An Experimental Investigation into the Evaluation of Explainability Methods for Computer Vision

Stassin, Sédrick; Englebert, Alexandre; Albert, Julien; Nanfack, Geraldin; Versbraegen, Nassim; Frénay, Benoît; Peiffer, Gilles; Doh, Miriam; Riche, Nicolas; De Vleeschouwer, Christophe

Published in:
Communications in Computer and Information Science

Publication date:
2023

Document Version
Peer reviewed version

An Experimental Investigation into the Evaluation of Explainability Methods for Computer Vision

Sédrick Stassin[1], Alexandre Englebert[2,3], Géraldin Nanfack[4], Julien Albert[4], Nassim Versbraegen[5,6], Gilles Peiffer[2], Miriam Doh[8,9], Nicolas Riche[7], Benoît Frenay[4], and Christophe De Vleeschouwer[2]

[1] ILIA unit, Faculty of Engineering, University of Mons, Mons, Belgium
[2] Institute for Information and Communication Technologies, Electronics and Applied Mathematics, Louvain-la-Neuve, UCLouvain, Belgium
[3] Service de chirurgie orthopédique et traumatologie, Cliniques Universitaires Saint-Luc UCL & Neuro Musculo Skeletal Lab (NMSK), IREC, Woluwe-Saint-Lambert, UCLouvain, Belgium
[4] PReCISE Research Center, Namur Digital Institute (NADI), University of Namur, Belgium
[5] Interuniversity Institute for Bioinformatics in Brussels, ULB-VUB, Brussels, Belgium
[6] Machine Learning Group, Université Libre de Bruxelles, Brussels, Belgium
[7] Multitel, Mons, Belgium
[8] IRIDIA, Université Libre de Bruxelles, Brussels, Belgium
[9] ISIA unit, University of Mons, Mons, Belgium

Abstract. E**X**plainable **A**rtificial **I**ntelligence (XAI) aims to help users to grasp the reasoning behind the predictions of an Artificial Intelligence (AI) system. Many XAI approaches have emerged in recent years. Consequently, the subfield related to the evaluation of XAI methods has gained considerable attention, with the aim of determining which methods provide the best explanation using various approaches and criteria. However, the literature lacks a comparison of the evaluation metrics themselves that could be used to evaluate XAI methods. This work aims to partially fill this gap by comparing 14 different metrics when applied to nine state-of-the-art XAI methods and three dummy methods (e.g., random saliency maps) used as references. Experimental results on image data show which of these metrics produce highly correlated results, indicating potential redundancy. We also demonstrate the significant impact of varying the baseline hyperparameter on the evaluation metric values. Finally, we use dummy methods to assess the reliability of metrics in terms of ranking, pointing out their limitations.

1 Introduction

E**X**plainable **A**rtificial **I**ntelligence (XAI) aims to provide accurate predictions and explanations of the predictions of an AI system in humanly understandable terms [9]. One of the most ubiquitous types of XAI methods in computer vision is that of attribution-based methods, which provide a weighted set of relevant features concerning the prediction of the model. These methods allow for relevant visualizations through saliency maps and are sometimes referred to as "saliency methods." While a broad class of such methods exists, evaluating them remains a significant challenge because of the lack of ground truth explanations [27]. Several properties and requirements (e.g., faithfulness), leading to the formulation of quantitative metrics, have been proposed to assess the

quality of saliency methods [3, 6]. While the existence of multiple metrics could allow one to obtain a better overview of the reliability of a given method, newly proposed saliency methods are often only compared to existing work using a limited subset of the available evaluation metrics. Furthermore, there is little to no quantitative analysis or detailed investigation into the potential redundancy of XAI evaluation metrics. This paper proposes to fill this gap by studying the concordance and redundancy of XAI evaluation metrics. To do so, we perform statistical analysis on various metrics after evaluating nine state-of-the-art XAI methods for convolutional neural networks (CNN) applied to computer vision and three dummy methods which are not real saliency methods and are used as sanity checks. In addition, we evaluate the influence of the baseline hyperparameter of the metrics (e.g. black or white pixels to simulate the deletion/insertion of pixels). Previous work has shown that the choice of the baseline has a big influence on the results of XAI methods [34]. However, to the best of our knowledge, no work has explored the impact of the baseline for the metrics, or its impact on the produced scores. Finally, thanks to the dummy methods, we discuss the reliability of metrics. The remainder of the paper is organized as follows. Section 2 presents an overview of related work. Section 3 describes the experimental setup that led to the results discussed in Section 4. Section 5 concludes the paper.

2 State of the Art

2.1 Saliency Methods

There is a large variety of saliency methods. This work does not have the ambition of being exhaustive, therefore we focus on three families only: gradient-based methods, perturbation-based methods, and Class Activation Map (CAM) methods. The models considered are CNNs, for which the XAI literature is extensive. Saliency maps for new architectures, such as ViT [10] are not entirely interchangeable with respect to CNNs, apart from gradient-based methods and some perturbation methods. This section will evoke the principal characteristics of the methods used in our experiments. For a detailed description of the functioning of the different methods, we refer the reader to the indicated papers specific to each method.

Gradient-based methods Gradient-based methods use the gradient with respect to the inputs (or a modified, backpropagated version thereof), i.e., the derivative of the model output with respect to its input[1]. This derivative produces a saliency map. These methods were introduced in the pioneering work of Simonyan et al. [29] to highlight the most significant pixels in the decision-making of an image classification neural network. Here, we shall present three representative variants. *SmoothGrad* [32] was proposed to explicitly reduce noise in attributions. It generates multiple samples by adding Gaussian noise to the original image and averages the calculated gradients to produce a saliency map. *Guided backpropagation* [33] is similar to [29], with a specific handling of ReLu during the backpropagation. It keeps the non-negative values found in

[1] More precisely, the derivative of the preactivations of the softmax layer.

the forward and backward passes when back-propagating through the ReLU functions. The *Integrated Gradients* method [35] linearly interpolates the input image $\boldsymbol{x} \in \mathbb{R}^D$ with a baseline image $\boldsymbol{x}'$ over an $\alpha \in [0, 1]$ parameter, and averages the gradient for all of these interpolations. The baseline input represents the absence of a feature in the image (e.g., black or uniform, Gaussian noise, blurred image, etc.).

Perturbation-based Methods These methods analyze the evolution of the model's output as a function of variations in the input. They aim to measure whether information lost due to the perturbation is negatively or positively correlated to the prediction score. Among the most recent techniques, *RISE* [24] uses random occlusion patterns produced by bilinear upsampling of lower resolution binary masks to perturb the input. The model produces a probability score for each masked image for each class, the saliency map for the original image being produced for each class of interest by a linear combination of the masks, weighted by the corresponding class scores for each mask.

CAM Methods CAM methods use the activation from a convolutional layer to produce the saliency maps. The original method [39] uses the activations of the last convolutional layer and linearly balances them by the weights of a one-layer linear classifier. Since CAM has limited applicability across architectures, *Grad-CAM* [28] was introduced as a generalization, defining the weighting factors for the linear combination of the activations as the average of the gradient flowing through each channel of the last convolutional layer. *Score-CAM* [37] does not depend on gradients, as it generates the weighting factors to combine the activation map using its forward passing score on the target class. The saliency map is then computed by a linear combination of the obtained scores and activation maps. *Poly-CAM* [11] recursively multiplexes the high-resolution activation maps in the early layers with upsampled versions of the class-specific activation maps in the last layers. The weighting factors for the multiple CAMs are derived from scores computed similarly to perturbation methods and Score-CAM. *Layer-CAM* [17] gathers CAM from all layers of the CNN by directly multiplying the gradient element-wise with the activation and averaging over the channels instead of performing a linear combination. This allows object localization collection from rough spatial localization to fine-grained details. *CAMERAS* [16] fuses the activation maps and backpropagated gradients of a layer for different scaled versions of the same input image before performing an element-wise product of these fused activation maps and gradients, similarly to Layer-CAM on the last convolutional layer.

2.2 Evaluation Metrics

To evaluate and compare the different XAI methods presented in Section 2.1, there are several types of metrics. We consider the four distinct families of such metrics that are applicable to our dataset, based on the property of explainability they aim to characterize, following the formalism proposed in Quantus [14]. *Faithfulness metrics* measure how explanations follow the predictive behavior of the model. Seven of them are explored in this paper. Faithfulness Correlation [6] partitions the input image into subsets of features whose values are iteratively replaced by a baseline value (e.g., a black pixel),

and computes the Pearson correlation between the drop in classification probability and the sum of the relevance attribution of a subset. Similarly, Faithfulness Estimation [3] calculates the Pearson correlation between the drop in classification probability and feature relevance. Monotonicity Arya [4] measures the extent to which model performance increases (as measured by classification probability) when features of increasing significance are added, whereas Monotonicity Nguyen [23] follows the same idea, but measures the model performance increase through probability estimation uncertainty. Pixel Flipping [5], also called the Deletion metric by some authors [24, 37, 11] flips pixels with high relevance scores from the relevance heatmap and looks at the evolution of the probability score. Region Perturbation [27]) and Selectivity [22] both extend this methodology to areas of an image. *Robustness metrics* measure the stability of explanations to small input perturbations. Three of them are used in the following work. The Local Lipschitz Estimate [2] measures the consistency in explanations for adjacent samples. Max-Sensitivity and Avg-Sensitivity [38] quantify how explanations change when inputs are infinitesimally perturbed, through a Monte Carlo sampling-based approximation. *Complexity metrics* measure explanation conciseness. Three of them are compared in the experiments. Sparseness [8] uses the Gini index to measure whether only highly-attributed features are predictive of the model output. Complexity [6] measures the entropy of the fractional contributions to the total attribution, whereas Effective Complexity [23] looks at how many attributions exceed a certain threshold. *Randomization metrics* measure model deterioration as a function of parameter randomization. Two methods are explored. Model Parameter Randomization [1] quantifies the similarity between original explanations and explanations from sequential randomization of successive model layers. The Random Logit Test [31] computes the distance between the original explanation and the explanation obtained for another random class.

2.3 Benchmarking Saliency Evaluation Metrics

Previous works that quantitatively evaluate metrics of saliency methods are mainly found in XAI literature surveys. For example, general overviews of XAI method evaluations are available [7]. Beyond descriptions, Zhou et al. [40] couple XAI methods with corresponding evaluation metrics. In this paper, we assume families of metrics as described in Section 2.2. While several works in the literature discuss XAI evaluation metrics, very little is known about potential redundancies between metrics in the same family or not. Noteworthy is the work from Li et al. [21], which presents a review that encompasses a quantitative evaluation of seven saliency methods. These methods are compared using three types of metrics: faithfulness, localization, and robustness. Li et al. [21] point out that RISE and Grad-CAM perform well for most metrics. However, they also indicate that no particular method is consistently ranked as best with all metrics. Our work goes beyond the aforementioned approaches by offering an inquiry focused on metrics rather than methods. We evaluate and identify representative and potentially redundant metrics to obtain a comprehensive view of a method in terms of XAI evaluation.

3 Experimental Setup to Compare Evaluation Metrics

This section describes the experimental setup to obtain the results presented in Section 4. Section 3.1 presents the image data set and the deep learning models used to obtain the image classifications to be explained. Sections 3.2 and 3.3 introduce the selected saliency methods and evaluation metrics, respectively. Finally, Sections 3.4 and 3.5 describe how we process and analyze the results. The code is made available for reproducibility[2].

3.1 Data Set and Model for Image Classification

When evaluating saliency methods, it is common to use the ImageNet (2012 ILSVRC) validation set [26] (as in [37, 24]) or the CIFAR10 dataset [20] (as in [8, 38, 3]). In accordance with [37], we use a subset of 2000 randomly selected images from ImageNet, and we employed two widely known pretrained models (ResNet-50 [13] and VGG-16 [30]) from the PyTorch model zoo to obtain predictions, which are explained using the methods in Section 3.2. The preprocessing is done by scaling each image to a 224×224 resolution and normalizing RGB channels with a mean of $[0.485, 0.456, 0.406]$ and a standard deviation of $[0.229, 0.224, 0.225]$, as for the training set of ImageNet ([26]). For CIFAR10, the first 2000 images of the test dataset were utilized with a ResNet50 model pretrained on CIFAR10 available at [25]. Each image has an original size of 32×32 pixels and is normalized using a mean of $[0.4914, 0.4822, 0.4465]$ and a standard deviation of $[0.2023, 0.1994, 0.2010]$, following the preprocessed values utilized by the ResNet-50 pretrained model (and the other models) from [25].

3.2 Selected XAI Methods for Image Classification Explanation

For image classification explanation, we use state-of-the-art XAI methods that include Integrated Gradient [35], SmoothGrad [32], Guided Backpropagation [33], Grad-CAM [28], Score-CAM [37], Layer-CAM [17], Poly-CAM [11], RISE [24] and CAMERAS [16]. Three dummy saliency maps were added to the methods used to compare metrics: a randomly generated map (sampling a standard uniform distribution $U(0, 1)$ for each pixel), a Sobel filter and a map produced from a two-dimensional centered Gaussian ($\boldsymbol{\mu} = 0$, $\Sigma = I$). These three methods can be seen as an "ablation study" on metrics, which is a new contribution of our study since it has not been investigated before to the best of our knowledge. Indeed, they allow the comparison of rankings produced for a metric according to the expected ranking of dummy methods (w.r.t. the evaluated property such as faithfulness). This is shown in Section 4.3, on the reliability of the metrics. Gradient-based implementations stem from Captum ([19]), whereas CAM-based ones come from TorchCAM ([12]). The implementations of Poly-CAM, RISE as well as CAMERAS come from their authors[3]. Following common practices, the method

[2] `https://github.com/multitel-ai/Evaluation-XAI-Metrics-GD6-TRAIL`.

[3] `https://github.com/andralex8/polycam`, `https://github.com/eclique/RISE`, and `https://github.com/VisMIL/CAMERAS` resp.

parameters are fifty input perturbations (with unit variance Gaussian noise) for SmoothGrad, fifty interpolation steps for Integrated Gradient, four thousand perturbations for RISE and all the ResNet blocks for Layer-CAM and Poly-CAM.

3.3 Selected Evaluation Metrics to Compare XAI Methods

In order to evaluate the above XAI methods, evaluation metrics were selected from those introduced in Section 2.2. For faithfulness metrics, these include Faithfulness Correlation, Faithfulness Estimate, Pixel Flipping, Selectivity, Monotonicity Nguyen and Monotonicity Arya. We did not consider the Region Perturbation metric [27], as it behaves in the same way as the Selectivity metric [22]. For robustness metrics, the selected metrics include the Local Lipschitz Estimate, Max-Sensitivity and Avg-Sensitivity, which were applied with ten Monte Carlo samples (to have a reduced computational burden[4]). From the group of complexity metrics, those selected were Sparseness, Complexity and Effective Complexity. Random Logit and Model Parameter Randomization were included as randomization metrics. The latter was used with a top-to-down randomization scheme [1]. We used the metric implementations from Quantus [14].

For faithfulness metrics on ImageNet, the number of steps for each is 224 (in accordance with [11]), except for Selectivity employing a patch-based approach. In order to harmonize the number of patches and the number of steps, we set the size of these patches to 14×14. A parallel strategy is applied for the CIFAR10 dataset, where the patch size is updated to two pixels to maintain the same proportion (one-fourteenth of the image width). As for the Monotonicity Arya metric, we revisited its output to obtain a monotonicity ratio, replacing the binary outcome of true or false. The initial binary nature of the metric's output hindered meaningful comparison with other metric results. In addition, the original design led to consistently false outcomes since the majority, if not all, methods did not exhibit monotonic increments throughout the metric's duration. By presenting the output as a percentage, we achieve two objectives: firstly, enabling method ranking based on the percentage values, and secondly, facilitating an analysis of the degrees of monotonicity across methods. It is important to note that this enhancement does not alter significantly the core functionality of the metric.

Finally, one important consideration when evaluating faithfulness metrics is the simulation of *feature removal*, which is technically done by replacing a pixel value to a *baseline* value (e.g., black, white, or random). We studied the influence of this baseline hyperparameter on the scores generated by faithfulness metrics. As explained in Section 2.2, faithfulness metrics measure how explanations relate to the predictive behavior of the model by perturbing parts of each input with a chosen value. In our experiments, we used four baselines: black, white, random, and uniform values as implemented in Quantus. Specifically, the baseline replacement with random and uniform comes from the implementation of the random package in the Python Library ([36]). The values for each pixel with random are between 0 (included) and 1 (excluded), while the value chosen for uniform is between the minimum and the maximum in an image or patch.

[4] For example, evaluating Poly-CAM on Avg-Sensitivity with 10 MC samples took approximately 22 hours with five parallel jobs on an NVIDIA Geforce GTX 1080 Ti.

3.4 From Method Evaluation Scores to a Matrix of Scores

As a result of running all the aforementioned methods and evaluating them with the 14 listed metrics, we obtained, for each method either a score or a list of scores for each image, depending on the metric. From the listed metrics, only Pixel Flipping, Selectivity and Model Parameter Randomization returned a list of scores for each image. In these cases, it is common to aggregate the result by computing the AUC of the list of scores [37]. As our goal is to understand how metrics behave in relation to each other, we want to obtain a matrix $\boldsymbol{X}$ whose rows and columns correspond to methods and metrics, respectively. Each element of this matrix X_{ij} is obtained by averaging the scores of the j-th metric applied to the outputs (i.e., 2000 images) of the i-th method, similar to what Li et al. [21] did.

3.5 Comparison of Evaluation Metrics through Method Scores

In order to investigate pairwise comparisons of metrics, we compute the correlation coefficients on the scores produced by these metrics (i.e., the columns $X_{\cdot j}$ of $\boldsymbol{X}$ described above). The objective is to identify potential pairs of redundant metrics. Since we possibly have nonlinear relationships between scores, an adequate solution to assess correlations is to use Kendall's τ_b rank coefficients [18]. The use of p-values with Kendall's τ_b can reveal significant correlations. In the rest of this paper, Kendall's τ_b will simply be referred to as Kendall's τ. By multiplying pairwise comparisons, the risk of wrongly rejecting a null hypothesis (i.e., the absence of correlation) increases. To prevent this, we used Holm–Bonferroni-corrected p-values [15] with a family-wise error rate of 0.05. Beyond significance, it is important to remain cautious about the interpretation of correlations in the context of XAI methods. Correlated scores can still contain some major differences. Only (near-) perfectly correlated metrics, i.e., with a Kendall τ superior to 0.9, will be considered potentially redundant. Based on the definition of Kendall's τ

$$\tau = \frac{(p-q)}{\sqrt{(p+q+t)(p+q+u)}}, \tag{1}$$

with p the number of concordant pairs, q the number of discordant pairs, t the number of ties in the first ranking, and u the number of ties in the second ranking, and on the hypothesis that there are no ties in any of the two compared rankings ($t = u = 0$), the percentage of concordant pairs can be derived from Kendall's τ. The threshold of 0.9 that we use corresponds to $95\,\%$ matching pairs (i.e., fewer than 3 discordant pairs).

4 Experimental Results for Evaluation Metrics

This section analyzes the correlations between metrics, the influence of hyperparameters of metrics, and the reliability of evaluation metrics.

4.1 On the Correlation of Evaluation Metrics

Given pairs of metrics of the same type, we investigate which of these pairs achieve possible agreement. First, we find two couples of metrics that can be considered potentially redundant (based on the threshold of 0.9 defined in Section 3.5). For *complexity*

metrics (bottom right red square in Fig. 1a, 1b and 2), the Sparseness and Complexity metrics are significantly correlated with a Kendall's τ correlation coefficient of 0.97 (ResNet-50) and 1.0 (VGG-16) on ImageNet, and 0.94 (ResNet-50) on CIFAR10. For *robustness metrics* (second red square in Fig. 1a, 1b and 2), the results show that Max-Sensitivity and Avg-Sensitivity are correlated for CIFAR10 on ResNet-50 (($\tau = 0.88$) and largely correlated for both models on ImageNet ($\tau = 0.91$ for ResNet-50 and $\tau = 1.0$ for VGG-16) where they can be considered potentially redundant. Indeed, it is intuitive that Avg-Sensitivity and Max-Sensitivity are potentially redundant because the more similar the perturbed values in the neighborhood of the input are, the more the average values will be correlated to the maximum value. The choice is left to users as to whether they want to give importance to some maximum perturbation outliers and prefer the Avg-Sensitivity rather than the Max-sensitivity.

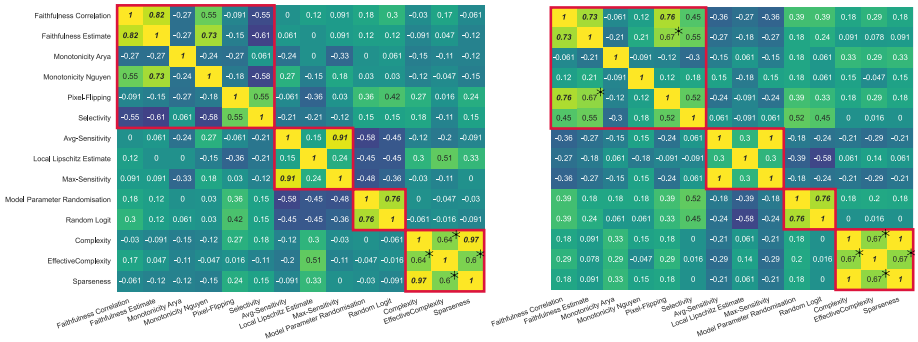
After analyzing the correlations above 0.9, we have the metrics that are significantly correlated below the 0.9 threshold defined in Section 3.5. For *randomization metrics* (third red square in Fig. 1a, 1b and 2), the results show that there is a moderately significant intra-group correlation between Model Parameter Randomization and Random Logit, i.e., 0.76 (ResNet-50) and 0.76 (VGG-16) on ImageNet, and 0.53 (ResNet-50) on CIFAR10. This suggests that even though the functioning of the randomization metric is significantly different (layer randomization versus explanation class randomization), explainability methods being sensitive to both types of randomization may be close.

Finally, for *faithfulness metrics* with black as a default baseline, there are no pairs of metrics that are significantly correlated on all models and datasets. We can only observe that in the case of ImageNet on both the Resnet-50 and VGG-16 models (Fig. 1) as well as on several baselines (Fig. 3), Faithfulness Correlation and Faithfulness Estimate are significantly correlated ($\tau = 0.82$ for Resnet-50, $\tau = 0.73$ for VGG-16). Furthermore, Pixel Flipping and Selectivity are significantly correlated on CIFAR10 ($\tau = 0.79$ on Resnet-50). We suggest that users use these metrics to analyze methods having similar maps (e.g., gradient-based methods only or CAM methods only), because methods such as the gradient-based methods produce sparser saliency maps (with high-frequency pixels) that will, most of the time, outperform any other methods. Beyond an intra-group analysis, it can be seen that there is no significant inter-group correlation between metrics for both models. This result is consistent with the fact that different types of metrics evaluate different properties of saliency methods.

4.2 On the Hyperparameters of Evaluation Metrics

In this section, we evaluate and discuss correlations for faithfulness metrics (with the ResNet-50 model) according to baseline values. All previous results were reported using black as the default baseline (value given to a pixel to simulate the *feature removal*) for all the faithfulness metrics that relied on this hyperparameter. We explore the impact of the baseline for the metrics and its impact on the produced scores.

Fig. 3 shows Kendall's τ correlation coefficients for the faithfulness metrics on different baselines. According to this figure, there are significant variations in the correlations calculated according to the baselines, such as the correlation between Monotonicity Arya and Pixel Flipping for a white and uniform baseline (-0.67 and -0.01 respectively). This is further emphasized in the results of Fig. 4, illustrating Kendall's



(a) ResNet-50 on ImageNet (b) VGG-16 on ImageNet

Fig. 1: Kendall's τ correlation coefficients for all metrics, with black as default baseline for faithfulness metrics. Groups of metrics are highlighted in red with (from left to right) Faithfulness, Randomisation, Robustness and Complexity [4]. Intra-group significant values are starred ("*"), based on Holm–Bonferroni-corrected p-values by taking into account the corresponding group sub-matrix.

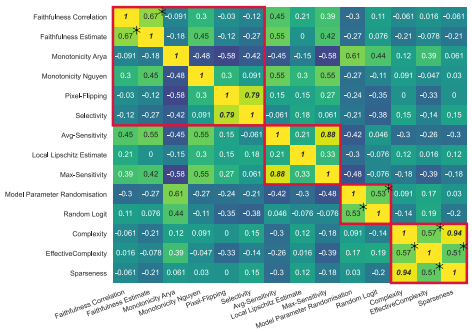
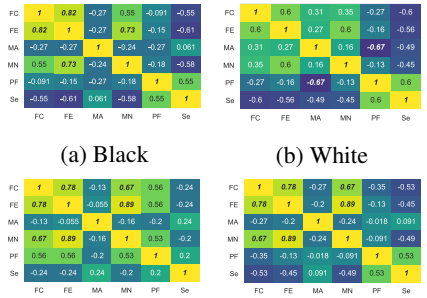


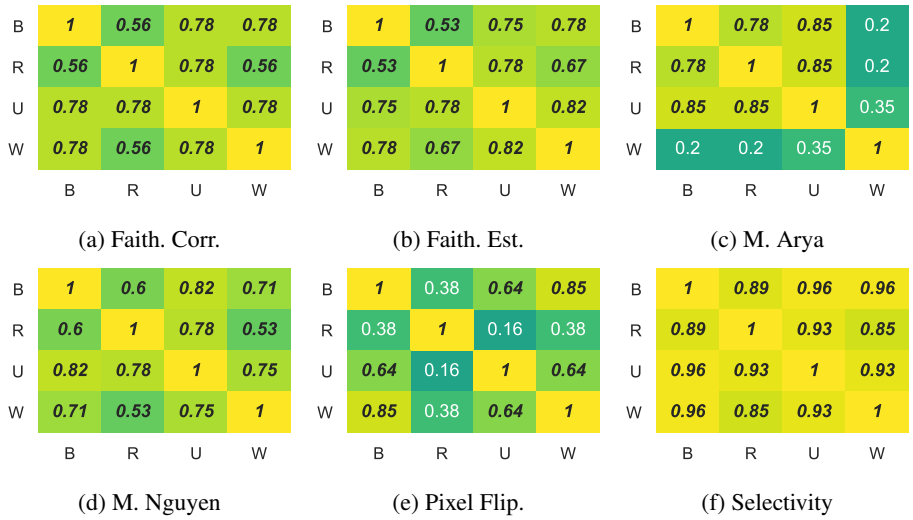
Fig. 2: Kendall's τ correlation coefficients for all metrics using ResNet-50 on CIFAR10, with black as default baseline for faithfulness metrics[4]. Intra-group significant values are starred ("*"), based on Holm–Bonferroni-corrected p-values by taking into account the corresponding group sub-matrix.



(a) Black (b) White

(c) Random (d) Uniform

Fig. 3: Kendall's τ correlations for faithfulness metrics using different baselines on ImageNet[4] using Resnet-50: Faithfulness Correlation (FC), Faithfulness Estimate (FE), Monotonicity Arya (MA), Monotonicity Nguyen (MN), Pixel Flipping (PF) and Selectivity (Se).



(a) Faith. Corr. (b) Faith. Est. (c) M. Arya

(d) M. Nguyen (e) Pixel Flip. (f) Selectivity

Fig. 4: Kendall τ correlation coefficients for the faithfulness metrics and different baselines on ImageNet [4]. Black (B), Random (R), Uniform (U), White (W) on ResNet-50.

τ correlation coefficients for each baseline and each faithfulness metric. Apart from Selectivity which shows resilience with correlation coefficients consistently over 0.85, the faithfulness metrics demonstrate high variability in the concordance of the scores depending on the baseline: all correlations are below 0.9, implying at least 3 discordant pairs in the scores according to the baseline (sometimes even below 0.2 such as for Pixel Flipping or Monotonicity Arya). Therefore, we recommend that new authors use multiple baselines in experiments using faithfulness metrics for a fair comparison. This becomes especially important considering faithfulness metrics are, to our knowledge, the most widely used in the state-of-the-art to evaluate XAI methods. Otherwise, it would mean drawing conclusions about potentially better XAI methods for a problem or dataset that may prove unreliable, as shown by the variability of results for different baselines.

4.3 On the Reliability of Metrics in Terms of Rankings

This section explores how the ranking attributed by evaluation metrics to the three dummy explanation methods defined in Section 3.2 (Sobel, Gaussian, and Random) compares to the ranking of state-of-the-art saliency methods. As the former are not valid explainability methods, their ranking in comparison to the ranking of the latter can unveil the strengths and weaknesses of the evaluation metrics. As each family of evaluation metrics is different, some may give dummy methods a low ranking, while others may provide a high ranking. This is discussed in detail below based on Fig. 5,

[4] Significant values are in italic bold, based on Holm–Bonferroni-corrected p-values by taking into account the complete matrix.

Fig. 6, and Fig. 7, which show the average ranking (rank 1 being the best) of the methods for each metric (based on the ranking for each image given by the metric scores) according to the model and dataset. We start the analysis by families of metrics from right to left in the figures.

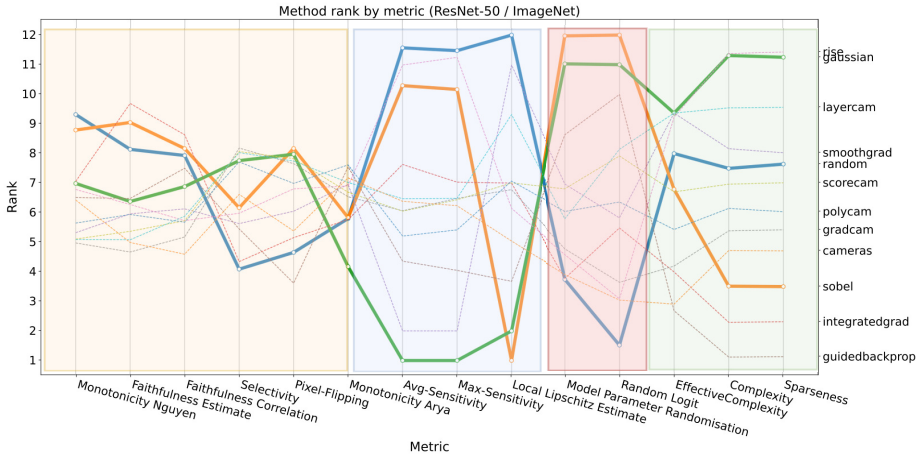


Fig. 5: Ranking of XAI methods with respect to metrics for ResNet-50 model on ImageNet (lower rank is better).

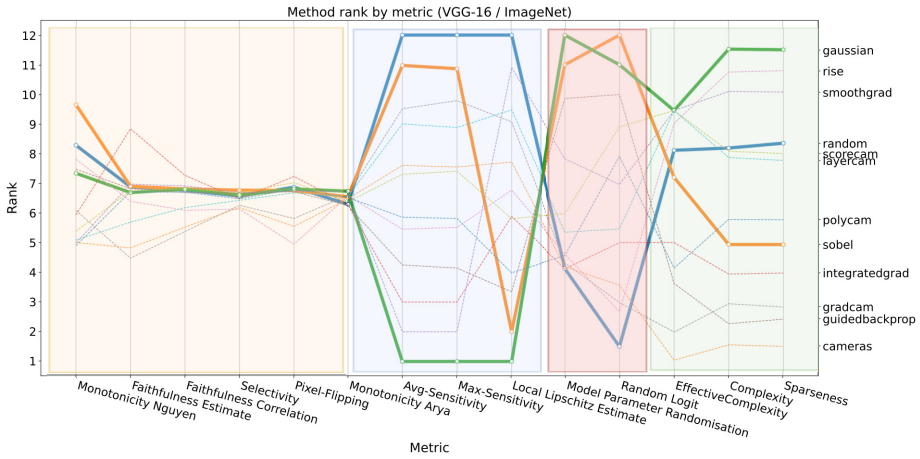


Fig. 6: Ranking of XAI methods with respect to metrics for VGG16 model on ImageNet (lower rank is better).

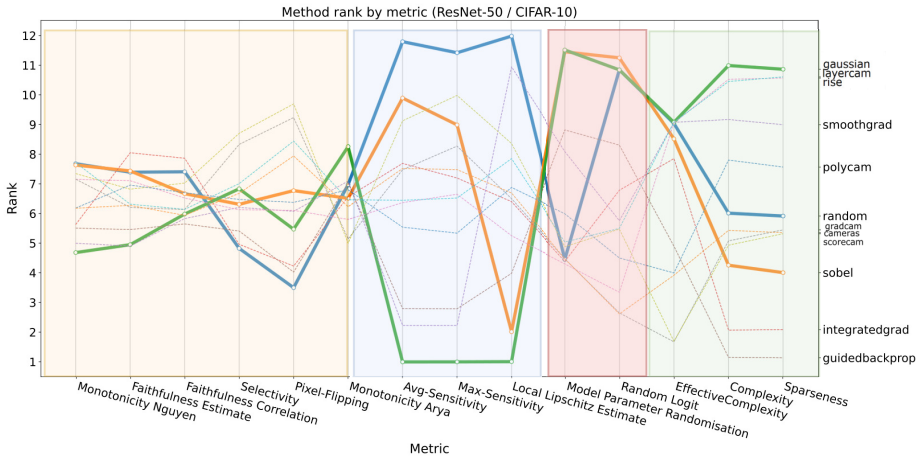


Fig. 7: Ranking of XAI methods with respect to metrics for Resnet-50 model on CIFAR10 (lower rank is better).

For *complexity* metrics, intuitively the complexity of the Sobel method depends on the input image, while that of Random is variable by definition. Thus, the ranking of these two dummy methods is not defined. However, the Gaussian method, with its constant and large map, should (and does) achieve a consistently poor ranking. The results obtained are consistent for all models and datasets.

For *randomization* metrics, Model Parameter Randomization (MPR) and Random Logit (RL) give a bad score to Sobel and Gaussian methods. This is the expected result since they do not depend on the model weights (MPR metric) or on the explanation (RL metric), and thus do not change with the randomization of the model or of the explanation class. On the other hand, as the Random dummy method produces a different result according to the model randomization (MPR metric) or the explanation class randomization (RL metric), it achieves an artificially good ranking by mimicking the big influence of the randomization. Note that this also shows that these metrics can be fooled by a nondeterministic explainability method. One result is, however, inconsistent between datasets because the Random Logit metric ranks the Random method with CIFAR10 (Fig. 7) very highly.

For *robustness* metrics, we expect Sobel and Random to have poor rankings, because the maps are completely different each time for Random, which makes them very unstable, and Sobel detects new edges, being sensitive to local perturbations. Since Gaussian is insensitive to any perturbations of the input, it provides trivial constant explanations and is the optimal choice as stated in the authors' paper [38]. This is what turns out to be the case for Max-Sensitivity and Avg-Sensitivity. However, the Local Lipschitz Estimate rates Sobel as one of the best methods, which is not coherent with the purpose of the metric.

For *faithfulness* metrics, the general goal is that perturbing/adding the most significant pixels or areas of an image has the most noticeable effect on the prediction for the

best explainability method (and vice versa). However, in addition to being the group of metrics with the weakest correlations in Section 4.1, we observe that across the three figures, it is also the only family where the results do not agree at all for different models and datasets. All XAI methods are extremely close in rankings for VGG-16, and we observe a lot of crossovers between rankings for all metrics in the other two figures. In terms of dummy methods, since none of the three dummy methods aim to select the most significant pixels, we initially expect them to be consistently ranked poorly. However, we find that Pixel-Flipping and Selectivity assign a good ranking to the Random method using Resnet50, or that Gaussian is assigned the best or worst ranking by Monotonicity Arya depending on the dataset used. Overall, there is no faithfulness metric that consistently gives a bad ranking to the three dummy methods for all models and datasets used, contrary to what one would logically expect from a reliable method.

5 Discussion and Conclusion

Despite the existence of multiple quantitative metrics assessing the quality of saliency methods [3, 6], the evaluation of these methods remains challenging in the absence of ground truth explanations [27]. There are also very few objective criteria for choosing methods and metrics. To the best of our knowledge, no previous work exists that focuses on metric comparison instead of saliency method comparison. This paper compares 14 metrics by evaluating nine state-of-the-art XAI methods, applied to 2000 images from the ImageNet validation set and 2000 images from the CIFAR10 test set, using pretrained CNN models.

We studied the pairwise correlation of metrics through Kendall's τ_b, followed by an analysis of the baseline hyperparameter for the faithfulness metrics, and finally a reliability analysis of the ranking given by the metrics to the XAI methods with the addition of 3 dummy methods used to detect potential issues. From experimental results, we unveil for complexity metrics that Sparseness and Complexity are very highly correlated on all datasets and therefore potentially redundant with each other to represent complexity. For robustness metrics, we find out that Max- and Avg-Sensitivity are strongly correlated. Moreover, as Local Lipschitz Estimate rated unexpectedly highly the Sobel dummy method in the reliability analysis, it would favor the choice of either Max- or Avg-Sensitivity to represent robustness metrics. Model Parameter Randomization and Random Logit demonstrate a significant correlation between each other without being redundant in the randomization metrics, with no consistent flaws found by the use of dummy methods, both representing well different aspects of the randomization. Last, though faithfulness metrics are arguably the most widely used group in the state-of-the-art, this metric group does not find any redundant metrics or significantly correlated ones on all datasets. The reliability analysis shows that the ranks given to the methods vary enormously across datasets and models and that for these we also cannot find a faithfulness method that gives the assumed rank (worst rank) to the dummy methods used. All of this is in addition to the fact that the ranks given by the faithfulness metrics vary significantly depending on the baselines used, as shown in our experiments. It is therefore essential to include a variation of this hyperparameter to draw reliable conclu-

sions, and to be cautious when comparing a small number of faithfulness metrics, given the inconsistent results obtained in this work.

Since there is no ground truth for explainability, the families of metrics each evaluate a different aspect of an explainability method. In real experiments, between the families, it is up to the user to decide which types of metrics they want to focus on (e.g., favor robustness over complexity). Within a family, we show which metrics are correlated and may have less interest to compute in common. In addition, the flaws identified using dummy methods indicate to the user which metrics they may want to avoid. Future work could extend the paper to analyze reliability under vision transformers or explore the impact of other hyperparameters than the baseline (e.g., the size of patches).

Acknowledgments

The authors acknowledge the various sources of funding for this research. Julien Albert, Nassim Versbraegen, and Miriam Doh are supported by the Service Public de Wallonie Recherche under grant n°2010235-ARIAC by DIGITALWALLONIA4AI. Nassim Versbraegen is also funded by Innoviris Joint R&D project Genome4Brussels [2020 RDIR 55b to N.V.]. The Research Foundation for Industry and Agriculture, National Scientific Research Foundation (FRIA-FNRS) funded this research as grants attributed to Alexandre Englebert and Gilles Peiffer, consisting in Ph.D. financing. Géraldin Nanfack is funded by the EOS-VeriLearn project number 30992574 of the Fonds de la Recherche Scientifique (F.R.S-FNRS). Sédrick Stassin thanks the support of the E-origin project funded by the Walloon Region within the pole of logistics in Wallonia.

Computational resources have been provided by the supercomputing facilities of the Université catholique de Louvain (CISM/UCL) and the Consortium des Equipements de Calcul Intensif en Fédération Wallonie Bruxelles (CECI) funded by the Fonds de la Recherche Scientifique de Belgique (F.R.S.-FNRS) under convention 2.5020.11 and by the Walloon Region.

References

1. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. In: Neurips (2018)
2. Alvarez-Melis, D., Jaakkola, T.S.: On the robustness of interpretability methods. arXiv:1806.08049 (2018)
3. Alvarez-Melis, D., Jaakkola, T.S.: Towards robust interpretability with self-explaining neural networks. In: Neurips (2018)
4. Arya, V., Bellamy, R.K., Chen, P.Y., Dhurandhar, A., Hind, M., Hoffman, S.C., Houde, S., Liao, Q.V., Luss, R., Mojsilović, A., et al.: One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. arXiv:1909.03012 (2019)
5. 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), 1–46 (2015)
6. Bhatt, U., Weller, A., Moura, J.: Evaluating and aggregating feature-based model explanations. In: IJCAI (2020)

7. Burkart, N., Huber, M.F.: A survey on the explainability of supervised machine learning. JAIR **70**, 245–317 (2021)
8. Chalasani, P., Chen, J., Chowdhury, A.R., Wu, X., Jha, S.: Concise explanations of neural networks using adversarial training. In: ICML (2020)
9. Das, A., Rad, P.: Opportunities and challenges in explainable artificial intelligence (xai): A survey. arXiv:2006.11371 (2020)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
11. Englebert, A., Cornu, O., Vleeschouwer, C.D.: Poly-CAM: High resolution class activation map for convolutional neural networks. arXiv:2204.13359v2 (2022)
12. Fernandez, F.G.: Torchcam: class activation explorer. `https://github.com/frgfm/torch-cam` (March 2020)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
14. Hedström, A., Weber, L., Bareeva, D., Motzkus, F., Samek, W., Lapuschkin, S., Höhne, M.M.C.: Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations. arXiv:2202.06861 (2022)
15. Holm, S.: A simple sequentially rejective multiple test procedure. Scand. J. Stat. **6**, 65–70 (1979)
16. Jalwana, M.A., Akhtar, N., Bennamoun, M., Mian, A.: Cameras: Enhanced resolution and sanity preserving class activation mapping for image saliency. In: CVPR (2021)
17. Jiang, P.T., Zhang, C.B., Hou, Q., Cheng, M.M., Wei, Y.: Layercam: Exploring hierarchical class activation maps for localization. IEEE Trans. Image Process. **30**, 5875–5888 (2021)
18. Kendall, M.G.: The treatment of ties in ranking problems. Biometrika **33**(3), 239–251 (1945)
19. 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. arXiv preprint arXiv:2009.07896 (2020)
20. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images. Tech. Rep. 0, University of Toronto, Toronto, Ontario (2009)
21. Li, X.H., Shi, Y., Li, H., Bai, W., Cao, C.C., Chen, L.: An Experimental Study of Quantitative Evaluations on Saliency Methods. In: KDD (2021)
22. Montavon, G., Samek, W., Müller, K.R.: Methods for interpreting and understanding deep neural networks. Digit. Signal Process. **73**, 1–15 (2018)
23. Nguyen, A.p., Martínez, M.R.: On quantitative aspects of model interpretability. arXiv:2007.07584 (2020)
24. Petsiuk, V., Das, A., Saenko, K.: Rise: Randomized input sampling for explanation of black-box models. In: BMC (2018)
25. Phan, H.: huyvnphan/pytorch_cifar10 (Jan 2021). https://doi.org/10.5281/zenodo.4431043
26. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. IJCV **115**(3), 211–252 (2015)
27. Samek, W., Binder, A., Montavon, G., Lapuschkin, S., Müller, K.R.: Evaluating the visualization of what a deep neural network has learned. IEEE Trans. Neural Netw. Learn. Syst. **28**, 2660–2673 (2017)
28. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: ICCV (2017)
29. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. ArXiv:1312.6034 [Cs] (2014)
30. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)

31. Sixt, L., Granz, M., Landgraf, T.: When explanations lie: Why many modified bp attributions fail. In: ICML (2020)
32. Smilkov, D., Thorat, N., Kim, B., Viégas, F., Wattenberg, M.: Smoothgrad: removing noise by adding noise. arXiv:1706.03825 (2017)
33. Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for simplicity: The all convolutional net. arXiv:1412.6806 (2014)
34. Sturmfels, P., Lundberg, S., Lee, S.I.: Visualizing the impact of feature attribution baselines. Distill **5**(1), e22 (2020)
35. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: ICML (2017)
36. Van Rossum, G.: The Python Library Reference, release 3.8.2. Python Software Foundation (2020)
37. Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., Hu, X.: Score-cam: Score-weighted visual explanations for convolutional neural networks. In: CVPR Worshop on TCV (2020)
38. Yeh, C.K., Hsieh, C.Y., Suggala, A., Inouye, D.I., Ravikumar, P.K.: On the (in) fidelity and sensitivity of explanations. In: Neurips (2019)
39. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: CVPR (2016)
40. Zhou, J., Gandomi, A.H., Chen, F., Holzinger, A.: Evaluating the quality of machine learning explanations: A survey on methods and metrics. Electronics **10**(5), 593 (2021)