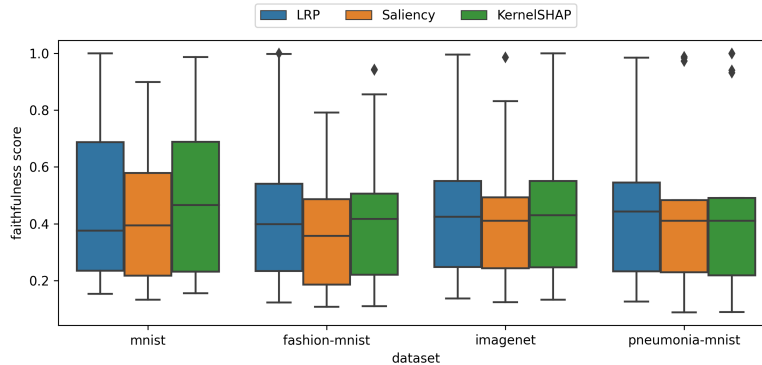

Fig. 2: Box plot showing faithfulness scores across all hyperparameter configurations in the feasible set for each dataset. The plot illustrates that the average faithfulness score is similar between different XAI methods across datasets. However the high variance enables a target manipulation. Note that the scores have been normalized dataset-wise by the highest score to allow for comparison across datasets.

9   Discussion and Limitations

The hyperparameters described in Sec. 7.1 could be extended to include other important choices such as the order of perturbation, i.e., descending or ascending [43] and the type of normalization function applied [11]. Also, in all our experiments we repeatedly perturb the input until the entire image is perturbed, which is the standard approach in faithfulness analysis. However, when the majority of pixels are removed there is danger of OOD effects (see e.g. Fig. 1), which can influence the evaluation outcome [22]. An alternative approach would be to only perturb parts of an image to avoid such OOD effects. One example is to perturb until the prediction changes and then stop [7]. But this introduces yet another hyperparamter, which further increases the scope for manipulation.

Our proposed MRR is a simple approach to combat the problem of manipulation, but it also has drawbacks. Most prominently, the computational cost rises quickly when more methods and hyperparameters are considered. Also, MMR requires domain expertise to determine the feasible set of hyperparameters. If the selection of the feasible set is done incorrectly, it might exacerbate the problem of manipulation since it can increase the amount of hyperparameters to choose from. MRR is also a ranking-based approach, where the scores depend on the set of explanation methods used in the analysis, including the cardinality of that set. Since the rankings are relative, they do not allow for meaningful comparisons across different tasks. To address this, we propose creating an open-source database, leveraging tools like Quantus [26] and OpenXAI [2], to efficiently store and standardise benchmarking results, thereby supporting researchers with the development and XAI evaluation. For future work, we further aim to expand the parameter sensitivity analysis to other families of quantitative measures such as randomisation [1, 25] and robustness [3, 17, 55] which rely on parameters such as segmentation masks and noise perturbation methods, respectively.

10   Conclusion

We have presented two general-purpose methods for manipulating the quantitative evaluation of explanation methods. Intra-manipulation which increases the performance of a single method and inter-manipulation which manipulates a comparative analysis of XAI methods. The motivation for these methods is based on the lack of ground truth explanations, which makes the selection of hyperparameters in quantitative evaluation for XAI challenging. We demonstrate the effectiveness of our manipulation strategies across numerous vision datasets and XAI methods for faithfulness metrics, with results indicating that there is significant room for manipulation of the evaluation outcome. This has potentially big implications for the XAI community, as it shows that evaluation outcomes cannot always be "taken at face value" and therefore, trusted. Lastly, we present a new ranking-based procedure that aims to improve the reliability of quantitative evaluation of XAI. We believe that this work highlights the difficulty of conducting reliable XAI evaluation and emphasizes the importance of a holistic and transparent approach to evaluation in XAI.

References

1. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. p. 9525–9536. NIPS'18, Curran Associates Inc., Red Hook, NY, USA (2018) 3, 14

2. Agarwal, C., Krishna, S., Saxena, E., Pawelczyk, M., Johnson, N., Puri, I., Zitnik, M., Lakkaraju, H.: OpenXAI: Towards a transparent evaluation of model explanations. In: Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2022), https://openreview.net/forum?id=MU2495w47rz 2, 3, 14

3. Alvarez Melis, D., Jaakkola, T.: Towards robust interpretability with self-explaining neural networks. In: Advances in Neural Information Processing Systems. pp. – (2018) 2, 3, 5, 6, 7, 8, 10, 14

4. Ancona, M., Ceolini, E., Öztireli, C., Gross, M.: Towards better understanding of gradient-based attribution methods for deep neural networks. In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net (2018) 2, 5, 6, 7, 8

5. Arras, L., Osman, A., Samek, W.: Clevr-xai: A benchmark dataset for the ground truth evaluation of neural network explanations. Information Fusion **81**, 14–40 (2022). https://doi.org/https://doi.org/10.1016/j.inffus.2021.11.008, https://www.sciencedirect.com/science/article/pii/S1566253521002335 6

6. Arya, V., Bellamy, R.K.E., Chen, P., Dhurandhar, A., Hind, M., Hoffman, S.C., Houde, S., Liao, Q.V., Luss, R., Mojsilovic, A., Mourad, S., Pedemonte, P., Raghavendra, R., Richards, J.T., Sattigeri, P., Shanmugam, K., Singh, M., Varshney, K.R., Wei, D., Zhang, Y.: One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques. CoRR **abs/1909.03012** (2019), http://arxiv.org/abs/1909.03012 7

7. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLOS ONE **10**(7), e0130140 (Jul 2015). https://doi.org/10.1371/journal.pone.0130140, https://doi.org/10.1371/journal.pone.0130140 1, 2, 4, 5, 6, 8, 10, 14

8. Bansal, N., Agarwal, C., Nguyen, A.: SAM: the sensitivity of attribution methods to hyperparameters. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2020, Seattle, WA, USA, June 14-19, 2020. pp. 11–21. Computer Vision Foundation / IEEE (2020) 3

9. Bhatt, U., Weller, A., Moura, J.M.F.: Evaluating and aggregating feature-based model explanations. In: Bessiere, C. (ed.) Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020. pp. 3016–3022. ijcai.org (2020) 2, 5, 6, 7, 8

10. Bhatt, U., Weller, A., Moura, J.M.F.: Evaluating and aggregating feature-based model explanations. In: International Joint Conference on Artificial Intelligence. pp. 3016–3022 (2020). https://doi.org/10.24963/ijcai.2020/417 3, 5, 6, 8, 10

11. Binder, A., Weber, L., Lapuschkin, S., Montavon, G., Müller, K.R., Samek, W.: Shortcomings of top-down randomization-based sanity checks for evaluations of deep neural network explanations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16143–16152 (2023). https://doi.org/10.1109/CVPR52729.2023.01549, https://doi.org/10.1109/CVPR52729.2023.01549 4, 9, 10, 14

12. Blücher, S., Vielhaben, J., Strodthoff, N.: Decoupling pixel flipping and occlusion strategy for consistent xai benchmarks (2024) 2, 4, 5
13. Brocki, L., Chung, N.C.: Evaluation of interpretability methods and perturbation artifacts in deep neural networks. CoRR **abs/2203.02928** (2022) 4
14. Brunke, L., Agrawal, P., George, N.: Evaluating input perturbation methods for interpreting CNNs and saliency map comparison. In: Computer Vision – ECCV 2020 Workshops, pp. 120–134. Springer International Publishing (2020) 4
15. Bykov, K., Hedström, A., Nakajima, S., Höhne, M.M.: Noisegrad - enhancing explanations by introducing stochasticity to model weights. In: Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022. pp. 6132–6140. AAAI Press (2022) 4, 10
16. Chalasani, P., Chen, J., Chowdhury, A.R., Wu, X., Jha, S.: Concise explanations of neural networks using adversarial training. In: III, H.D., Singh, A. (eds.) Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 1383–1391. PMLR (13–18 Jul 2020), https://proceedings.mlr.press/v119/chalasani20a.html 2, 3
17. Dasgupta, S., Frost, N., Moshkovitz, M.: Framework for evaluating faithfulness of local explanations. In: International Conference on Machine Learning. pp. 4794–4815. PMLR (2022) 2, 5, 6, 14
18. Deng, J., et al.: Imagenet: A large-scale hierarchical image database. In: Computer Vision and Pattern Recognition. pp. 248–255 (2009) 10
19. Deng, L.: The mnist database of handwritten digit images for machine learning research. IEEE Signal Processing Magazine **29**(6), 141–142 (2012) 10
20. Dolci, G., Cruciani, F., Galazzo, I.B., Calhoun, V.D., Menegaz, G.: Objective assessment of the bias introduced by baseline signals in XAI attribution methods. In: IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering, MetroXRAINE 2023, Milano, Italy, October 25-27, 2023. pp. 266–271. IEEE (2023) 4
21. Fong, R.C., Vedaldi, A.: Interpretable explanations of black boxes by meaningful perturbation. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 3449–3457 (2017). https://doi.org/10.1109/ICCV.2017.371 8
22. Hase, P., Xie, H., Bansal, M.: The out-of-distribution problem in explainability and search methods for feature importance explanations. In: Ranzato, M., Beygelzimer, A., Dauphin, Y.N., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual. pp. 3650–3666 (2021) 4, 14
23. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 CVPR. pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90 10
24. Hedström, A., Bommer, P.L., Wickstrøm, K.K., Samek, W., Lapuschkin, S., Höhne, M.M.: The meta-evaluation problem in explainable AI: Identifying reliable estimators with metaquantus. Transactions on Machine Learning Research (2023), https://openreview.net/forum?id=j3FKOOHyfU 2, 3, 9, 10
25. Hedström, A., Weber, L., Lapuschkin, S., Höhne, M.: A fresh look at sanity checks for saliency maps. In: Explainable Artificial Intelligence. pp. 403–420. Springer Nature Switzerland, Cham (2024) 2, 4, 14
26. Hedström, A., Weber, L., Krakowczyk, D., Bareeva, D., Motzkus, F., Samek, W., Lapuschkin, S., Hãhne, M.M.C.: Quantus: An explainable ai toolkit for re-

sponsible evaluation of neural network explanations and beyond. Journal of Machine Learning Research **24**(34), 1–11 (2023), http://jmlr.org/papers/v24/22-0142.html 2, 3, 8, 10, 14

27. Hooker, S., Erhan, D., Kindermans, P.J., Kim, B.: A benchmark for interpretability methods in deep neural networks. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019), https://proceedings.neurips.cc/paper_files/paper/2019/file/fe4b8556000d0f0cae99daa5c5c5a410-Paper.pdf 8

28. Karimi, A.H., Muandet, K., Kornblith, S., Schölkopf, B., Kim, B.: On the relationship between explanation and prediction: A causal view. In: XAI in Action: Past, Present, and Future Applications (2023), https://openreview.net/forum?id=ag1CpSUjPS 4

29. Kermany, D.S., Goldbaum, M., Cai, W., Valentim, C.C., Liang, H., Baxter, S.L., McKeown, A., Yang, G., Wu, X., Yan, F., Dong, J., Prasadha, M.K., Pei, J., Ting, M.Y., Zhu, J., Li, C., Hewett, S., Dong, J., Ziyar, I., Shi, A., Zhang, R., Zheng, L., Hou, R., Shi, W., Fu, X., Duan, Y., Huu, V.A., Wen, C., Zhang, E.D., Zhang, C.L., Li, O., Wang, X., Singer, M.A., Sun, X., Xu, J., Tafreshi, A., Lewis, M.A., Xia, H., Zhang, K.: Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell **172**(5), 1122–1131.e9 (2018). https://doi.org/https://doi.org/10.1016/j.cell.2018.02.010, https://www.sciencedirect.com/science/article/pii/S0092867418301545 10

30. Kindermans, P.J., Hooker, S., Adebayo, J., Alber, M., Schütt, K.T., Dähne, S., Erhan, D., Kim, B.: The (Un)reliability of Saliency Methods, p. 267–280. Springer International Publishing (2019). https://doi.org/10.1007/978-3-030-28954-6_14 3

31. Koenen, N., Wright, M.N.: Toward understanding the disagreement problem in neural network feature attribution. CoRR **abs/2404.11330** (2024) 4

32. Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., Yan, S., Reblitz-Richardson, O.: Captum: A unified and generic model interpretability library for pytorch (2020) 10

33. Krishna, S., Han, T., Gu, A., Pombra, J., Jabbari, S., Wu, S., Lakkaraju, H.: The disagreement problem in explainable machine learning: A practitioner's perspective. CoRR **abs/2202.01602** (2022), https://arxiv.org/abs/2202.01602 2, 4

34. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998). https://doi.org/10.1109/5.726791 10

35. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. p. 4768–4777. NIPS'17, Curran Associates Inc., Red Hook, NY, USA (2017) 10

36. Mamalakis, A., Barnes, E.A., Ebert-Uphoff, I.: Carefully choose the baseline: Lessons learned from applying XAI attribution methods for regression tasks in geoscience. CoRR **abs/2208.09473** (2022) 4

37. Montavon, G., Samek, W., Müller, K.R.: Methods for interpreting and understanding deep neural networks. Digital Signal Processing **73**, 1–15 (2018). https://doi.org/https://doi.org/10.1016/j.dsp.2017.10.011, https://www.sciencedirect.com/science/article/pii/S1051200417302385 2, 5, 8

38. Morch, N., et al.: Visualization of neural networks using saliency maps. In: International Conference on Neural Networks. pp. 2085–2090 (1995) 1, 10

39. Nguyen, A., Martínez, M.R.: On quantitative aspects of model interpretability. CoRR **abs/2007.07584** (2020), https://arxiv.org/abs/2007.07584 2, 6, 7

40. Rieger, L., Hansen, L.K.: IROF: a low resource evaluation metric for explanation methods. CoRR **abs/2003.08747** (2020), https://arxiv.org/abs/2003.08747 2, 5, 6, 7, 8, 10

41. Rong, Y., Leemann, T., Borisov, V., Kasneci, G., Kasneci, E.: A consistent and efficient evaluation strategy for attribution methods. In: Proceedings of the 39th International Conference on Machine Learning. pp. 18770–18795. PMLR (2022) 2, 8

42. Rong, Y., Leemann, T., Borisov, V., Kasneci, G., Kasneci, E.: A consistent and efficient evaluation strategy for attribution methods. In: International Conference on Machine Learning. pp. 18770–18795 (2022) 4, 5, 7, 8

43. Samek, W., Binder, A., Montavon, G., Lapuschkin, S., Müller, K.: Evaluating the visualization of what a deep neural network has learned. IEEE Trans. Neural Networks Learn. Syst. **28**(11), 2660–2673 (2017) 2, 4, 5, 6, 7, 8, 14

44. Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., Müller, K.R. (eds.): Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Springer International Publishing (2019). https://doi.org/10.1007/978-3-030-28954-6, http://dx.doi.org/10.1007/978-3-030-28954-6 1

45. Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for simplicity: The all convolutional net. In: ICLR Workshop (2015) 4

46. Sturmfels, P., Lundberg, S., Lee, S.I.: Visualizing the impact of feature attribution baselines. Distill (2020). https://doi.org/10.23915/distill.00022, https://distill.pub/2020/attribution-baselines 2, 3, 5, 10

47. Su, J., Vargas, D.V., Sakurai, K.: One pixel attack for fooling deep neural networks. IEEE Transactions on Evolutionary Computation **23**(5), 828–841 (2019). https://doi.org/10.1109/TEVC.2019.2890858 8

48. Sundararajan, M., Taly, A.: A note about: Local explanation methods for deep neural networks lack sensitivity to parameter values. CoRR **abs/1806.04205** (2018) 4

49. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017. Proceedings of Machine Learning Research, vol. 70, pp. 3319–3328. PMLR (2017), http://proceedings.mlr.press/v70/sundararajan17a.html 3

50. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017. Proceedings of Machine Learning Research, vol. 70, pp. 3319–3328. PMLR (2017) 4

51. Theiner, J., Müller-Budack, E., Ewerth, R.: Interpretable semantic photo geolocalization. CoRR **abs/2104.14995** (2021), https://arxiv.org/abs/2104.14995 3, 6

52. Tomsett, R., Harborne, D., Chakraborty, S., Gurram, P., Preece, A.D.: Sanity checks for saliency metrics. In: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020. pp. 6021–6029. AAAI Press (2020) 4

53. Wickstrøm, K.K., Trosten, D.J., Løkse, S., Boubekki, A., Mikalsen, K.Ø., Kampffmeyer, M.C., Jenssen, R.: RELAX: representation learning explainability. Int. J. Comput. Vis. pp. 1584–1610 (2023) 1, 8, 10
54. Xiao, H., Rasul, K., Vollgraf, R.: Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms (2017) 10
55. Yeh, C.K., Hsieh, C.Y., Suggala, A.S., Inouye, D.I., Ravikumar, P.: On the (in)fidelity and sensitivity of explanations. In: Neural Information Processing Systems (2019) 2, 14