Look at the Variance! Efficient Black-box Explanations with Sobol-based Sensitivity Analysis

Abstract

We describe a novel attribution method using Sobol indices derived from sensitivity analysis. Beyond modeling the individual contributions of image pixels, Sobol indices provide an efficient way to capture higher-order interactions between image regions and their contributions to the system's prediction through the lens of variance. We further describe an approach that makes the computation of these indices efficient for high dimensional problems by using perturbation masks coupled to efficient estimators to handle the high dimensionality of images. Importantly, we show that the proposed method leads to favorable scores on standard benchmarks for vision (and language models) while drastically reducing the computing time compared to other black-box methods – even approaching the accuracy of state-of-the-art white-box methods which require access to internal representations. The code to replicate our results is available: https://anonymous.4open.science/r/Sobol-Attribution-Method-AF4B

1   Introduction

Deep neural networks are now being deployed in numerous domains including medicine, transportation, security or finances with broad societal implications [1, 2]. Yet, these networks have become nearly inscrutable, and for most real-world applications, these systems are used to make critical decisions – often without any explanation. In recent years, numerous explainability methods have been proposed [3, 4, 5, 6, 7, 8, 9, 10]. In addition to helping improve people's trust in these systems, these methods have helped identify and correct biases in datasets [6, 7] – which has helped, in turn, improve these systems' robustness and accelerate their broad deployment. An important limitation of these explainability methods is that they require access to the system's internal states including hidden layer activations or input gradients [3, 7, 8, 11]. As a result, these white-box methods cannot be applied in the most general situations for which the internal states of these networks are not publicly accessible. Only a few black-box methods have been proposed in this common and challenging context [4, 6, 10]. It is thus critical to develop more general methods that can reliably interpret and characterize the underlying decision processes of a wider array of models.

Common approaches to explaining a model's prediction consists in attributing a score for each input dimension such as image pixels for computer vision systems or individual words for natural language processing. Shown in Figure 1 is an example image and associated importance map for an image categorization system, whereby scores for individual pixels are displayed as heatmaps where hotter locations correspond to pixels that contribute the most to the system's final prediction. In the context of black-box models, a core challenge is to derive these heatmaps using only the output predictions available through the forward pass. A naive approach consists in applying a given perturbation at a specific location on the input image to then measure how the corresponding prediction is affected. In the case of image classifiers, a pixel is replaced by a baseline intensity such as a pure black or

Figure 1: **(Left) Sobol attribution method overview:** From an input image and a set of real-valued masks $\mathbf{M}$ drawn from a Quasi-Monte Carlo (QMC) sequence, we form perturbed inputs $\mathbf{X}$ using a perturbation function $\mathbf{\Phi}$ (shown here for a simple *Inpainting* function). From the model predictions to these image perturbations, we estimate total order Sobol indices $\mathcal{S}_{T_i}$ to characterize the importance of each pixel: the first-order Sobol indices $\mathcal{S}_i$ reflect the importance of a pixel in isolation while $\mathcal{S}_{T_i}$ encompasses the effects of first and all higher-order non-linear interactions between pixels. The indices are introduced in Section 3.2 and their estimator in Section. 3.3. **(Right)** Sample explanations for ResNet50V2. Pixel regions deemed important in $\mathcal{S}_{T_i}$ but not in $\mathcal{S}_i$ are those that contribute to the model decision – not individually – but through their interactions with other image regions (e.g., lynx tips are important but conditional on the presence of other features such as eyes).

gray value; In the case of language models, individual words are instead removed from the text [12, 13]. However, evaluating the impact of these perturbations one dimension at a time fails to unravel the non-linear interactions between input variables that are known to prevail in a complex system such as a deep neural network. Also, applying perturbations at multiple locations at a time quickly becomes combinatorially intractable. Recent methods have tried to address these issues by grouping dimensions together, for instance pixels belonging to a same neighborhood of the image (such as a superpixel) [4] and additionally sampling perturbation masks that affect multiple regions of the input at a time [6, 10]. A first limitation includes the use of Monte-Carlo sampling methods which require a high number of forward passes – rendering these approaches very slow. A second limitation of recent methods is that they rely on relatively simple perturbation such as flipping pixels on or off [4, 6, 10]. This severely constrains the space of perturbations considered and limits the efficiency with which the space of perturbations can be explored.

We address these limitations by introducing an attribution method that leverages variance-based Sensitivity Analysis techniques, and more specifically Sobol indices [14]. It was introduced to identify the input variables of a non-linear mathematical system that have the most influence on the variance of the output. It was traditionally used in physics or economics to estimate the part of variance induced by each coalition of variables on the output. Our main contribution is to propose a general framework to adapt Sobol indices in order to explain predictions from black-box models through the use of perturbation masks. Contrarily to previous work, our attribution method is designed to support real-valued intensity perturbation to generate a continuous range of perturbations. Additionally, our attribution method is faster than other comparable state-of-the-art methods as we leverage an efficient estimator introduced in the Sensitivity Analysis literature and Quasi-Monte Carlo sequences for sampling. Our method can thus be efficiently applied in the context of high-dimensional inputs such as the pixels space of a modern image classifier. Another benefit of our method is that it allows to characterize not only first-order interaction between image regions but also higher-order interactions by decomposing the variance of the model's prediction using the Sobol indices. We run extensive experiments to demonstrate the interest of our method on several recent benchmarks that are complementary such as Pointing Game and Deletion. We mainly focus on image classifiers but also report results on text classifiers.

2   Related work

Attribution methods for white-box models   A widely used family of attribution methods uses gradient computations via the backpropagation algorithm. These methods aimed to probe a model's

decision in the local neighborhood of the input image. The first such method was introduced in [15] and later refined in [3] in the context of deep convolutional networks. The idea is to backpropagate the gradient from a classification unit all the way down to the input image. The resulting gradient image indicates which pixels affect the decision score the most. Other gradient-based methods, DeConvNet [4] and Guided Backprop [5], were specifically developed to deal with specific activation functions such as ReLU. However, this family of methods focuses on the local influence of pixels, which can be a problem. For instance, it has been shown that gradients often vanish when the associated prediction score is high. Integrated Gradient [9] partially addresses this issue by accumulating gradients along a straight interpolation path from a baseline state to the original image. The baseline is often set to zero, representing the complete absence of information.

Another class of methods [8, 16] uses the gradients differently – by iteratively optimizing a perturbation mask over regions of the image. It leads to a different explanation by outputting the optimized perturbation mask. However, such an approach turns out to be computationally very heavy – limiting its broad applicability. In comparison, our proposed approach also works on image regions but it does not require any gradient computation – and it is thus computationally more efficient.

Another family of attribution methods relies on a network's activations. Popular examples include CAM [17] which computes an attribution score based on a weighted sum of feature channels activities – right before the classification layer. GradCAM [7] extends it by using the gradients to weights each feature channels and refine the attribution scores. These methods also work on image regions because they are computed on the output of feature maps which are themselves at a coarser resolution than the original image. In comparison, our proposed approach is model-agnostic and hence does not require access to internal computations.

Attribution methods for black-box models   Most similar to our approach are attribution methods that can be used to explain the predictions of truly black-box models. They probe the model responses to perturbations over image regions and combine the resulting predictions into an influence score for each individual pixel. The simplest method, Occlusion [4], occludes individual image regions – one at a time – with an occluding mask set to a baseline value and assigns the corresponding prediction scores to all pixels within the occluded region. Occlusion does not account for the joint (higher-order) interactions between multiple image regions. For instance, occluding two image regions – one at a time – may only decrease the model's prediction minimally (say a single eye or mouth component on a face) while occluding these two regions together may yield a substantial change in the model's prediction if these two regions interact non-linearly as is expected for a deep neural network.

Our work, together with related methods such as LIME [6] and RISE [10], addresses this problem by randomly perturbing the input image in multiple regions at a time. Obviously, perturbing multiple image locations simultaneously leads to a combinatorial explosion in the number of combinations and methods have been proposed to make these approaches more tractable. For instance, a popular method, LIME [6] takes superpixels as regions to perturb instead of individual pixels. An influence score is then computed for a set of connected pixel patches indicating how strong a patch is correlated to the model predictions.

RISE [10] relies on Monte Carlo sampling to generate a set of binary masks, each value in the masks representing a pixel region. By probing the model with the randomly masked version of the input, it produces a heatmap explaining the model's prediction by taking the average of the masks weighted by their associated prediction score. Instead of using binary masks, our method considers a continuous range of perturbations which allows for a finer exploration of the model response.More importantly, the aforementioned methods lack a rigorous framework. Here, we introduce a theoretical framework that decomposes the influence score of each individual region between orders of influence. The first-order approximates Occlusion [4] by considering the influence of one region at a time, while the second-order considers two regions at a time. The decomposition also includes higher-orders.

Variance-based sensitivity analysis   Our attribution method builds on the variance-based sensitivity analysis framework. The approach was introduced in the seventies [18] and reached a cornerstone with the Sobol indices [19]. Sobol indices are currently used in many fields such as the analysis of physical phenomena. They are used to identify the input dimensions that have the most influential effect on the output of a model or a mathematical system. Several statistical estimators of these indices are available [20, 21, 22, 23, 24] and have asymptotic guarantees [22, 25, 26]. We build on this literature by adapting these Sobol indices in the context of black-box models to compute the influence of regions of an image on the output predictions using perturbation masks.

3 Sobol attribution method

In this work, we consider a feature attribution problem which consists in quantifying the contribution of a collection of real-valued features $\tilde{\boldsymbol{x}} = (\tilde{x}_1, ..., \tilde{x}_d)$ with respect to a model decision. Specifically, we consider a **black-box** decision function $\boldsymbol{f} : \mathcal{X} \to \mathbb{R}^k$ whose internal states and analytical form are unknown. (For instance, $\boldsymbol{f}$ can be a probability score for the input to belong to a specific class.) Contrary to previous attribution methods, our goal is not only to quantify the independent importance of each feature to the decision score $\boldsymbol{f}(\tilde{\boldsymbol{x}})$, but also the conjunctive contribution of these features. To capture these higher-order interactions, our method consists in estimating the Sobol indices of the collection $\tilde{\boldsymbol{x}}$ by applying random perturbations to the original input and measuring their effects on the prediction of the black-box model as can be seen in Figure 1.

Considering variations of $\boldsymbol{f}(\tilde{\boldsymbol{x}})$ in response to meaningful perturbations of the input $\tilde{\boldsymbol{x}}$ is a natural way to interpret the local behavior of a the decision function $\boldsymbol{f}$ around $\tilde{\boldsymbol{x}}$. Several methods build on this idea, e.g., by removing one or a group of input features [4, 6, 8, 10, 16] or by back-propagating the gradient in the input space through the model [3, 9, 27, 7]. Most of these methods use the model's internal representations and/or require computing the gradient w.r.t. input, which makes them unusable in a black-box setting. Moreover, these methods focus on estimating the intrinsic contribution of each feature, neglecting the combinatorial components. Our method applies perturbations directly on the input in order to deal with a black-box scenario, and allows us to estimate higher-order interactions between $\tilde{x}_1, ..., \tilde{x}_d$ that contributes to $\boldsymbol{f}(\tilde{\boldsymbol{x}})$.

3.1 Random Perturbation

Formally, let us define a probability space $(\Omega, \mathcal{X}, P)$ of possible input perturbations and a random vector $\boldsymbol{X} = (\boldsymbol{X}_1, ..., \boldsymbol{X}_d)$ as a stochastic perturbation of the original input $\tilde{\boldsymbol{x}}$. There are several ways to define random perturbations which explore differently the data manifold around $\tilde{\boldsymbol{x}}$. For instance, we can consider the perturbation mask operator $\boldsymbol{\Phi} : \mathcal{X} \times \mathcal{M} \to \mathcal{X}$ which combines a stochastic mask $\mathbf{M} = (\boldsymbol{M}_1, ..., \boldsymbol{M}_d)$ (i.e., an i.i.d sequence of real-valued random variables on $[0, 1]^d$) with the original input $\tilde{\boldsymbol{x}}$. This formulation encompass *Inpainting* perturbations: $\boldsymbol{\Phi}(\tilde{\boldsymbol{x}}, \mathbf{M}) = \tilde{\boldsymbol{x}} \odot \mathbf{M} + (\mathbf{1} - \mathbf{M})\mu$ with $\mu \in \mathbb{R}$ a baseline value, and $\odot$ the Hadamard product. This consists in linearly varying the pixel intensities toward a baseline value such as zero [8, 6, 4, 10]. Similarly, *Blurring* consists in applying a blur to certain regions of the image with various intensities [8]. Different perturbation domains can be considered for other types of data such as textual or tabular that we discuss further in the experimental section.

3.2 Sensitivity analysis using Sobol indices

We first briefly review the classical Sobol-Hoeffding decomposition from [28] and introduce the Sobol indices. We consider that $(\boldsymbol{X}_1, ..., \boldsymbol{X}_d)$ are independent variables and assume that $\boldsymbol{f}$ belongs to $\mathbb{L}^2(\mathcal{X}, P)$. Moreover we denote $\mathcal{U} = \{1, ..., d\}$ and $\boldsymbol{u}$ a subset of $\mathcal{U}$, $\sim\!\mathbf{u}$ its complementary onto $\mathcal{U}$ and $\mathbb{E}(\cdot)$ the expectation over the perturbation space. The Hoeffding decomposition allows us to express the function $\boldsymbol{f}$ into summands of increasing dimension, denoting $\boldsymbol{f}_{\boldsymbol{u}}$ the partial contribution of variables $\boldsymbol{X}_{\boldsymbol{u}} = (\boldsymbol{X}_i)_{i \in \boldsymbol{u}}$ to the score $\boldsymbol{f}(\boldsymbol{X})$:

$$\boldsymbol{f}(\boldsymbol{X}) = \boldsymbol{f}_\emptyset + \sum_i \boldsymbol{f}_i(\boldsymbol{X}_i) + \sum_{1 \leqslant i < j \leqslant d} \boldsymbol{f}_{i,j}(\boldsymbol{X}_i, \boldsymbol{X}_j) + \cdots + \boldsymbol{f}_{1,...,d}(\boldsymbol{X}_1, ..., \boldsymbol{X}_d)$$
$$= \sum_{\mathbf{u} \subseteq \mathcal{U}} \boldsymbol{f}_{\boldsymbol{u}}(\boldsymbol{X}_{\boldsymbol{u}}) \tag{1}$$

Eq. 1 consists of 2^d terms and is unique under the following orthogonality constraint.

$$\forall (\boldsymbol{u}, \boldsymbol{v}) \subseteq \mathcal{U}^2 \text{ s.t. } \boldsymbol{u} \neq \boldsymbol{v}, \ \ \mathbb{E}(\boldsymbol{f}_{\boldsymbol{u}}(\boldsymbol{X}_{\boldsymbol{u}})\boldsymbol{f}_{\boldsymbol{v}}(\boldsymbol{X}_{\boldsymbol{v}})) = 0 \tag{2}$$

Furthermore, orthogonality yields the characterization $\boldsymbol{f}_{\boldsymbol{u}}(\boldsymbol{X}) = \mathbb{E}(\boldsymbol{f}(\boldsymbol{X})|\boldsymbol{X}_{\boldsymbol{u}}) - \sum_{\boldsymbol{v} \subset \boldsymbol{u}} \boldsymbol{f}_{\boldsymbol{v}}(\boldsymbol{X})$ and allows us to decompose the model variance as:

$$\mathrm{Var}(\boldsymbol{f}(\boldsymbol{X})) = \sum_i^n \mathrm{Var}(\boldsymbol{f}_i(\boldsymbol{X}_i)) + \sum_{1 \leqslant i < j \leqslant d} \mathrm{Var}(\boldsymbol{f}_{i,j}(\boldsymbol{X}_i, \boldsymbol{X}_j)) + ... + \mathrm{Var}(\boldsymbol{f}_{1,...,d}(\boldsymbol{X}_1, ..., \boldsymbol{X}_d))$$
$$= \sum_{\mathbf{u} \subseteq \mathcal{U}} \mathrm{Var}(\boldsymbol{f}_{\boldsymbol{u}}(\boldsymbol{X}_{\boldsymbol{u}})) \tag{3}$$

Building from Eq. 3, it is natural to characterize the influence of any input subset $\boldsymbol{u}$ as its own variance w.r.t. the total variance. This yields, after normalization by $\text{Var}(\boldsymbol{f}(\boldsymbol{X}))$, the general definition of Sobol indices.

Definition 1 (Sobol indices [19]) *The sensitivity index $\mathcal{S}_{\boldsymbol{u}}$ which measures the contribution of the variable set $\boldsymbol{X}_{\boldsymbol{u}}$ to the model response $\boldsymbol{f}(\boldsymbol{X})$ in terms of fluctuation is given by:*

$$\mathcal{S}_{\boldsymbol{u}} = \frac{\text{Var}(\boldsymbol{f}_{\boldsymbol{u}}(\boldsymbol{X}_{\boldsymbol{u}}))}{\text{Var}(\boldsymbol{f}(\mathbf{X}))} = \frac{\text{Var}(\mathbb{E}(\boldsymbol{f}(\boldsymbol{X})|\boldsymbol{X}_{\boldsymbol{u}})) - \sum_{\boldsymbol{v} \subset \boldsymbol{u}} \mathbb{E}(\boldsymbol{f}_{\boldsymbol{v}}(\boldsymbol{X}))}{\text{Var}(\boldsymbol{f}(\boldsymbol{X}))} \tag{4}$$

Sobol indices give a quantification of the importance of any subset of features with respect to the model decision, in the form of a normalized measure of the model output deviation from $f(\boldsymbol{X})$. Thus, Sobol indices sum to one : $\sum_{\boldsymbol{u} \subseteq \mathcal{U}} \mathcal{S}_{\boldsymbol{u}} = 1$.

For each subset of variables $\boldsymbol{X}_{\boldsymbol{u}}$, the associated Sobol index $\mathcal{S}_{\boldsymbol{u}}$ describes the proportion of the model's output variance explained by this subset. In particular, the first-order Sobol indices $\mathcal{S}_i$ capture the intrinsic share of total variance explained by a particular variable, without taking into account its interactions. Many attribution methods construct such intrinsic importance estimator. However, the framework of Sobol indices enables us to capture higher-order interactions between features. In this view, we define the Total Sobol indices.

Definition 2 (Total Sobol indices [29]) *The total Sobol index $\mathcal{S}_{T_i}$ measuring the contribution to the model output variance of the variable X_i as well as its interactions of any order, with any other input variables, is given by:*

$$\mathcal{S}_{T_i} = \sum_{\substack{\boldsymbol{u} \subseteq \mathcal{U} \\ i \in \boldsymbol{u}}} \mathcal{S}_{\boldsymbol{u}} = 1 - \frac{\text{Var}(\mathbb{E}(\boldsymbol{f}(\boldsymbol{X})|\boldsymbol{X}_{\sim i}))}{\text{Var}[\boldsymbol{f}(\boldsymbol{X})]} = \frac{\mathbb{E}_{\boldsymbol{X}_{\sim i}}(\text{Var}(\boldsymbol{f}(\boldsymbol{X})|\boldsymbol{X}_{\sim i}))}{\text{Var}(\boldsymbol{f}(\boldsymbol{X}))} \tag{5}$$

Where $\mathbb{E}_{\boldsymbol{X}_{\sim i}}(\text{Var}(\boldsymbol{f}(\boldsymbol{X})|\boldsymbol{X}_{\sim i}))$ is the expected variance that would be left if all variables but $\boldsymbol{X}_i$ were to be fixed. (See Supplementary for derivation). $\mathcal{S}_{T_i}$ is the sum of the Sobol indices for the all the possible variable coalitions where i appears, i.e. first and higher order interactions of variable X_i. We note that the following inegality property can easily be deduced: $\forall i, 0 \leq \mathcal{S}_i \leq \mathcal{S}_{T_i} \leq 1$. These indices satisfy several interesting properties which can be derived from 3 and 4 that will make them relevant to the explainability of a black box model:

These statistics enable us to quantify the intrinsic (first-order indices) and relational (total indices) impact of each variable to the model output: A variable with a low total Sobol index is therefore marginally important to explain the model decision, while the difference between the first-order index and the total Sobol index gives the interaction strength of this variable. When $\mathcal{S}_{T_i} \approx \mathcal{S}_i$ then the variable has a weak interaction, and on the contrary, if the difference is strong, it means that the variable has a strong potential for interaction. Thus the following decomposition is an efficient tool to describe a black box model decision. We next present an efficient method to estimate these indices.

3.3 Efficient estimator

As models are becoming more and more complex, the proposed estimator must take into account the computational cost of model evaluation. Efficient estimators have been proposed in the literature [30]. We consider in this work one method proposed by Jansen [31] combined with a Quasi-Monte Carlo sampling strategy. Quasi-Monte Carlo methods improve over Monte Carlo methods by using low-discrepancy sample sequences which allow a better convergence rate [32]. For more information see the appendix B.

We start by drawing two independent matrices of size $N \times d$ of perturbation masks from a Sobol low discrepancy LP_{τ} sequences. Once the perturbation operator is applied to $\tilde{\boldsymbol{x}}$ with these masks, we obtain two matrices $\mathbf{A}$ and $\mathbf{B}$ of the same size of perturbed inputs (e.g. partially masked images). We note $\mathbf{A}_{ji}$ and $\mathbf{B}_{ji}$ the elements of the matrices such that $i = 1, ..., d$ the number of variables studied and $j = 1, ..., N$ the number of samples in each matrix. We form the new matrix $\mathbf{C}^{(i)}$ similar to $\mathbf{A}$ except for the column giving the variable i replaced by the column of $\mathbf{B}$. We denote $f_{\emptyset} = \frac{1}{N} \sum_{j=0}^{N} \boldsymbol{f}(\mathbf{A}_j)$ and the empirical variance $\hat{\mathbf{V}} = \frac{1}{N-1} \sum_{j=0}^{N} (\boldsymbol{f}(\mathbf{A}_j) - f_{\emptyset})^2$. The empirical estimators for first ($\hat{\mathcal{S}_i}$) and total order ($\hat{\mathcal{S}_{T_i}}$) can be formulated as:

$$\hat{\mathcal{S}}_i = \frac{\hat{\mathbf{V}} - \frac{1}{2N} \sum_{j=1}^{N} (\boldsymbol{f}(\mathbf{B}_j) - \boldsymbol{f}(\mathbf{C}_j^{(i)}))^2}{\hat{\mathbf{V}}}$$

$$\hat{\mathcal{S}}_{T_i} = \frac{\frac{1}{2N} \sum_{j=1}^{N} (\boldsymbol{f}(\mathbf{A}_j) - \boldsymbol{f}(\mathbf{C}_j^{(i)}))^2}{\hat{\mathbf{V}}}$$

(6)

Hence, the set of first order and total indices can be computed for $N(d + 2)$ evaluation of the model.

3.3.1 Signed estimator

Although the proposed Sobol' Attribution Method allows us to determine the impact of the variables for a given prediction and thus to locate those on which we are interested, it lacks the ability to highlight the type of contribution made, whether positive or negative. Simple methods such as occlusion comprise this information. In order to take advantage of this property, we propose a variant that combines the importance scores of the total Sobol indices to the sign of the occlusion. We computes the difference in score between the prediction on the original input $\tilde{\boldsymbol{x}}$ and a partial version $\tilde{\boldsymbol{x}}_i$ with variable $\boldsymbol{X}_i$ occluded.

$$\hat{\mathcal{S}}_{T_i}^{\Delta} = \hat{\mathcal{S}}_{T_i} \times \text{sign}(\boldsymbol{f}(\mathbf{x}) - \boldsymbol{f}(\tilde{\boldsymbol{x}}_i))$$

(7)

Intuitively, this provides a direction to the variations generated by the variables studied with respect to a reference state.

4 Experiments

To evaluate the benefits and reliability of the Sobol attribution method, we performed a series of experiments on vision and natural language models using common explainability metrics.

For our vision experiments, we compare the plausibility of the explanations produced on the Pointing Game [33] benchmark. We evaluate the fidelity of our explanations using the deletion metric for 4 representative models commonly used in explainability studies: ResNet50V2 [34] , VGG16 [35], EfficientNet [36] and MobileNetV2 [37] trained on ILSVRC-2012 [38]. In addition, we also compare the speed of convergence of the proposed estimator with that of RISE, a leading approach, on these same models. For our NLP experiments, we fine-tuned a Bert model and trained a bi-LSTM on the IMDB sentiment analysis dataset [39] before comparing fidelity scores using word-deletion for representative methods.

Throughout this work, explanations were generated using the Sobol total estimator $\hat{\mathcal{S}}_{T_i}$ on the target class output. For the experiments involving images, the masks were generated at a resolution of $d' = 11 \times 11$ pixels, then upsampled with a nearest-neighbor interpolation before being applied with the inpainting perturbation function. Finally, N was set to 32 which is equivalent to 3936 forward passes (see section 3.3 for details). For $\hat{\mathcal{S}}_{T_i}^{\Delta}$, an occlusion using the same resolution as the masks is used to sign $\hat{\mathcal{S}}_{T_i}$, with zero as baseline. For RISE, we have followed the recommendations of the original paper with 8000 forward passes on all the models.

4.1 Pointing game

Different evaluation methods have been proposed to compare attribution methods and their explanations [40, 41, 42, 43]. The first common approach consists in measuring the plausibility of an explanation as the correlation between attribution maps and human-provided semantic annotations. Here, we focus on the Pointing Game used in [33, 8, 16, 10]. For a given attribution method, we compute a contribution score for each pixel given a certain class of object appearing in the image. We then calculate the percentage of times the pixel with the highest score is included in the bounding box surrounding the object of interest. Intuitively, a good attribution method should point to the most important evidence of the object appearance in accordance with a human user.

A summary of our results for the Pascal VOC [44] and MS COCO [45] datasets and for VGG16 [35] and ResNet50[34] is provided in Table 1 . In the last column we report the corresponding computation times for each method averaged over 100 MS COCO samples for the ResNet50 model. We subdivided allocation methods into two categories: white-box methods which require access to the backpropagation, such as Gradient [4] and ExtremalPerturbation [16], and/or access to the internal states of the model, such as GradCAM [7] vs. black-box methods such as Occlusion, RISE [10], or

	Method	VOC07 Test (All/Diff)		COCO14 Val (All/Diff)		Time (s)
		VGG16	ResNet50	VGG16	ResNet50	
	Baseline Center	69.6 / 42.4	69.6 / 42.4	27.8 / 19.5	27.8 / 19.5	-
White box	Saliency [3]	76.3 / 56.9	72.3 / 56.8	37.7 / 31.4	35.0 / 29.4	0.031
	DeconvNet [4]	67.5 / 44.2	68.6 / 44.7	30.7 / 23.0	30.0 / 21.9	0.029
	Guided-Backprop. [5]	75.9 / 53.0	77.2 / 59.4	39.1 / 31.4	42.1 / 35.3	0.051
	MWP [33]	77.1 / 56.6	84.4 / 70.8	39.8 / 32.8	49.6 / 43.9	0.039
	cMWP [33]	79.9 / 66.5	**90.7 / 82.1**	49.7 / 44.3	**58.5 / 53.6**	0.040
	GradCAM [7]	86.6 / 74.0	90.4 / 82.3	**54.2 / 49.0**	57.3 / 52.3	0.015
	ExtremalPerturbation [16]	**88.0 / 76.1**	88.9 / 78.7	51.5 / 45.9	56.5 / 51.5	26.48
Black box	RISE [10]	86.9 / 75.1	86.4 / 78.8	50.8 / 45.3	54.7 / 50.0	13.19
	Sobol ($\hat{\mathcal{S}}_{T_i}$) (ours)	87.7 / 75.3	**89.6 / 80.2**	54.6 / 49.6	57.3 / 52.7	6.381
	Sobol signed ($\hat{\mathcal{S}}_{T_i}^{\Delta}$) (ours)	**89.8 / 82.4**	86.7 / 73.0	**55.2 / 49.9**	**57.5 / 53.1**	6.721

Table 1: **Pointing game.** Accuracy over the full test set and a subset of difficult images (defined in [33]). The first and second best results are respectively in **bold** and underlined. Results are based on PyTorch re-implementations using TorchRay package. The reported execution time is an average over 100 runs on ResNet50 using an Nvidia Tesla V100 on Google Colab.

the proposed Sobol method $\mathcal{S}_{T_i}$ which only require the final model prediction. The proposed method outperforms RISE on all of the tested cases, while reducing the number of forward passes by half. Surprisingly, white-box methods do not always lead to better scores, and indeed $\hat{\mathcal{S}}_{T_i}^{\Delta}$ is the leading method for Pascal VOC / VGG16 and our two estimators $\hat{\mathcal{S}}_{T_i}^{\Delta}, \hat{\mathcal{S}}_{T_i}$ prevail on CoCo / VGG16. Also note that the signed version of the estimator obtains better scores. This could be explained by the fact that the datasets tested generally contain several classes and that maximum variance in the output is not always induced by the object of interest. This is consistent with the fact that this does not seem to be the case for the PascalVOC ResNet50, suggesting that it is indeed the object of interest that generally guides the most variance in the output.

4.2 Fidelity

There is broad consensus that measuring the plausibility of an explanation alone is insufficient [46, 47]. Indeed, if an explanation is used to make a critical decision, users expect an explanation to reflect the true underlying decision process of the model and incorrect explanations could have disastrous consequences. One major limitation of evaluation methods based on human ground-truthing such as the pointing game is that they rely on the implicit assumption that models and humans leverage the same image regions for recognition. When humans and models leverage different strategies, a good attribution method should be able to correctly explain a model decision even when the model made an incorrect decision. We thus add an additional complementary type of evaluation that better reflects the fidelity of the explanations. The fidelity metric used assumes that the more faithful an explanation is, the faster the prediction score should drop when pixels that are considered important are reset to a baseline value (e.g., gray values). Note that for images, region removal requires an appropriate baseline state which is still an open problem [48]. Here, we used a Deletion Metric [10] (or $1 - AOPC$ [40]) on 4 different pre-trained models: ResNet50 [34] , VGG16 [35], EfficientNet [36] and MobileNet [37] on 2,000 images sampled from the ImageNet validation set. TensorFlow [49] and the Keras [50] API were used to run the models.

We observe that $\mathcal{S}_{T_i}$ appears to be the most faithful method for black-box methods on all tested models. In general, the white-box methods lead with Integrated Gradients [9] and SmoothGrad [27] as the most faithful white-box methods with one exception: ResNet50V2 where $\mathcal{S}_{T_i}$ obtains the best score among all the methods. In terms of fidelity, compared to the pointing game, the signed version of the Sobol attribution method reaches a lower result than the standard Sobol. This can be explained by the fact that the dataset tested contains only one class and therefore the main variance area generally coincides with the class to be explained. This may suggest that in case of a single class, the signed estimator is to be avoided. Another metric has been proposed by the authors of RISE [10] called Insertion. It consists in starting from a baseline image and inserting the variables on the original image, starting with those considered more important. Nevertheless the score on this

	Method	*ResNet50V2*	*VGG16*	*EfficientNet*	*MobileNetV2*
	Baseline Random (ours)	0.235	0.168	0.124	0.137
White box	Saliency [3]	0.174	0.134	0.105	0.125
	Guided-Backprop. [5]	0.142	0.138	0.105	0.102
	DeconvNet [4]	0.159	0.146	0.105	0.111
	Grad.-Input [51]	0.140	0.096	0.093	0.103
	Integ.-Grad. [9]	**0.123**	**0.095**	**0.091**	**0.093**
	SmoothGrad [27]	0.130	0.106	0.094	0.098
	GradCAM [7]	0.141	0.118	0.130	0.122
Black box	Occlusion [4]	0.350	0.357	0.252	0.357
	RISE [10]	0.127	0.121	0.119	0.114
	Sobol ($\hat{\mathcal{S}}_{T_i}$) (ours)	**0.121**	**0.109**	**0.104**	**0.107**
	Sobol signed ($\hat{\mathcal{S}}_{T_i}^{\Delta}$) (ours)	0.145	0.114	0.147	0.141

Table 2: **Deletion** scores obtained on 2000 ImageNet validation set images. Lower is better. Random consists in removing randomly among the pixels remaining at each step. The first and second best results are respectively in **bold** and underlined.

greedy path is highly influenced by the first inserted pixels which contribute disproportionately to the final score. A good score on this metric therefore requires exploring a region very far from the original image and therefore less relevant. It is possible to modify the distribution of masks in order to explore this region more intensively and thus obtain a better score, the explanation of which is nevertheless less interesting.

4.3 Efficiency

The black-box methods presented so far compete with those that do not require access to the internal representation of the model at the cost of a large number of forwards. This weakness leads us to take a more serious look at the performance of these methods. We believe it is important for current and future methods to not simply increase the number of calls but to make them more efficient We propose an experiment to show that beyond producing higher quality explanations, our estimator converges quickly. We compare the explanation generated at different number of forward to the explanation generated with a forward level of $10,000$. To allow a fair comparison, both methods use masks generated in 7×7 dimensions (as recommended by RISE [10]). For each step, the test is performed on 500 images from the ImageNet validation set.

Figure 2: Spearman rank correlation on $\mathcal{S}_{T_i}$ and RISE explanations generated with different levels of forward on 500 images from the ImageNet dataset. The generated explanations are compared to those generated for 10000 forwards. The more efficient the method, the higher the correlation

This procedure requires defining a measure of similarity between two explanations. Since the proper interpretation method is to rank the features most sensitive to the model's decision, it seems natural to consider the Spearman rank correlation [52] to compare the similarity between explanations. Prior work has provided theoretical and experimental arguments in line with this choice [47, 46, 53, 43]. However, it is important to note that the problem of assessing the similarity between explanations remains an open problem. Empirically, our Sobol attribution method associated with Quasi-Monte Carlo sampling converges faster and is more stable than the RISE estimator. This implies that the number of forwards currently in use can be greatly reduced. Indeed, RISE proposes to use $8,000$ forwards, but our method is faster and reaches better results with 2 times less forwards. Similar results are reported for other models tested (see Appendix E).

4.4 Word deletion

In the context of NLP, the concept of a black box requires to perturb directly individual words and therefore the *Inpainting* perturbation which uses continuous masks is not applicable. We use instead

	Saliency	Grad-Input	SmoothGrad	Integ-Grad	Occlusion	$\hat{\mathcal{S}}_{T_i}$	$\hat{\mathcal{S}}_{T_i}^{\Delta}$
BERT	0.684	0.682	0.682	0.689	**0.531**	0.662	0.598
LSTM	0.541	0.529	0.541	0.538	**0.440**	0.523	0.461

Table 3: **Word deletion** scores, obtained on 1000 sentences. Delete up to 20 words per sentence accordingly to their relevance and track the impact on the classification performance. Lower is better. The first and second best results are respectively in **bold** and underlined.

a perturbation that binarizes the masks used such that $\Phi(\tilde{\boldsymbol{x}}, \mathbf{M}) = \tilde{\boldsymbol{x}} \odot \lceil \mathbf{M} - 0.5 \rceil$, i.e., if the value is greater than 0.5 the word remains present, otherwise it is removed. We then quantitatively validate the fidelity score by ensuring that the explanation identifies words that support a specific decision of the classifier. The models are a bi-LSTM and a BERT model (for details, see appendix G) fine-tuned on the IMDB Review dataset. Similar to the Deletion benchmark, the scores on the original sentence serve as an indicator of relevance. More precisely, from these scores, we successively remove the most relevant words and measure the drop in score when this word is removed. For the experiments, we consider $1,000$ sentences from the validation dataset for which we removed up to 20 words each.

Surprisingly, the different methods tested so far are less efficient on this benchmark than a very simple method: Occlusion (also called Omit-1). One explanation for this phenomenon is that the standard distribution of masks used in our method is centered around 0.5, which consists in removing on average half of the words. However, this perturbation moves us away from the original sentence – preventing us from revealing the important word elements in the context of the initial sentence. We verified this hypothesis by running the same experiment with a different threshold: 0.05 which allowed us to remove far fewer words. As a result, we observed comparable results with Occlusion such as 0.527 for BERT.

5   Conclusion

We have presented a novel explainability method to study and understand the predictions of a black-box model. This approach is grounded theoretically within the Sensitivity Analysis framework using Sobol indices. A non-trivial contribution of this work was to make the approach tractable and efficient for high dimensional data such as images. For this purpose, we have introduced a method using perturbation masks coupled with efficient estimators from the Sensitivity Analysis literature. One additional benefit of the approach is that it provides a way to study the importance of not just first-order interactions between input variables but also higher-order interactions. We showed that our method can be efficiently used on image classifiers. It reaches as good or better results than the best black-box method while being twice faster. It even reaches as good or better results than the best white-box method that requires access to internal states. We also showed that our method could be applied on language models and reported some early competitive results. We hope that this work will guide further efforts in the search for more general and efficient explainability methods for black-box models and that further links will be made with the field of Sensitivity Analysis.

Broader Impact

This work is part of the burgeoning field of explainable AI which has a potential positive impact on society including transportation, science and medicine. There is now broad consensus that many (if not all) the AI systems deployed in real-world settings exhibit significant biases including gender and racial biases. Understanding how these systems arrive at their decisions is a necessary first step before these biases can be corrected. We described a method that provides explanations for predictions made by black-box models that are as faithful as possible (in the sense that it reflects the true inner-working of the model). We thus feel it is important for any explainability method to be built on a rigorous theoretical framework. Here, we borrowed methods from Sensitivity Analysis which has been extensively used to evaluate critical systems and that we showed compare favorably to other approaches on fidelity benchmarks. Nevertheless, progress in explicability should not result in a blind trust in the explanations of the models, which should always be used in the knowledge of their associated flaws. Finally, the attribution methods presented in this work are sensitive to adversarial attacks that can be used to hide the behavior of a model [54, 47] which is still an open problem in the frame of attribution methods.

References

[1] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 2015. 1

[2] Thomas Serre. Deep learning: The good, the bad, and the ugly. *Annual review of vision science*, 2019. 1

[3] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Workshop Proceedings of the International Conference on Learning Representations (ICLR)*, 2014. 1, 3, 4, 7, 8, 16

[4] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, 2014. 1, 2, 3, 4, 6, 7, 8, 16

[5] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. In *Workshop Proceedings of the International Conference on Learning Representations (ICLR)*, 2014. 1, 3, 7, 8, 16

[6] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016. 1, 2, 3, 4

[7] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 1, 3, 4, 6, 7, 8, 16

[8] Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 1, 3, 4, 6

[9] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2017. 1, 3, 4, 7, 8, 16

[10] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018. 1, 2, 3, 4, 6, 7, 8, 16

[11] Andrei Kapishnikov, Tolga Bolukbasi, Fernanda Viégas, and Michael Terry. Xrai: Better attributions through regions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[12] Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. " what is relevant in a text document?": An interpretable machine learning approach. *PloS one*, 12(8):e0181142, 2017. 2

[13] Leila Arras, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. Explaining recurrent neural network predictions in sentiment analysis. 2017. 2

[14] I.M Sobol. Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. *Mathematics and Computers in Simulation*, 2001. 2, 14

[15] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. How to explain individual classification decisions. *The Journal of Machine Learning Research*, 11:1803–1831, 2010. 3

[16] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 3, 4, 6, 7

[17] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[18] RI Cukier, CM Fortuin, Kurt E Shuler, AG Petschek, and JH Schaibly. Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. i theory. *The Journal of chemical physics*, 59(8):3873–3878, 1973. 3

[19] Ilya M Sobol. Sensitivity analysis for non-linear mathematical models. *Mathematical modelling and computational experiment*, 1:407–414, 1993. 3, 5

[20] Andrea Saltelli, Paola Annoni, Ivano Azzini, Francesca Campolongo, Marco Ratto, and Stefano Tarantola. Variance based sensitivity analysis of model output. design and estimator for the total sensitivity index. *Computer physics communications*. 3

[21] Amandine Marrel, Bertrand Iooss, Beatrice Laurent, and Olivier Roustant. Calculations of sobol indices for the gaussian process metamodel. *Reliability Engineering & System Safety*, 2009. 3

[22] Alexandre Janon, Thierry Klein, Agnes Lagnoux, Maëlle Nodet, and Clémentine Prieur. Asymptotic normality and efficiency of two sobol index estimators. *ESAIM: Probability and Statistics*. 3

[23] Art B Owen. Better estimation of small sobol'sensitivity indices. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 23(2):1–17, 2013. 3

[24] Stefano Tarantola, Debora Gatelli, and Thierry Alex Mara. Random balance designs for the estimation of first order global sensitivity indices. *Reliability Engineering & System Safety*, 2006. 3

[25] Sébastien Da Veiga and Fabrice Gamboa. Efficient estimation of sensitivity indices. *Journal of Nonparametric Statistics*, 2013. 3

[26] Jean-Yves Tissot and Clémentine Prieur. Bias correction for the estimation of sensitivity indices based on random balance designs. *Reliability Engineering & System Safety*, 2012. 3

[27] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. In *Workshop on Visualization for Deep Learning, ICML*, 2017. 4, 7, 8, 16

[28] Wassily Hoeffding. A class of statistics with asymptotically normal distribution. In *Annals of Mathematical Statistics*. 1948. 4

[29] Toshimitsu Homma and Andrea Saltelli. Importance measures in global sensitivity analysis of nonlinear models. *Reliability Engineering & System Safety*, 52(1):1–17, 1996. 5

[30] Bertrand Iooss and Paul Lemaître. A review on global sensitivity analysis methods. *Uncertainty management in Simulation-Optimization of Complex Systems: Algorithms and Applications*, 2015. 5

[31] Michiel J.W. Jansen. Analysis of variance designs for model output. *Computer Physics Communications*, 1999. 5

[32] Mathieu Gerber. On integration methods based on scrambled nets of arbitrary size. *Journal of Complexity*, 2015. 5, 14

[33] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. In *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, 2016. 6, 7

[34] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, 2016. 6, 7

[35] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. 2015. 6, 7

[36] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, 2019. 6, 7

[37] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 6, 7

[38] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 6

[39] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. 6

[40] 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. 6, 7

[41] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2019. 6

[42] Umang Bhatt, Adrian Weller, and José M. F. Moura. Evaluating and aggregating feature-based model explanations. 2020. 6

[43] Thomas Fel and David Vigouroux. Representativity and consistency measures for deep neural network explanations. *arXiv preprint arXiv:2009.04521*, 2020. 6, 8

[44] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. Springer, 2010. 6

[45] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*. Springer, 2014. 6

[46] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 7, 8

[47] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2017. 7, 8, 9

[48] Pascal Sturmfels, Scott Lundberg, and Su-In Lee. Visualizing the impact of feature attribution baselines. 2020. 7

[49] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org. 7

[50] François Chollet et al. Keras. https://keras.io, 2015. 7

[51] Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713*, 2016. 8, 15, 16

[52] Charles Spearman. The proof and measurement of association between two things. *American Journal of Psychology*, 1904. 8

[53] Richard Tomsett, Dan Harborne, Supriyo Chakraborty, Prudhvi Gurram, and Alun Preece. Sanity checks for saliency metrics. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2019. 8

[54] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 180–186, 2020. 9

[55] Il'ya Meerovich Sobol'. On the distribution of points in a cube and the approximate evaluation of integrals. *USSR Computational Mathematics and Mathematical Physics*, 1967. 14

[56] Gunther Leobacher and Friedrich Pillichshammer. *Introduction to quasi-Monte Carlo integration and applications*. Springer, 2014. 14

[57] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018. 15

[58] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2017. 15

[59] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 17

[60] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 1997. 16

[61] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014. 16

[62] Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020. 16

[63] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014. 16