Saliency Maps Give a False Sense of Explanability to Image Classifiers: An Empirical Evaluation across Methods and Metrics

Hanwei Zhang ZHANG@DEPEND.UNI-SAARLAND.DE
Saarland University, Saarbrücken, Germany; Institute of Intelligent Software, Guangzhou, China

Felipe Torres FELIPE.TORRES@LIS-LAB.FR
LIS lab, École Centrale de Marseille, France

Holger Hermanns HERMANNS@CS.UNI-SAARLAND.DE
Saarland University, Saarbrücken, Germany

Editors: Vu Nguyen and Hsuan-Tien Lin

Abstract

The interpretability of deep neural networks (DNNs) has emerged as a crucial area of research, particularly in image classification tasks where decisions often lack transparency. Saliency maps have been widely used as a tool to decode the inner workings of these networks by highlighting regions of input images deemed most influential in the classification process. However, recent studies have revealed significant limitations and inconsistencies in the utility of saliency maps as explanations. This paper aims to systematically assess the shortcomings of saliency maps and explore alternative approaches to achieve more reliable and interpretable explanations for image classification models. We carry out a series of experiments to show that 1) the existing evaluation does not provide a fair nor meaningful comparison to the existing saliency maps; these evaluations have their implicit assumption and are not differentiable; 2) the saliency maps do not provide enough information on explaining the accuracy of network, the relationship between classes and the modification of the images.

Keywords: Interpretability; Saliency Maps; Image Classification

1. Introduction

Deep learning models achieved remarkable success in various machine learning tasks and their use is starting to pervade high-risk AI systems, such as autonomous driving and medical analysis, which urges the development of explainable AI to build trust and help in validating deep learning models.[1] According to the survey (Zhang et al., 2021), most of the research focuses on providing passive attribution as local explanations, *i.e.* providing insight into the decision-making process of the model by highlighting the most relevant regions of the input. Such explanations in computer vision are widely known as *saliency map*, providing visual indications what regions of an image contribute most to a decision. However, when we examine the various existing methods to generate saliency maps, we face severe difficulties in evaluating their adequacy as practical means to identify "the main elements of the decision taken", rooted in the diversity of definitions and proposals for methods and metrics.

Saliency maps are commonly used in weakly supervised object localization and segmentation (Zhang et al., 2017b; Zhou et al., 2016). In this paper, we call any method generating a saliency

[1] Article 86 of the new European AI Act, for instance, mandates that persons affected adversely by AI have the right to obtain a clear and meaningful explanations of the role of the AI system in the decision-making procedure and the main elements of the decision taken.

map a “saliency method”. The implicit assumption here is that the model’s prediction is based solely on the object itself. However, in reality, intrinsic object features such as color, shape, parts, and the background also contribute to recognition (Zhang et al., 2023). Saliency object detection and fixation (Bylinskii et al., 2018; Judd et al., 2012) are also related to localization but use human focus as ground truth. These evaluations assess how well the models align with human fixation but are not necessarily related to explainability. The ability to provide visual feedback is a crucial asset of saliency maps, albeit being challenging to evaluate quantitatively. Therefore evaluations based on human perception, such as crowd-sourced use cases (Zeiler and Fergus, 2014; Samuel et al., 2021) have been introduced. However, these methods are typically costly and subject to bias (Gonzalez-Garcia et al., 2018). Plus, the assumption that model predictions align with human perception can be misleading, as humans and networks do not necessarily focus on the same regions (Das et al., 2017). Discriminative ability (Li et al., 2019; Wang and Vasconcelos, 2023) suggests that explanations can be evaluated based on whether the saliency maps provide effective discriminative information. However, while this approach makes sense, it lacks generalisability. In practice, datasets with specific annotations are required for evaluating discriminative ability.

Since evaluations based on human perception are costly and subjective, mathematical evaluation appears as a more practical, stable, and objective approach for assessing how well saliency maps reveal model predictions. Researchers in the field of explainable AI are working to establish benchmarks for evaluating saliency maps based on several key metrics, including fidelity (Gomez et al., 2022; Han et al., 2022), robustness (Samek et al., 2016; Dasgupta et al., 2022), complexity (Bhatt et al., 2020; Nguyen and Martínez, 2020), and randomization (Adebayo et al., 2018b; Hedström et al., 2024), as indicated by recent work (Hedström et al., 2023). However, instead of a systematic evaluation across metrics, there is a tendency of cherry-picking a few metrics when proposing a new saliency method, or to only compare on toy datasets like MNIST or CIFAR-10, rather than on more realistic datasets such as ImageNet. A notable exception in this fragmented research landscape is the work by Li et al. (2021) that provides a comparative study of two particular saliency map methods across different metrics and on two realistic datasets. This is a highly valuable step towards clarifying the actual performance of saliency methods. However, the authors focus on comparing methods in terms of localization and false-positive results using the eBAM dataset, which contains both scene and object class labels rooted in human perception. As discussed above, we instead prefer mathematically defined quality indicators, such as fidelty and other metrics. However, this ambition is challenged further by the diverse purposes and scenarios for which saliency maps and assessment metrics are designed. Each of the existing mathematically defined metrics is based on different assumptions and experimental setups, complicating comparisons and making it difficult to define satisfactory explanations.

The present paper reports on orchestrated efforts to clarify the strengths and weaknesses of saliency maps as explanations in computer vision, choosing the archetypal task of image classification as the use case for investigation. By comparing the various saliency methods over various mathematically defined metrics, we systematically assess the limitations and deficiencies of saliency mapping as well as the evaluation metrics, offering both empirical evidence and insights to illuminate their shortcomings. Furthermore, our focus is on presenting compelling arguments regarding explainability in image classification tasks, thereby driving advancements in the field of interpretable AI. Our work can be considered to complement and complete the initial work of Li et al. (2021) (that is consistent with our findings) regarding the benchmarking of saliency methods as explanations in image classification. Furthermore, we present compelling arguments regarding explainability in image classification tasks, thereby driving advancements in the field of interpretable AI.

2. Preliminaries

Classification. Consider a classifier network $f : \mathcal{X} \to \mathbb{R}^C$ that maps an input image $\mathbf{x} \in \mathcal{X}$ to a logit vector $\mathbf{y} = f(\mathbf{x}) \in \mathbb{R}^C$, where $\mathbf{x}$ with size $W \times H \times 3$, $\mathcal{X}$ is the image space and C is the number of classes. We denote the predicted logit as $y_c = f(\mathbf{x})_c$, the predicted probability for the class c as $p_c = \mathrm{softmax}(\mathbf{y})_c := e^{y_c}/\sum_j e^{y_j}$, and the prediction function of the classifier as $\phi(\mathbf{x}; f) := \arg\max_k f(\mathbf{x})_k$. For layer l, we denote the corresponding feature map of the model by $A^l := f_l(\mathbf{x}) \in \mathbb{R}^{w_l \times h_l \times k_l}$, which is a tensor with spatial resolution $w_l \times h_l$ and k_l channels.

Post Hoc Explanation. Consider a post hoc explanation $\mathbf{e} := E(f; \mathbf{x}; c)$ to the black box model f recognizing the input $\mathbf{x}$ as class c generated by the explanation approach E, which we usually take to be a part (*i.e.* a selected subset) of the input. If the explanation $\mathbf{e}$ is a sufficient explanation, then we expect to have $g(f; \mathbf{x}; \mathbf{e}) \equiv g(f; \mathbf{x}; \mathbf{e} \cup \mathbf{e}')$, where $\mathbf{e}'$ is any subset of the input, and $g(f; \mathbf{x}; \mathbf{e})$ denotes a function to evaluate the performance of $\mathbf{e}$ w.r.t. f and $\mathbf{x}$. If the explanation $\mathbf{e}$ is a necessary explanation, then we expect to have $g(f; \mathbf{x}; \mathbf{e} \setminus \hat{\mathbf{e}}) \ll g(f; \mathbf{x}; \mathbf{e})$, where $\hat{\mathbf{e}}$ is any nonempty subset of $\mathbf{e}$. This formalization of post hoc explanation is inspired by Huang and Marques-Silva (2024).

Saliency Map as Explanations. Assume a saliency mapping approach S producing a saliency map $\mathbf{m} := S(f, \mathbf{x}, c)$, where $\mathbf{m}$ is a matrix of size $W \times H$ with values in $[0, 1]$, as a post hoc explanation to the black box model f predicting the input $\mathbf{x}$[2] as class c. The saliency map $\mathbf{m}$ indicates which pixel contributes to prediction and gives a value in $[0, 1]$ to indicate how much it contributes. For $\mathbf{m}$ to be a sufficient explanation, we need to have $\phi(\mathbf{x}; f) \equiv \phi(\mathbf{x} \odot \mathbf{m}; f) \equiv \phi(\mathbf{x} \odot \mathbf{m}'; f)$ for each $\mathbf{m}' \succeq \mathbf{m}$, where $\odot$ denotes the point-wise multiplication and $\succeq$ is the partial order obtained by pointwise lifting of $\geq$ to matrices. For $\mathbf{m}$ to be a necessary explanation, we require $\phi(\mathbf{x} \odot \mathbf{m}'; f) \neq \phi(\mathbf{x}; f)$ whenever $\mathbf{m}' \prec \mathbf{m}$. Owed to the quantification over candidate maps $\mathbf{m}'$ it is obviously hard to evaluate whether in reality a saliency map is a sufficient or necessary explanation. Plus the effect of pixels is not independent of each other. As a result, a series of tailored properties and corresponding metrics have been proposed to measure the quality of saliency maps as explanations approximately.

3. Saliency Methods

Unlike text and table data, images are composed of a collection of pixels with inherent connections among them. This connectivity is particularly relevant for Convolutional Neural Networks (CNNs), which identify patterns based on regions, *i.e.* subsets of these pixels. To account for the high dimensionality and pixel correlations in images, a series of specialized techniques have been developed other than general attribution methods, collectively referred to as saliency maps. We categorize these techniques into four families, which we introduce in the following sections.

[2] Given that $\mathbf{x} \in \mathbb{R}^{W \times H \times 3}$ and $\mathbf{m} \in \mathbb{R}^{W \times H}$, to perform point-wise multiplication $\odot$ between $\mathbf{x}$ and $\mathbf{m}$, we replicate $\mathbf{m}$ to match the shape of $\mathbf{x}$. For simplicity, this replication step is omitted in the notation.

3.1. Gradient-based Methods

Gradient-based methods (Adebayo et al., 2018a; Springenberg et al., 2015; Baehrens et al., 2010) uses the gradient of a target class score with respect to the input to measure the effect of different image regions on the prediction. Intuitively, the gradient indicates those input features that influence the output with respect to the corresponding class c, thus Simonyan et al. (2014) treats gradients as a saliency map, *i.e.* $m_i^c := \frac{\partial y_c}{\partial x_i}$. Inspired by DeconvNet (Zeiler and Fergus, 2014), Guided Backpropagation (GuidedBP) (Springenberg et al., 2015) permits only the flow of positive gradients. Therefore, we define the saliency map using intermediate recurrence, denoted as $R_i^l := (A_i^l > 0)(R_i^{l+1} > 0)R_i^{l+1}$ for the layer l. For the last layer L, we have $R_i^L := \frac{\partial y_c}{\partial A_i^L}$. To achieve implementation invariance, Integrated Gradients (IG) (Sundararajan et al., 2017) computes saliency maps by multiplying the input variable element-wise with the average partial derivatives as the input transitions from a baseline to its final value. The mathematical form of it can be $m_i^c := (x_i - \bar{x}_i) \int_{\alpha=0}^{1} \frac{\partial f(\bar{\mathbf{x}}+\alpha(\mathbf{x}-\bar{\mathbf{x}})_c)}{\partial x_i} \, d\alpha$, where $\bar{\mathbf{x}}$ is a baseline input. Other methods (Shrikumar et al., 2017; Zhang et al., 2017b; Bastings and Filippova, 2020), inspired by Layer-wise Relevance Propagation (LRP) (Bach et al., 2015), calculate the relevance scores by backpropagating the final prediction to the input using their specific rules. SmoothGrad (Smilkov et al., 2017) accumulate gradients into saliency maps. A different approach is to use adversarial attacks (Elliott et al., 2021; Jalwana et al., 2020).

Gradient-based methods treat each pixel as an element to calculate importance, resulting in a different texture compared to other families of methods. Researchers from Human-Computer Interaction (HCI) prefer saliency maps generated by gradient-based methods because they reveal the shape of the context within an image and are considered understandable and informative by that community (Wang and Yin, 2021). However, computer vision experts tend to criticize gradient-based methods because some of them fail to satisfy the fundamental property of implementation invariance (Sundararajan et al., 2017) and do not pass sanity checks (Adebayo et al., 2018b).

3.2. CAM-based Methods

Class activation maps (CAM) (Zhou et al., 2016) is a visualization method that highlights the image regions most relevant to a target class by a linear combination of feature maps. *CAM-based methods* (Selvaraju et al., 2017; Chattopadhay et al., 2018; Zhang et al., 2023), especially proposed for images, replicate this process without requiring additional training or specific architectures. Given a layer l and a class of interest c, CAM-based saliency maps generated can be formulated as $S_l^c(\mathbf{x}) := \sum_k w_k^c A_k^l$ where w_k^c are weights defining a linear combination over channels. Several variants use different definitions for the weights. Grad-CAM (Selvaraju et al., 2017) defines weights for any layer l as $w_k^c := GAP\left(\frac{\partial y_c}{\partial A_k^l}\right)$ where GAP is the global average pooling. Grad-CAM++ (Chattopadhay et al., 2018) defines weights as $w_k^c := \sum_i \sum_j \left(\left[\frac{\frac{\partial^2 y_c}{(\partial A_k^l)^2}}{2\frac{\partial^2 y_c}{(\partial A_k^l)^2} + \sum_i \sum_j (A^k \frac{\partial^3 y_c}{(\partial A_k^l)^3})} \right] ReLU\left(\frac{\partial y_c}{\partial A_k^l}\right) \right)$, where i and j denotes the spatial indices of the features. Score-CAM (Wang et al., 2020) defines weights as $w_k^c := softmax\left(f(\mathbf{x} \odot n(up(A^l)))_c - f(\bar{\mathbf{x}})_c\right)_k$, where $n(A) := \frac{A - minA}{maxA - minA}$ is a normalization of matrix A into $[0, 1]$, $up(\cdot)$ is an up-sampling function to make the feature has the same shape as images, and $\bar{\mathbf{x}}$ is baseline inputs. Because CAM-based methods consistently use bilinear up-sampling and max-min normalization, the resulting saliency maps are always smooth and have values between zero and one.

3.3. Occlusion-based Methods

Occlusion-based methods (Petsiuk et al., 2018; Fong and Vedaldi, 2017; Schulz et al., 2020) use a number of candidate masks, measure their effect on the prediction, then combine them in a single saliency map. LIME (Ribeiro et al., 2016) relies on superpixels to define the fundamental elements for perturbation and masking. We generate instances around $\mathbf{x}$ by randomly selecting nonzero elements of $\mathbf{x}$. Given a perturbed sample $\mathbf{z}' \in \{0,1\}^d$, which includes a fraction of the nonzero elements of $\mathbf{x}$, we reconstruct the sample in its original form $\mathbf{z} \in R^d$ and compute $f(\mathbf{z})$ to serve as the label for the explanation model. We then approximate the saliency map by using K-LASSO to solve the following problem $\mathbf{m}^c := \arg\min_{\mathbf{w}} \sum_{\mathbf{z},\mathbf{z}' \in \mathcal{Z}} \pi_{\mathbf{x}}(\mathbf{z})(f(\mathbf{z})_c - \mathbf{w}\mathbf{z}')^2 + \Omega(\mathbf{w})$, where $\pi_{\mathbf{x}}(\mathbf{z}) := \exp(-\frac{\sum(\mathbf{x}-\mathbf{z})^2}{\sigma^2})$, Ω is a complexity measure, σ is the width for the exponential kernel defined on L_2 distance function. RISE (Petsiuk et al., 2018) introduces a function $\mathcal{M} : \mathcal{X} \to \{0,1\}^{W \times H \times 3}$ to generate random binary mask with distribution $\mathcal{D}$, then the saliency map can be approximate $\mathbf{m}^c := \frac{1}{N}\frac{1}{\mathbb{E}[M]}\sum_{i=1}^{N} f(\mathbf{x} \odot M_i)_c M_i$, where M denotes a binary mask generated by function $\mathcal{M}$, *i.e.* averaging the influence of random masks weighted by the class score. *Meaningful perturbations* (Fong and Vedaldi, 2017) and *extremal perturbations* (Fong et al., 2019) directly optimize the mask in the image space by using gradients. They require a large number of parameters as well as regularizers, *e.g.* for smoothness. *Information bottleneck attribution* (IBA) (Schulz et al., 2020) optimizes the mask in the feature space as a tensor instead. Score-CAM (Wang et al., 2020) is also an occlusion-based method, using individual feature maps as candidate masks.

3.4. Learning-based Methods

Learning-based methods (Chang et al., 2019; Dabkowski and Gal, 2017; Phang et al., 2020; Zolna et al., 2020) use an additional network or branch and they train it on extra data and image-level labels to predict a saliency map given an input image. This includes for example generators (Chang et al., 2019) or auto-encoders (Dabkowski and Gal, 2017; Phang et al., 2020). This approach may be compared with weakly-supervised object detection (Bilen and Vedaldi, 2016), segmentation (Kolesnikov and Lampert, 2016) or instance segmentation (Ahn et al., 2019). Information Bottlenecks for Attribution (IBA) (Schulz et al., 2020) includes a learning-based approach in the feature space. Apart from requiring extra data, it is not satisfying in the sense that the learned decoder would need to be explained too. IBA applies a linear interpolation between signal of feature map A^l and noise $\epsilon \sim \mathcal{N}(\mu_{A^l}, \sigma^2_{A^l})$. This is represented as $\mathbf{z} := MA^l + (1-M)\epsilon$ where M denotes mask in $[0,1]$ with the same dimensions as A^l. Let $Q(\mathbf{z}) := \mathcal{N}(\mu_{A^l}, \sigma_{A^l})$ denote a variational approximation which assumes that all dimensions of $\mathbf{z}$ are distributed normally and independent as feature maps after linear or convolutional layers tend to have a Gaussian distribution. To train a distinct neural network by optimizing the cross-entropy loss $\mathcal{L}_{CE}$ of the classification and the mutual information loss, we can obtain the saliency map by $\mathbf{m} := \arg\min_M \mathcal{L}_{CE} + \beta\mathbb{E}_{A^l}\left[D_{KL}\left[P(\mathbf{z}|A^l)\|Q(\mathbf{z})\right]\right]$, where the parameter β controls the relative importance of the two objectives, and D_{KL} denotes the KL-divergence.

Discussion. Given the existing variety of the saliency mapping approaches reviewed above, the following difficulties are obviously present when aiming at a comparative evaluation: 1) The methods have *incompatible value ranges*; for instance, *CAM-based methods* generate single-channel saliency maps with values in the range $[0, 1]$ due to bilinear interpolation and normalized with max-min scaling, while the normalization is not detailed for *gradient-based methods* and the saliency map lacks a clear value limit. 2) The methods are characterised by *different textures*; for instance, some *Occlusion-based methods* like LIME (Ribeiro et al., 2016) use a group of pixels as the basic element resulting in discontinuous saliency maps, *CAM-based methods* treat a channel of feature maps as the basic element resulting in smooth saliency maps, *gradient-based methods* consider a channel of a pixel as the basic element resulting scattered saliency maps, (*learning-based methods* can generate continuous and smooth or binary saliency maps, dependent on how they are set up); 3) They *differ in scope*; for instance, *learning-based methods* are generally model-agnostic while *gradient-based* and *CAM-based methods* are mainly model-specific.

Takeaway: *Saliency maps as explanation providers for images differ from those for text or table data. The inputs are high-dimensional, making it challenging to disentangle the relationships between different pixels. Additionally, the variety of value ranges, textures, and scope makes it a priori difficult to compare saliency maps obtained from different method categories.*

4. Saliency Metrics

This section looks at the zoo of evaluation metrics and their relation to explanatory properties.

4.1. Fidelity

Fidelity, also known as faithfulness or correctness, is associated with the capability of the explanation to approximate the prediction of the black-box model (Carvalho et al., 2019). Fidelity is a crucial concept for an explanation since it is meant to answer the question of whether an explanation faithfully reproduces the dynamics of the underlying model (Alvarez Melis and Jaakkola, 2018). It is however difficult to capture formally.

Fidelity *vs*. Accuracy. To assess how well an explanation approximates a model's overall behavior, it is natural to associate fidelity with accuracy. A local explanation $\mathbf{m}$ with good fidelity satisfies *Local Accuracy* (LA) (Lundberg and Lee, 2017), *i.e.* $|f(\mathbf{x})_c - f(\mathbf{x} \odot \mathbf{m})_c| = 0$ for image classification. The corresponding metric is $LA := \frac{1}{N}\sum_{i=1}^{N} |f(\mathbf{x}_i)_c - f(\mathbf{x}_i \odot \mathbf{m}_i)_c|$. This implies that a good saliency map should retain the information necessary for the network to give exactly the same prediction as it would with the entire input. In computer vision, several metrics are proposed under the assumption that a better saliency map will increase confidence in the explanation. These metrics include *Average Drop* (AD), *Average Increase* (AI) (Chattopadhay et al., 2018) and *Average Gain* (AG) (Zhang et al., 2023). AD measures the negative impact on predicted class probabilities when the input image is masked with the saliency map, with a smaller value indicating better performance. It is calculated as $\mathrm{AD}(\%) := \frac{1}{N}\sum_{i=1}^{N} \frac{[f(\mathbf{x}_i)_c - f(\mathbf{x}_i \odot \mathbf{m}_i)_c]_+}{f(\mathbf{x}_i)_c} \cdot 100$. While AI and AG evaluate the positive effect, but AI only cares if the saliency map improves the prediction, calculated as $\mathrm{AI}(\%) := \frac{1}{N}\sum_{i}^{N} \mathbb{1}_{f(\mathbf{x}_i)_c < f(\mathbf{x}_i \odot \mathbf{m}_i)_c} \cdot 100$. Since a trivial increase in prediction can be detected by AI, it has a fundamental flaw. To address this issue, AG is proposed to evaluate the magnitude of the increase. It is calculated as $\mathrm{AG}(\%) := \frac{1}{N}\sum_{i=1}^{N} \frac{[f(\mathbf{x}_i \odot \mathbf{m}_i)_c - f(\mathbf{x}_i)_c]_+}{1 - f(\mathbf{x}_i)_c} \cdot 100$. Higher values of AI and AG indicate a more favorable saliency map.

Given the high-dimensionality of images and the intricate relationships among pixels, determining whether a good saliency map should improve accuracy or merely mirror changes in accuracy is challenging. LA does not differentiate between positive and negative influences on prediction. The popularity of AI and AD in computer vision suggests that enhancing accuracy is more attractive to researchers in the field.

Sufficiency and Necessity. To elaborate on the definition of fidelity, it is crucial to assess how well an explanation includes all important information, *i.e.* sufficiency, and accurately identifies truly insignificant features as insignificant, *i.e.* necessity (DeYoung et al., 2019; Luss et al., 2019). The metrics designed based on this principle and commonly used for evaluating saliency maps for images are *Insertion* (I) and *Deletion* (D) (Petsiuk et al., 2018). Insertion and deletion sequentially add or remove pixels in decreasing order of saliency and observe the effect on the prediction. Deletion measures the decrease in the probability p_c of class c when pixels are removed in decreasing order of saliency, with removal being equivalent to setting pixel values to zero. Conversely, insertion measures the increase in probability when adding pixels back. This process starts with a version of the image that is distorted by Gaussian blur relative to the original image. These operations result in out-of-distribution (OOD) images (Gomez et al., 2022), and the metrics tend to favor small and compact regions Zhang et al. (2023).

There are metrics satisfied both sufficient and necessary under the assumption *monotonicity* or *completeness*. Monotonic-increase Arya et al. (2019) measure the monotonic increase in classification probability due to incremental inclusion of pixels in increasing order of saliency map. It defines as $Monotonicity := Corr_{i\in\{1\cdots K\}}\left(\mathbf{m}_i, f(\mathbf{x}\odot\mathbf{m}_i)_c\right)$, where $Corr$ denotes the Pearson's correlation, and $\mathbf{m}_i$ denotes the i^{th} set of pixels in saliency maps sorted in increasing order. Both monotonicity (Arya et al., 2019; Nguyen and Martínez, 2020) and completeness (Shrikumar et al., 2017; Ancona et al., 2017; Sundararajan et al., 2017) assume that inputs with higher values in saliency maps have a greater impact on the output compared to those with lower values. Monotonicity considers the relative values of the saliency maps, whereas completeness considers their actual values. Existing experiments (Zhang et al., 2023) on failure cases of I/D show that when saliency maps highlight multiple regions, the values are high, but the impact on the output drops significantly. This occurs because the added pixels form discontinuous regions, hindering model recognition. This finding indicates that the assumptions of monotonicity and completeness are not realistic for image classifiers.

4.2. Robustness

The robustness, also known as sensitivity, evaluates the similarity of explanations under changes to the input point. We classify such robustness into two categories: 1) robustness with general perturbations measures the change in explanation as a function of change in input where this change in input is described by local perturbation of a given radius (Yeh et al., 2019; Agarwal et al., 2022; Montavon et al., 2018); 2) robustness with adversarial perturbations is a special case where the perturbations are intended to cause the change in explanation while the input is perturbed imperceptibly (Ghorbani et al., 2019). *Max sensitivity (MS)* (Yeh et al., 2019) measures the maximum change in the saliency map under a small general perturbation. It defines as $MS := \max_{\|\mathbf{x}'-\mathbf{x}\|\leq\epsilon}\|S(f,\mathbf{x},c)-S(f,\mathbf{x}',c)\|$ where ϵ is a spare of radius around the input to bound the perturbation. However, maximum sensitivity is unrepresentative of the behavior of the saliency map method if the saliency map varies smoothly in a region of perturbation, except for a few isolated points. To address it, Bhatt et al. (2020) proposed *average sensitivity (AS)* as $AS := \int_{\mathbf{x}' \in \mathcal{R}} \|S(f, \mathbf{x}, c) - S(f, \mathbf{x}', c)\| p(\mathbf{x}')\, d\mathbf{x}'$ where $\mathcal{R}$ is perturbation region, $p(\mathbf{x})$ is uniformly distributed in a sphere with radius ϵ around $\mathbf{x}$, *i.e.* $p := U(\mathbf{x}' | \|\mathbf{x}' - \mathbf{x}\| \leq \epsilon)$.

4.3. Complexity

The explanations should capture to what extent explanations are concise, *i.e.* few features are used to explain a model prediction. Bhatt et al. (2020) evaluates the complexity by entropy complexity, *i.e.* measuring the entropy of the fractional contribution distribution of the input element. It can be formulated as $Complexity := \mathbb{E}[-\ln(\mathbf{p}_m(\mathbf{x}))] = -\sum_i \mathbf{p}_m(\mathbf{x})_i \ln(\mathbf{p}_m(\mathbf{x})_i)$, where $\mathbf{p}_m(\mathbf{x})_i := \frac{|S(f,\mathbf{x},c)_i|}{\sum_j |S(f,\mathbf{x},c)_j|}$. While Nguyen and Martínez (2020) defines *effective complexity (EC)* as the minimum number of the important input element retrained in the inputs such that the conditional expected loss over model performance does not exceed a given tolerance. Then we have $EC := \arg\min_{k \in [1, \cdots, K]} |M_k|$ subject to $\mathbb{E}\left(\mathcal{L}(f(\mathbf{x}), f(\mathbf{x} \cdot M_k) | \mathbf{x} \odot M_k)\right)$, where M_k denotes the salience maps of top k set of pixels and $\mathcal{L}$ denotes the least squares loss. Besides, works (Samek et al., 2016) evaluate complexity as sparsity and they evaluate the number of nonzero coefficients in the function, for instance, L_1 regularization is used as sparseness in Chalasani et al. (2020).

4.4. Localization

Localization metrics are a specific type of metric for image explanations that evaluate how well saliency maps align with object bounding boxes. This property is highly valued by researchers in computer vision because localization is a crucial task in the field and is intuitive for human understanding. However, evaluating localization as part of the explanation implies that the model should interpret images in the same way humans do. Additionally, some researchers argue that focusing on localization as an explanation often overlooks the importance of background context (Shetty et al., 2019; Rao et al., 2022). According to the experiments conducted by Zhang et al. (2023), it is evident that the region of the object is not the only area influencing the network's decision. Thus, localization metrics based on ground truth bounding boxes may not be appropriate for evaluating the quality of saliency maps as explanations.

4.5. Sanity

Sanity checking, also known as Randomization, is another commonly used approach for evaluating saliency maps as explanations in computer vision. This metric implies that the quality of the saliency map should degrade as the evaluation problem deteriorates, induced by increasing randomization of model parameters (Adebayo et al., 2018b; Hedström et al., 2024; Sixt et al., 2020). So, the sensitivity of saliency maps with respect to model parameters is evaluated, assuming that saliency maps should exhibit similar sensitivity to noise as the model's accuracy does with respect to changes in parameters. Intuitively, fidelity entails sanity. Sanity checks are popular across computer vision, signifying the shortcomings of existing fidelity metrics and the conceptual confusion in the field.

Takeaway: *(1) Fidelity is the most critical property for evaluating saliency maps as explanations, but for images, the assumptions behind the metrics are not always valid. (2) Robustness and complexity are important mathematical properties that a good saliency map method should possess, but they do not directly address explanability. (3) Localization is not necessarily linked to explanation. (4) The sanity check is considered a subset concept of fidelity.*

5. Experiments

In this section, we first introduce the experimental setting. Next, we focus on the most important metrics, namely fidelity, and evaluate the performance of existing metrics to determine if they fulfill their intended purpose. We then assess the performance of existing metrics in terms of robustness and complexity, and explore the performance of saliency maps with respect to transformations.

5.1. Settings

Network and Dataset. We use the pretrained ResNet50 (He et al., 2016) from the Pytorch model zoo[3]. We use the validation set of ImageNet ILSVRC 2012 (Krizhevsky et al., 2012; Russakovsky et al., 2015), which contains $50,000$ images evenly distributed over the $1,000$ categories.

Saliency Methods. We select Gradient (Simonyan et al., 2014), IG (Sundararajan et al., 2017) and GuidedBackprop (Springenberg et al., 2015) as representatives of gradient-based methods; Grad-CAM (Selvaraju et al., 2017), Grad-CAM++ (Chattopadhay et al., 2018) and ScoreCAM (Wang et al., 2020) for CAM-based methods; RISE (Petsiuk et al., 2018) and LIME (Ribeiro et al., 2016) for occlusion-based methods; and IBA (Schulz et al., 2020) for learning-based methods. Additionally, we tested on Fake-CAM (Poppi et al., 2021), which defines a constant, entirely uninformed, saliency map by setting the top-left pixel to zero while keeping all other values uniform.

Normalization. The range of values in saliency maps significantly influences evaluation metrics that involve masking of images, *e.g.* AI, AG, or AD. Gradient-based methods generate saliency maps of size $W \times H$ in three unconstrained color dimensions. This needs to induce some unfairness when comparing to value-constraint single-channel saliency maps. To address this, we explored existing normalization practices for gradient-based methods, and decided to employ max-min normalization to confine values between zero and one, and then (if applicable) to aggregate three channels into one through the $\max$ operation, all this being inspired by the code of SmoothGrad[4],

Evaluation Metrics. We choose AI, AD (Chattopadhay et al., 2018), AG (Zhang et al., 2023), I, D (Petsiuk et al., 2018), Monotonicity (Nguyen and Martínez, 2020) and Completeness (Sundararajan et al., 2017) as representative of Fidelity metrics. We choose MS and AS (Yeh et al., 2019) as representative of robustness, and Sparsenes (Chalasani et al., 2020), Complexity (Bhatt et al., 2020) and EC (Nguyen and Martínez, 2020) as representative of Complexity. AI/AD/AG are implemented according to the definition, I/D is from the official implementation of RISE [5]. All the other implementations are taken from Quantus (Hedström et al., 2023)[6].

5.2. Fidelity

We evaluate popular fidelity metrics across the various saliency methods, results are shown in Table 1. Gradient-based methods perform poorly on fidelity metrics, which echoes the result of (Adebayo et al., 2018b) that they fail in sanity checking. On average, CAM-based methods can be considered to perform best. The fact that Fake-CAM ranks in the top two for AI/AD/Monotonicity

[3] https://pytorch.org/vision/0.8/models.html
[4] https://github.com/kazuto1011/smoothgrad-pytorch/blob/master/smooth_grad.py
[5] https://github.com/eclique/RISE
[6] https://github.com/understandable-machine-intelligence-lab/Quantus

Methods		AI	$\overline{\text{AD}}$	AG	I	$\overline{\text{D}}$	Monotonicity
Uninformed	Fake-CAM	47.3	99.2	1.6	60.1	61.6	0.57
Gradient	Gradient	2.2	4.0	0.0	51.4	81.3	-0.19
	IG	2.5	5.1	0.1	53.5	88.4	0.36
	GuidedBP	2.6	4.8	0.0	53.5	86.9	0.08
CAM	Grad-CAM	44.4	87.6	17.8	66.8	86.8	0.53
	Grad-CAM++	42.3	87.0	16.3	66.2	86.5	0.59
	Score-CAM	50.0	89.3	21.7	66.7	85.1	0.56
Occlusion	RISE	39.0	86.0	13.9	65.0	80.4	0.35
	LIME	9.7	31.8	2.9	64.7	84.0	0.24
Learning	IBA	36.8	82.3	14.5	66.5	85.7	0.55

Table 1: Evaluation of selected saliency mapping methods for different fidelity metrics w.r.t. the respective ground truth classes, where $\overline{\mathrm{AD}} = 100 - \mathrm{AD}$ and $\overline{\mathrm{D}} = 100 - \mathrm{D}$. This adjustment aligns all metrics so that higher values correspond to better performance.

highlights the shortcomings of current fidelity metrics for image classification. All the methods fail to satisfy the completeness assumption (not shown in the table).

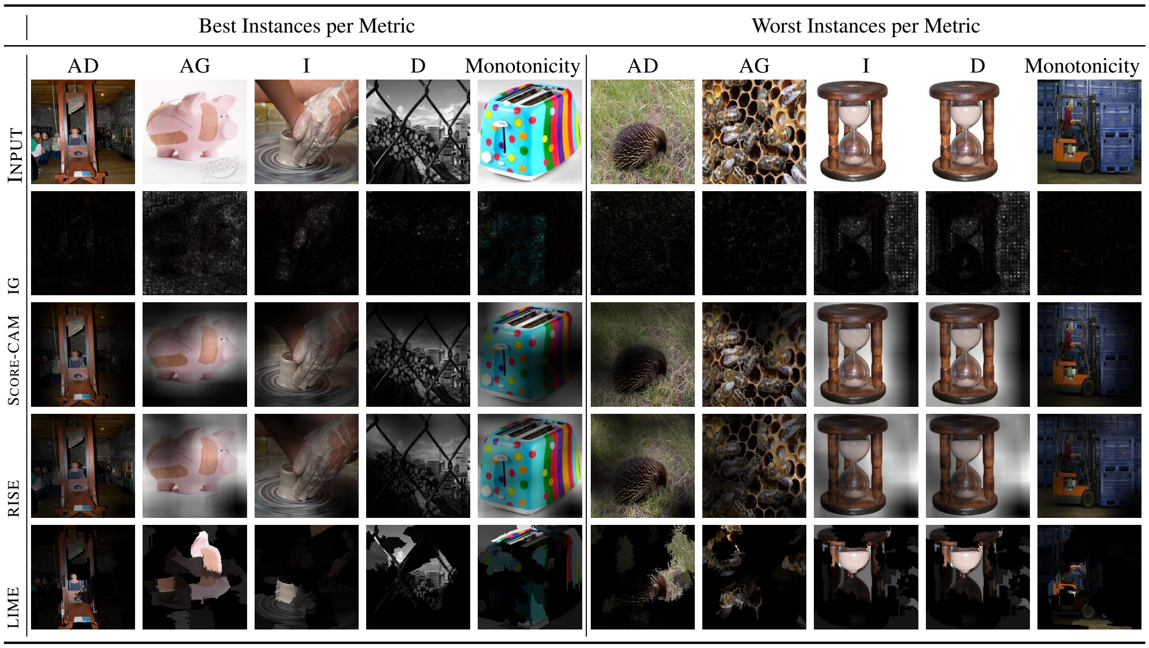

Figure 1: Images achieving best/worst performance w.r.t. each fidelity metric.

We now discuss individual images where best and worst performance are achieved: For each metric, we select the method performing best/worst on the corresponding metric and identify the images with the highest/lowest score. We display those together with the saliency maps obtained by various saliency methods in Figure 1[7]. Due to space constraints (and because the performances

[7] The visual appeareance of IG maps is worse than in the original paper due to different normalization techniques. This is a result of our intention to provide a mathematically sound comparison across methods rather than visualization.

within the set of gradient-based methods do not differ considerably, and similarly for CAM-based methods), we chose IG and Score-CAM as the respective representatives. From the visualizations presented, we observe that if considered together with the original images, gradient-based methods (IG) appear to provide the least information. However, if looked at without the original images, only the gradient-based method IG could be considered to produce meaningful structural information on its own, instead of highlighting continuous yet unspecific regions. Overall, fidelity metrics tend to favor saliency maps that highlight large, continuous regions. The worst cases for I and D metrics are rooted in the same image, where the IG method (errorneously) highlights a cylindrical shape on the right.

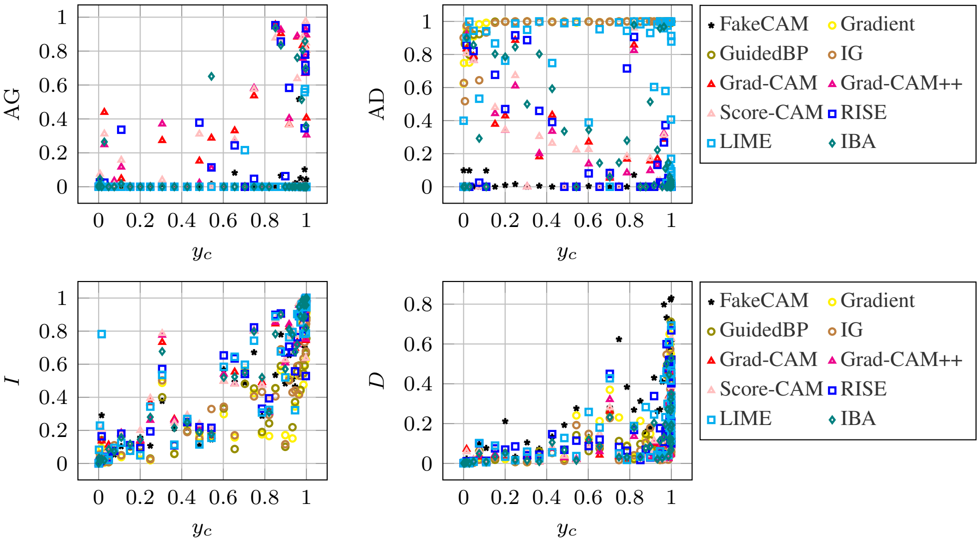

Figure 2: Relation between the probability y_c of ground truth and respecitve fidelity metrics.

Fidelity *vs*. Prediction. We now consider the probability value y_c with respect to the ground truth class c in relation to the corresponding fidelity metrics (*i.e.*, AG/AD/I/D). In Figure 2, we rank all the samples according to their probability values, downsampling uniformly to get the samples displayed. For CAM-based methods, there is a clear positive correlation between AG and y_c, and a negative correlation between AD and y_c (so a positive correlation for $\overline{\text{AD}}$). However, such correlations are not as apparent for other types of saliency methods. Generally, there is a positive correlation between I and y_c, and while there is also a positive correlation between D and y_c, it is not as pronounced as that for I.

To investigate the descriptive capabilities of saliency methods in correlation to their class specific behaviour, we generate attributions for the ground-truth label, the top-1 predicted class and the least probable class across the validation dataset. In this experiment, we expect to observe a steady decline in the quality of explanations as the behaviour is modified, particularly when performance shifts from 100% accuracy to 0. As shown in Table 2, the overall performance of interpretability metrics is optimal when considering the ground truth class, as expected. Moreover, performance is only slightly decreased when the attributions are generated for the top-1 predicted class.

When the target class is potential, fidelity metrics for saliency maps do not provide additional information. However, if the target class is impossible for the images (*i.e.*, least probable), we can distinguish this from fidelity metrics by observing higher values for AI/AD/D and extremely lower values for AG/I. It also indicates that existing fidelity metrics manage to approximate the model's prediction to some extent, but there is still room for improvement.

Methods		Ground Truth					Predicted Class					Least Probable				
		AI	$\overline{\text{AD}}$	AG	I	$\overline{\text{D}}$	AI	$\overline{\text{AD}}$	AG	I	$\overline{\text{D}}$	AI	$\overline{\text{AD}}$	AG	I	$\overline{\text{D}}$
Uninformed	Fake-CAM	47.3	99.2	1.6	60.1	61.6	46.8	99.5	1.8	66.9	57.4	53.3	100	0	0	100
Gradient	Gradient	2.2	4.0	0	51.4	81.3	0.1	0.7	0	55.8	78.3	100	100	0	0	100
	IG	2.5	5.1	0.1	53.4	88.4	0.1	1.3	0	57.8	85.9	99.9	99.9	0	0	100
	GuidedBP	2.6	4.8	0.0	53.5	86.9	0	0.9	0	57.8	84.9	100	100	0	0	100
CAM	Grad-CAM	44.4	87.6	17.8	66.8	86.8	38.7	85.9	18.7	75.2	85.7	100	100	0.1	0	100
	Grad-CAM++	42.3	87.0	16.3	66.2	86.5	37.3	85.7	17.3	74.5	85.3	79.5	88.5	0	0	100
	Score-CAM	50.0	89.3	21.7	66.7	85.1	45.2	88.2	23.0	75.7	83.8	90.4	94.4	0	0	100
Occlusion	RISE	39.0	86.0	13.9	65.0	80.4	27.1	79.6	12.1	71.9	79.5	81.0	88.3	0	0	100
	LIME	9.7	31.8	2.9	64.7	84.0	4.8	27.1	2.4	72.2	82.5	98.5	99.1	0	0	100
Learning	IBA	36.8	82.3	14.5	66.5	85.7	30.6	79.7	14.9	74.9	84.4	87.5	92.4	0	0	100

Table 2: Evaluation of fidelity metrics with respect to different classes.

Takeaway: *(1) Gradient-based methods perform badly in most fidelity metrics. (2) CAM-based methods (notably including Fake-CAM) perform best overall across fidelity metrics. (3) Among existing metrics of fidelity, AG and I (and to a lesser extent D) are not fooled by Fake-CAM. (4) I and D show a positive correlation with the probability of ground truth. (5) Overall, the fidelity metrics are not discrimitative enough to meaningfully tell apart performance of specific saliency methods.*

5.3. Robustness, Complexity, and Sensitivity

Robustness and Complexity. Although robustness and complexity do not directly demonstrate the explanatory capability of saliency maps, these mathematical attributes nevertheless appear crucial for saliency methods. We evaluate the corresponding metrics in Table 3. The smaller value of MS/AS/Complexity/EC is better while the larger value of Sparseness is better. The results indicate that most methods exhibit similar values with respect to these metrics, suggesting that these metrics cannot effectively distinguish which saliency methods outperform others. These results support our assertion that while robustness and complexity are important mathematical properties, they do not adequately address explainability.

Sensitivity to Transformation. Another meaningful dimension of evaluation for saliency maps focuses on their ability to explain modified images. Current model interpretation techniques assume that images for which we seek explanations are of high quality and typically in *natural* orientations. In this section, we investigate the interpretability properties of saliency maps generated on modified images. To do this, we select four popular image augmentation methods, *i.e.* MixUp, Resize, Rotation and Crop , and apply four variations of each to a small subset of $1,000$ randomly selected images from the ImageNet validation set. Additionally, we present results for the ground truth labels. MixUp (Zhang et al., 2017a) is an augmentation technique that generates a synthetic image by interpolating two images from different classes. The label for this synthetic image is a one-hot encoding representing the probability of belonging to the two different categories.

Methods		MS	AS	Sparseness	Complexity	EC
Uninformed	Fake-CAM	0.96	0.96	0.0	10.8	50175.0
Gradient	Gradient	0.93	0.91	42.4	10.5	50174.9
	IG	0.95	0.92	48.4	10.4	50174.8
	GuidedBP	1.22	1.14	39.3	10.5	50174.9
CAM	Grad-CAM	0.92	0.92	39.6	10.5	49540.4
	Grad-CAM++	0.93	0.92	36.7	10.6	50055.2
	Score-CAM	0.93	0.92	36.7	10.6	50060.8
Occlusion	RISE	0.93	0.92	26.0	10.7	50175.0
	LIME	0.93	0.93	72.3	9.7	26506.4
Learning	IBA	0.91	0.91	44.6	10.5	50172.6

Table 3: Evaluation of robustness and complexity metrics.

Table 4 presents these results. When using Fake-CAM as a baseline, we observe that AI increases significantly with Rotation, while AD drops dramatically with Resize. Values of I generally decrease while values of D increase, with less magnitude in Crop. These changes are linked to the variation in prediction probability y_c for the transformations. The performance of gradient-based methods varies greatly depending on the transformation, whereas CAM-based methods perform more stably. Regarding the metrics, we find that only AG remains stable despite the transformations.

Methods	MixUp					Resize					Rotation					Crop				
	AI	$\overline{\text{AD}}$	AG	I	$\overline{\text{D}}$	AI	$\overline{\text{AD}}$	AG	I	$\overline{\text{D}}$	AI	$\overline{\text{AD}}$	AG	I	$\overline{\text{D}}$	AI	$\overline{\text{AD}}$	AG	I	$\overline{\text{D}}$
Fake-CAM	49.6	97.8	0.6	20.9	86.6	47.4	2.0	90.5	24.8	86.2	65.6	97.0	4.0	24.3	82.6	47.5	98.5	1.7	49.9	71.2
Gradient	40.8	44.6	0.1	17.6	95.6	34.7	37.8	0.1	22.0	95.2	8.5	13.2	0.0	18.9	95.8	3.0	5.2	0.0	43.3	92.5
IG	39.7	43.7	0.1	18.8	95.3	34.2	37.3	0.0	22.2	95.1	7.7	11.8	0.0	21.8	96.7	3.1	5.5	0.0	43.5	92.8
GuidedBP	43.1	47.4	0.1	17.2	95.8	34.8	38.2	0.2	21.7	95.2	9.1	14.1	0.1	19.4	96.9	3.1	5.7	0.0	42.9	93.3
Grad-CAM	64.8	84.8	7.0	25.1	96.3	47.7	71.7	6.3	31.3	93.8	54.2	79.2	9.7	29.5	96.9	37.3	78.2	11.4	55.3	78.8
Grad-CAM++	54.1	79.1	6.6	24.9	96.0	46.2	71.0	5.8	30.9	91.4	51.7	78.8	11.5	29.1	96.8	35.5	76.0	10.3	52.4	90.4
Score-CAM	61.5	83.4	8.3	24.1	96.0	42.9	75.7	11.5	36.7	91.5	57.4	81.8	14.9	29.5	96.5	41.7	80.4	14.5	53.8	89.5
RISE	55.7	79.6	5.3	22.1	93.7	39.6	66.0	4.6	26.4	92.2	45.4	73.6	8.6	27.7	94.5	27.8	69.4	7.5	51.9	84.2
LIME	41.9	51.1	0.9	23.1	95.3	29.9	40.6	2.7	27.2	93.1	15.3	26.2	1.5	27.8	95.1	7.2	18.4	1.5	53.8	88.6
IBA	54.6	77.1	5.2	19.3	93.6	33.6	57.4	4.9	26.9	93.1	45.3	72.4	9.2	29.3	96.1	26.3	65.0	7.2	54.0	89.4

Table 4: Evaluation of fidelity metrics of saliency maps for transformed input images.

Takeaway: *(1) The metrics of robustness and complexity primarily identify extreme cases and lack differentiation in most saliency methods. (2) All saliency methods are sensitive to transformation, with gradient-based methods being particularly sensitive. (3) Fidelity metrics are also sensitive to transformation, though AG is the most stable among them.*

6. Conclusion

In this work, we have focussed on image classification to review existing research on saliency maps. Specifically, we examine the properties that researchers believe saliency maps should possess as explanations. We highlight the challenges of comparing saliency maps in the field of image analysis and attempt to normalize them for fair comparisons. We evaluate these metrics and conduct experiments, revealing that not all concepts are essential in this field such as localization. Furthermore, existing metrics turn out to not always be realistic for images, such as Monotonicy and Completeness. We have pinpointed obvious room for improvement regarding the metrics. On the other hand, our empirical results also indicate that saliency methods often lack sufficient class-specific information. We strongly encourage researchers in the field to develop saliency maps that not only highlight where the models assign importance to, but also clarify what factors contribute to these predictions and why. Notably, the results of Li et al. (2021) support our deeper observation that fidelity metrics are not sufficiently discriminative to effectively differentiate the performance of specific saliency methods. Their experiments use completely different datasets and an additional network, which suggests that our conclusions are likely generalising to other models and datasets as well.

Acknowledgement This work has received financial support by VolkswagenStiftung as part of Grant AZ 98514 – EIS and by DFG under grant No. 389792660 as part of TRR 248 – CPEC. This work also has received funding from the UnLIR ANR project (ANR-19-CE23-0009), the Excellence Initiative of Aix-Marseille Universite - A*Midex, a French“Investissements d'Avenir programme” (AMX-21-IET-017).

References

Julius Adebayo, Justin Gilmer, Ian J. Goodfellow, and Been Kim. Local explanation methods for deep neural networks lack sensitivity to parameter values. *ICLR Workshop*, 2018a.

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018b.

Chirag Agarwal, Nari Johnson, Martin Pawelczyk, Satyapriya Krishna, Eshika Saxena, Marinka Zitnik, and Himabindu Lakkaraju. Rethinking stability for attribution-based explanations. *arXiv preprint arXiv:2203.06877*, 2022.

Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *CVPR*, 2019.

David Alvarez Melis and Tommi Jaakkola. Towards robust interpretability with self-explaining neural networks. *Advances in neural information processing systems*, 31, 2018.

Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. *arXiv preprint arXiv:1711.06104*, 2017.

Vijay Arya, Rachel KE Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Q Vera Liao, Ronny Luss, Aleksandra Mojsilović, et al. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. *arXiv preprint arXiv:1909.03012*, 2019.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 2015.

David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. How to explain individual classification decisions. *J. MLR*, 2010.

Jasmijn Bastings and Katja Filippova. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? *EMNLP Workshop*, 2020.

Umang Bhatt, Adrian Weller, and José MF Moura. Evaluating and aggregating feature-based model explanations. *arXiv preprint arXiv:2005.00631*, 2020.

Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *CVPR*, 2016.

Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *IEEE transactions on pattern analysis and machine intelligence*, 41(3):740–757, 2018.

Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832, 2019.

Prasad Chalasani, Jiefeng Chen, Amrita Roy Chowdhury, Xi Wu, and Somesh Jha. Concise explanations of neural networks using adversarial training. In *ICML*, pages 1383–1391. PMLR, 2020.

Chun-Hao Chang, Elliot Creager, Anna Goldenberg, and David Duvenaud. Explaining image classifiers by counterfactual generation. *ICLR*, 2019.

A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In *WACV*, 2018.

Piotr Dabkowski and Yarin Gal. Real time image saliency for black box classifiers. *NIPS*, 2017.

Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? *CVIU*, 163:90–100, 2017.

Sanjoy Dasgupta, Nave Frost, and Michal Moshkovitz. Framework for evaluating faithfulness of local explanations. In *ICML*, pages 4794–4815. PMLR, 2022.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. Eraser: A benchmark to evaluate rationalized nlp models. *arXiv preprint arXiv:1911.03429*, 2019.

Andrew Elliott, Stephen Law, and Chris Russell. Explaining classifiers using adversarial perturbations on the perceptual ball. In *CVPR*, 2021.

Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2950–2958, 2019.

Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *ICCV*, 2017.

Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3681–3688, 2019.

Tristan Gomez, Thomas Fréour, and Harold Mouchère. Metrics for saliency map evaluation of deep learning explanation methods. In *International Conference on Pattern Recognition and Artificial Intelligence*, pages 84–95. Springer, 2022.

Abel Gonzalez-Garcia, Davide Modolo, and Vittorio Ferrari. Do semantic parts emerge in convolutional neural networks? *International Journal of Computer Vision*, 126:476–494, 2018.

Tessa Han, Suraj Srinivas, and Himabindu Lakkaraju. Which explanation should i choose? a function approximation perspective to characterizing post hoc explanations. *Advances in neural information processing systems*, 35:5256–5268, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

Anna Hedström, Leander Weber, Daniel Krakowczyk, Dilyara Bareeva, Franz Motzkus, Wojciech Samek, Sebastian Lapuschkin, and Marina Marina M.-C. Höhne. Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond. *Journal of Machine Learning Research*, 24(34):1–11, 2023.

Anna Hedström, Leander Weber, Sebastian Lapuschkin, and Marina Höhne. Sanity checks revisited: An exploration to repair the model parameter randomisation test. *arXiv preprint arXiv:2401.06465*, 2024.

Xuanxiang Huang and Joao Marques-Silva. On the failings of shapley values for explainability. *International Journal of Approximate Reasoning*, page 109112, 2024.

Mohammad A. A. K. Jalwana, Naveed Akhtar, Mohammed Bennamoun, and Ajmal Mian. Attack to explain deep representation. In *CVPR*, 2020.

Tilke Judd, Frédo Durand, and Antonio Torralba. A benchmark of computational models of saliency to predict human fixations. In *MIT Technical Report*, 2012.

Alexander Kolesnikov and Christoph H. Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *ECCV*, 2016.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *NIPS*, 2012.

Ruiyi Li, Yangzhou Du, Zhongchao Shi, Yang Zhang, and Zhiqiang He. Discrimination assessment for saliency maps. In *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8*, pages 628–636. Springer, 2019.

Xiao-Hui Li, Yuhan Shi, Haoyang Li, Wei Bai, Caleb Chen Cao, and Lei Chen. An experimental study of quantitative evaluations on saliency methods. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 3200–3208, 2021.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *NIPS*, 2017.

Ronny Luss, Pin-Yu Chen, Amit Dhurandhar, Prasanna Sattigeri, Karthik Shanmugam, and Chun-Chen Tu. Generating contrastive explanations with monotonic attribute functions. *arXiv preprint arXiv:1905.12698*, 3, 2019.

Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15, 2018. ISSN 1051-2004.

An-phi Nguyen and María Rodríguez Martínez. On quantitative aspects of model interpretability. *arXiv preprint arXiv:2007.07584*, 2020.

Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *BMVC*, 2018.

Jason Phang, Jungkyu Park, and Krzysztof J Geras. Investigating and simplifying masking-based saliency methods for model interpretability. *arXiv preprint arXiv:2010.09750*, 2020.

Samuele Poppi, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Revisiting the evaluation of class activation mapping for explainability: A novel metric and experimental analysis. In *CVPR*, pages 2299–2304, 2021.

Sukrut Rao, Moritz Böhle, and Bernt Schiele. Towards better understanding attribution methods. In *CVPR*, pages 10223–10232, 2022.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *SIGKDD*, KDD '16, 2016. ISBN 9781450342322.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 115(3):211–252, 2015.

Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, 28(11):2660–2673, 2016.

Sam Zabdiel Sunder Samuel, Vidhya Kamakshi, Namrata Lodhi, and Narayanan C Krishnan. Evaluation of saliency-based explainability method. *arXiv preprint arXiv:2106.12773*, 2021.

Karl Schulz, Leon Sixt, Federico Tombari, and Tim Landgraf. Restricting the flow: Information bottlenecks for attribution. *arXiv preprint arXiv:2001.00396*, 2020.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *CVPR*, 2017.

Rakshith Shetty, Bernt Schiele, and Mario Fritz. Not using the car to see the sidewalk–quantifying and controlling the effects of context in classification and segmentation. In *CVPR*, pages 8218–8226, 2019.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *ICML*, 2017.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *ICLR Workshop*, 2014.

Leon Sixt, Maximilian Granz, and Tim Landgraf. When explanations lie: Why many modified bp attributions fail. In *ICML*, pages 9046–9057. PMLR, 2020.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda B. Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *CoRR*, abs/1706.03825, 2017.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. Striving for simplicity: The all convolutional net. *ICLR*, 2015.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *ICML*, 2017.

Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-CAM: Score-weighted visual explanations for convolutional neural networks. In *CVPR Workshop*, 2020.

Pei Wang and Nuno Vasconcelos. A generalized explanation framework for visualization of deep learning model predictions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Xinru Wang and Ming Yin. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*, pages 318–328, 2021.

Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David I Inouye, and Pradeep K Ravikumar. On the (in) fidelity and sensitivity of explanations. *Advances in Neural Information Processing Systems*, 32, 2019.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 818–833. Springer, 2014.

Hanwei Zhang, Felipe Torres, Ronan Sicre, Yannis Avrithis, and Stephane Ayache. Opti-cam: Optimizing saliency maps for interpretability. *arXiv preprint arXiv:2301.07002*, 2023.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017a.

Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *IJCV*, 126:1084–1102, 2017b.

Yu Zhang, Peter Tiňo, Aleš Leonardis, and Ke Tang. A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(5):726–742, 2021.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016.

Konrad Zolna, Krzysztof J. Geras, and Kyunghyun Cho. Classifier-agnostic saliency map extraction. *CVIU*, 196:102969, 2020. ISSN 1077-3142.