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

Thomas Fel, Rémi Cadène, Mathieu Chalvidal, Matthieu Cord, David Vigouroux, Thomas Serre

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

Thomas Fel[1,3,4 *] **Rémi Cadène**[1,2 *†] **Mathieu Chalvidal**[1,3]
Matthieu Cord[2,5] **David Vigouroux**[3,4] **Thomas Serre**[1,3]
[1]Carney Institute for Brain Science, Brown University, USA [2]Sorbonne Université, CNRS, France
[3]Artificial and Natural Intelligence Toulouse Institute, Université de Toulouse, France
[4] Institut de Recherche Technologique Saint-Exupery, France [5]Valeo.ai
{thomas_fel,remi_cadene}@brown.edu

Abstract

We describe a novel attribution method which is grounded in Sensitivity Analysis and uses Sobol indices. Beyond modeling the individual contributions of image regions, Sobol indices provide an efficient way to capture higher-order interactions between image regions and their contributions to a neural network's prediction through the lens of variance. We describe an approach that makes the computation of these indices efficient for high-dimensional problems by using perturbation masks coupled with 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 surpassing the accuracy of state-of-the-art white-box methods which require access to internal representations. Our code is freely available: github.com/fel-thomas/Sobol-Attribution-Method.

1 Introduction

Deep neural networks are now being deployed in numerous domains including medicine, transportation, security or finances with broad societal implications. 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 [1–9]. In addition to helping improve people's trust in these systems, these methods have helped identify and correct biases in datasets [4, 5, 10]. This has, in turn, helped improve these systems' robustness and accelerate their broad deployment. An important limitation of standard explainability methods is that they require access to the system's internal states including hidden layer activations or input gradients [1, 5, 6, 11]. As a result, these so-called white-box methods cannot be applied in the most general situations for which the internal states of network are not publicly accessible. For instance, it is common for companies to use neural networks provided by third parties (e.g., through web APIs or specialized hardware). However, only a handful of so-called black-box methods have been proposed to address this challenge with limited successes [2, 4, 9]. 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 of attributing a score for each input dimension such as image pixels for computer vision systems or individual words for natural language processing. Shown in Fig. 1 is an example image and associated importance map for an image categorization system, whereby scores for individual pixels are displayed as heatmaps where

[*] Equal contribution [†] Work done before April 2021 and joining Tesla

Preprint. Under review.

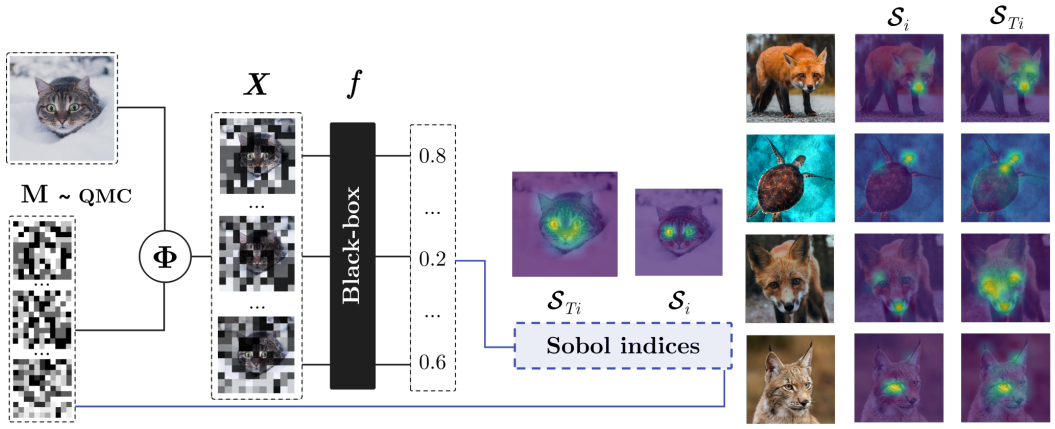

Figure 1: **(Left) Sobol Attribution Method overview.** Our method aims to explain the prediction of a black-box model for a given image. We first sample a set of real-valued masks $\boldsymbol{M}$ drawn from a Quasi-Monte Carlo (QMC) sequence. We apply these masks to the input image through a perturbation function $\boldsymbol{\Phi}$ (here the *Inpainting* function) to form perturbed inputs $\boldsymbol{X}$ that we forward to the black box $\boldsymbol{f}$ to obtain prediction scores. Using the masks $\boldsymbol{M}$ and the associated prediction scores, we finally produce an explanation $\mathcal{S}_{T_i}$ which characterizes the importance of each region by estimating the total order Sobol indices. While $\mathcal{S}_{T_i}$ encompasses the effects of first and all higher-order non-linear interactions between pixel regions, we can also produce the first-order Sobol indices $\mathcal{S}_i$ that reflect the importance of a region in isolation (e.g., the eyes of the cats). **(Right) Sample explanations for ResNet50V2.** Comparing explanations produced with $\mathcal{S}_i$ and $\mathcal{S}_{T_i}$ helps highlight the importance of individual image regions in isolation vs. jointly (e.g., the lynx tips are important but conditioned on the presence of the presence of an eye).

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 network's forward pass. A simple 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 models, pixel intensities are simply set to a default value corresponding to a pure black or gray value; in the case of language models, individual words are removed entirely from the text [12, 13]. However, evaluating the impact of these perturbations one dimension at a time fails to identify all of the non-linear interactions between input variables that are known to prevail in a complex system such as a deep neural network. However, estimating the combined effect of perturbations across multiple locations quickly becomes combinatorially intractable. Methods have been recently proposed to try to address some of these issues by grouping dimensions together, such as by grouping pixels within a neighborhood of the image (superpixel) [2] or sampling perturbation masks that affect multiple regions of the input [4, 9]. A first limitation of these approaches includes the use of Monte Carlo sampling methods which require a high number of forward passes – making these approaches computationally expensive. A second limitation is that they rely on relatively simple perturbations such as flipping pixels on or off [2, 4, 9]. 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]. These methods were initially introduced to help identify the input variables that have the most influence on the variance of the output of a non-linear mathematical system [15]. These were traditionally used in physics and economics to estimate the part of the variance induced by a group of variables on a system's output [16–18]. Our main contribution is a general framework to explain predictions from black-box models by adapting Sobol indices to be used in conjunction with perturbation masks. One of the originalities of our attribution method is that it is designed to support standard perturbations and real-valued intensity perturbations that can generate a continuous range of perturbations. Our second contribution is a tractable method for the calculation of Sobol indices. This is done by first sampling the perturbation masks following Quasi-Monte Carlo sequences, which efficiently covers the space of perturbations. This can be done best by leveraging the most efficient estimator borrowed from the sensitivity analysis literature [19–22]. As a result, the proposed method can be efficiently applied

to high-dimensional inputs such as images. It produces on par with or better explanations than the state-of-the-art with at least half the number of forward passes. Another benefit of our method is that it allows to characterize not only the main effect of image regions but also higher-order interactions by decomposing the variance of the system's prediction using the Sobol indices. We run extensive experiments to demonstrate the benefits of the proposed method on several recent complementary benchmarks including the "Pointing Game" and "Deletion". Our primary focus is image classification, but our method is general and we include results on text classification benchmarks as well.

2 Related work

Attribution methods for white-box models A widely used family of attribution methods uses gradient computations via the backpropagation algorithm. The first such method was introduced in [23] and later refined in [1] 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 pixels. The resulting gradient image indicates which pixels affect the decision score the most. Other gradient-based methods including DeConvNet [2] and Guided Backprop [3] were specifically developed to deal with certain activation functions such as ReLU. However, this family of methods are limited by the fact that they focus on the influence of individual pixels in an infinitesimal neighborhood around the input image in the image space. For instance, it has been shown that gradients often vanish when the prediction score to be explained is near the maximum value [8]. Integrated Gradient [8] partially addresses this issue by accumulating gradients along a straight interpolation path from a baseline state to the original image.

Yet another class of gradient-based methods [6, 24] iteratively optimize a perturbation mask over regions of an image. It leads to a different type of explanation given by the final perturbation mask. Such an approach might leverage more meaningful perturbations but it turns out to be computationally very slow with 20 seconds per explanation – limiting its broad applicability. In comparison, our proposed approach also works on image regions but it does not require iterative gradient computation – and can thus be parallelized.

Another family of attribution methods relies on the neural network's activations. Popular examples include CAM [25] which computes an attribution score based on a weighted sum of feature channels activities – right before the classification layer. GradCAM [5] extends CAM via the use of gradients to reweigh each feature channel to take into account their importance for the predicted class. 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. These methods probe a neural network's responses to perturbations over image regions and combine the resulting predictions into an influence score for each individual pixel or group of pixels. The simplest method, "Occlusion" [2], masks 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. Then the explanation is given by these prediction scores and can be easily interpreted. However, occlusion fails to 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 [4] and RISE [9], addresses this problem by randomly perturbating the input image in multiple regions at a time. Obviously, perturbating 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 [4], takes superpixels as regions to perturbate instead of individual pixels. An influence score is then computed for a set of connected pixel patches indicating how strongly a patch is correlated to the model predictions.

RISE [9] 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 randomly masked versions of the input, RISE [9] produces a importance map by considering the average of the masks weighted by their

associated prediction scores. Instead of using binary masks, our method considers a continuous range of perturbations which allows for a finer exploration of the model's response. Our method can still use the same perturbations as used in Occlusion [2], LIME [4] and RISE [9], but it also enables the use of more advanced perturbation functions that take continuous inputs.

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 multiple orders of influence. The first-order approximates Occlusion [2] by considering the influence of one region at a time, while the second-order considers two regions at a time, etc. 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 70s [26] and reached a cornerstone with the Sobol indices [27]. Sobol indices are currently used in many fields (including those that are said to be safety-critical), especially for the analysis of physical phenomena [17]. More recently, connections have been successfully made between these indices and existing metrics of fairness [28].

They are used to identify the input dimensions that have the highest influence on the output of a model or a mathematical system. Several statistical estimators to compute these indices are available [16, 29, 21, 30, 31] and have asymptotic guarantees [21, 32, 33]. 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 formulate the feature attribution problem as quantifying the contribution of a collection of d real-valued features $\boldsymbol{x} = (x_1, ..., x_d)$ 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 score the probability for the input to belong to a specific class). Our goal is to quantify the importance of each feature to the decision score $\boldsymbol{f}(\boldsymbol{x})$, not just individually but also collectively. To capture these higher-order interactions, our method consists in estimating the Sobol indices of the features $\boldsymbol{x}$ by randomly perturbating them and evaluating the impact of these perturbations on the prediction of the black-box model (Fig. 1).

Considering variations of $\boldsymbol{f}(\boldsymbol{x})$ in response to meaningful perturbations of the input $\boldsymbol{x}$ is a natural way to interpret the local behavior of a the decision function $\boldsymbol{f}$ around $\boldsymbol{x}$. Several methods build on this idea, e.g., by removing one or a group of input features [2, 4, 6, 9, 24] or by back-propagating the gradient to the input space through the model [1, 8, 34, 5]. Most of these methods use the model's internal representations and/or require computing the gradient w.r.t. the 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 the features $\boldsymbol{x} = (x_1, ..., x_d)$ that contribute to $\boldsymbol{f}(\boldsymbol{x})$.

3.1 Random Perturbation

Formally, let us define a probability space $(\Omega, \mathcal{X}, P)$ of possible input perturbations and a random vector $\boldsymbol{X} = (X_1, ..., X_d)$ as a stochastic perturbation of the original input $\boldsymbol{x}$ (see Fig. 1). There are several ways to define random perturbations corresponding to different coverage of the data manifold around $\boldsymbol{x}$. For instance, we can consider the perturbation mask operator $\boldsymbol{\Phi} : \mathcal{X} \times \mathcal{M} \to \mathcal{X}$ which combines a stochastic mask $\boldsymbol{M} = (M_1, ..., M_d) \in \mathcal{M}$ (i.e., an i.i.d sequence of real-valued random variables on $[0, 1]^d$) with the original input $\boldsymbol{x}$. This formulation encompasses *Inpainting* perturbations: $\boldsymbol{\Phi}(\boldsymbol{x}, \boldsymbol{M}) = \boldsymbol{x} \odot \boldsymbol{M} + (\boldsymbol{1} - \boldsymbol{M})\mu$ with $\mu \in \mathbb{R}$ a baseline value, and $\odot$ the Hadamard product. This consists in linearly varying the pixel intensities towards a baseline intensity such as a pure black with a value of zero [6, 4, 2, 9]. Similarly, *Blurring* consists of applying a blur operator with various intensities to certain regions of the image [6]. Different perturbation domains can be considered for other types of data such as textual or tabular data that we discuss further in the experimental section. In the next section, we explain how we adapt the Sobol-based sensitivity analysis using a class of perturbations to explain the predictions of a black-box model.

3.2 Sensitivity analysis using Sobol indices

We first briefly review the classical Sobol-Hoeffding decomposition from [35] and introduce the Sobol indices. Let $(X_1, ..., X_d)$ be independent variables and assume that $\boldsymbol{f}$ belongs to $\mathbb{L}^2(\mathcal{X}, P)$. Moreover we denote the set $\mathcal{U} = \{1, ..., d\}$, $\boldsymbol{u}$ a subset of $\mathcal{U}$, its complementary $\sim\mathbf{u}$ 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 $\boldsymbol{f_u}$ the partial contribution of variables $\boldsymbol{X}_u = (X_i)_{i\in\boldsymbol{u}}$ to the score $\boldsymbol{f}(\boldsymbol{X})$:

$$\begin{aligned}\boldsymbol{f}(\boldsymbol{X}) &= \boldsymbol{f}_\emptyset + \sum_i^d \boldsymbol{f}_i(X_i) + \sum_{1\leqslant i<j\leqslant d} \boldsymbol{f}_{i,j}(X_i, X_j) + \cdots + \boldsymbol{f}_{1,\dots,d}(X_1, ..., X_d) \\ &= \sum_{\mathbf{u}\subseteq\mathcal{U}} \boldsymbol{f_u}(\boldsymbol{X_u})\end{aligned} \tag{1}$$

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

$$\forall(\boldsymbol{u}, \boldsymbol{v}) \subseteq \mathcal{U}^2 \; s.t. \; \boldsymbol{u} \neq \boldsymbol{v}, \;\; \mathbb{E}(\boldsymbol{f_u}(\boldsymbol{X_u})\boldsymbol{f_v}(\boldsymbol{X_v})) = 0 \tag{2}$$

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

$$\begin{aligned}\mathrm{Var}(\boldsymbol{f}(\boldsymbol{X})) &= \sum_i^d \mathrm{Var}(\boldsymbol{f}_i(X_i)) + \sum_{1\leqslant i<j\leqslant d} \mathrm{Var}(\boldsymbol{f}_{i,j}(X_i, X_j)) + ... + \mathrm{Var}(\boldsymbol{f}_{1,\dots,d}(X_1, ..., X_d)) \\ &= \sum_{\mathbf{u}\subseteq\mathcal{U}} \mathrm{Var}(\boldsymbol{f_u}(\boldsymbol{X_u}))\end{aligned} \tag{3}$$

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 $\mathrm{Var}(\boldsymbol{f}(\boldsymbol{X}))$, the general definition of Sobol indices.

Definition 1 (Sobol indices [27]) *The sensitivity index $\mathcal{S}_{\boldsymbol{u}}$ which measures the contribution of the variable set $\boldsymbol{X_u}$ to the model response $\boldsymbol{f}(\boldsymbol{X})$ in terms of fluctuation is given by:*

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

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 $\boldsymbol{f}(\boldsymbol{X})$. Thus, Sobol indices sum to one : $\sum_{\boldsymbol{u}\subseteq\mathcal{U}} \mathcal{S}_{\boldsymbol{u}} = 1$.

For each subset of variables $\boldsymbol{X_u}$, 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 [36]) *The total Sobol index $\mathcal{S}_{T_i}$ which measures the contribution of the variable X_i as well as its interactions of any order with any other input variables to the model output variance is given by:*

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

Where $\mathbb{E}_{\boldsymbol{X}_{\sim i}}(\mathrm{Var}_{X_i}(\boldsymbol{f}(\boldsymbol{X})|\boldsymbol{X}_{\sim i}))$ is the expected variance that would be left if all variables but X_i were to be fixed. $\mathcal{S}_{T_i}$ is the sum of the Sobol indices for the all the possible groups of variables where i appears, i.e. first and higher order interactions of variable X_i.

Since the total interaction index contains the first order index, it is natural that it is greater than or equal to the first order index. We thus note the property which can easily be deduced: $\forall i, 0 \leq \mathcal{S}_i \leq \mathcal{S}_{T_i} \leq 1$. We will now see why these two indices and the difference between them make them relevant for the explainability of a black-box model.

These statistics 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 not important to explain the model decision. Also, a variable has a weak interaction with other variables when $\mathcal{S}_{T_i} \approx \mathcal{S}_i$, while it has a strong interaction when the difference between its two indices is high. A strong interaction means that the effect of one variable on the variation of the model output depends on other variables. Thus, using Sobol indices allows to describe fine grained interactions between inputs which leads to the 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. Many efficient estimators have been proposed in the literature [17]. In this work, we use the Jansen [20] estimator which is often considered as one of the most efficient [22]. Jansen is typically used with a Monte Carlo sampling strategy. We improve over Monte Carlo by using a Quasi-Monte Carlo (QMC) sampling strategy which generates low-discrepancy sample sequences allowing a faster and more stable convergence rate [37]. For more information, see appendix D. We will now describe the procedure to implement these estimators.

We start by drawing two independent matrices of size $N \times d$ of perturbation masks from a Sobol low discrepancy LP_τ sequences. N will be our number of designs and we recall that d is our dimensions (e.g, 121 for 11 by 11 masks). Once the perturbation operator is applied to $\boldsymbol{x}$ with these masks, we obtain two matrices $\mathbf{A}$ and $\mathbf{B}$ of the same size as the perturbed inputs (i.e., 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 $\mathbf{C}^{(i)}$ in the same way as $\mathbf{A}$ except for the fact that the column corresponding to the variable i is now 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 $\hat{\mathrm{V}} = \frac{1}{N-1}\sum_{j=0}^{N}(\boldsymbol{f}(\mathbf{A}_j) - f_\emptyset)^2$. The empirical estimators for first ($\hat{\mathcal{S}}_i$) and total order ($\hat{\mathcal{S}}_{T_i}$) can be formulated as:

$$\hat{\mathcal{S}}_i = \frac{\hat{\mathrm{V}} - \frac{1}{2N}\sum_{j=1}^{N}(\boldsymbol{f}(\mathbf{B}_j) - \boldsymbol{f}(\mathbf{C}_j^{(i)}))^2}{\hat{\mathrm{V}}} \qquad \hat{\mathcal{S}}_{T_i} = \frac{\frac{1}{2N}\sum_{j=1}^{N}(\boldsymbol{f}(\mathbf{A}_j) - \boldsymbol{f}(\mathbf{C}_j^{(i)}))^2}{\hat{\mathrm{V}}} \tag{6}$$

Hence, to compute the set of first order and total indices, it is necessary to perform $N(d+2)$ forwards of the model. We study in section 3.3 how to choose a sufficient number of forwards (N). To ease understanding and demonstrate that these estimators can be easily implemented, we show in Algorithm 1 a minimal pythonic implementation of the total order estimator that outputs $\hat{\mathcal{S}}_{T_i}$ indices. The input `Y` contains the prediction scores of the $N * (d + 2)$ forwards. The scores are ordered following the same QMC sampling ordering of their associated mask. The output `STis` contains d importance scores, one for each dimension of the mask. In the case of images, we obtain our final explanation map by applying a bilinear upsampling to match the dimensions of the input image.

```
def total_order_estimator(Y, N=32, d=11*11):
  fA, fB = Y[:N], Y[N:N*2]
  fC = [Y[N*2+N*i:N*2+N*(i+1)] for i in range(d)]
  f0 = mean(fA)
  V = sum([(val - f0)**2 for val in fA]) / (len(fA) - 1)
  STis = [sum((fA - fC[i])**2) / (2 * N) / V for i in range(d)]
  return STis
```

Algorithm 1: Pythonic implementation of the Total Order indices ($\hat{\mathcal{S}}_{T_i}$) calculation.

3.3.1 Signed estimator

Although the proposed Sobol-based attribution method allows us to determine the impact of any variables for a given prediction and thus to identify diagnostic ones, it lacks the ability to highlight the type of contributions made, whether positive or negative. Simple methods such as "Occlusion" typically include this information. Hence, we propose a variant that combines the importance scores of

		Pointing Game	Deletion	*Time (s)*
	Baseline Center	27.8	0.235	-
White box	Saliency [1]	37.7	0.174	0.031
	Guided-Backprop. [3]	39.1	0.142	0.051
	MWP [38]	39.8	-	0.039
	cMWP [38]	49.7	-	0.040
	Integ.-Grad. [45]	49.7	<u>0.123</u>	0.040
	GradCAM [5]	<u>54.2</u>	0.141	0.015
	ExtremalPerturbation [24]	51.5	-	26.48
Black box	Occlusion	35.6	0.350	1.134
	RISE [9]	50.8	0.127	13.19
	Sobol ($\hat{\mathcal{S}}_{T_i}$) (ours)	**54.6**	**0.121**	6.381

Table 1: **Pointing game.** Accuracy over the full test set and a subset of difficult images (defined in [38]). The first and second best results are **bolded** and <u>underlined</u>. Results are based on PyTorch re-implementations using the TorchRay package. The reported execution time is an average over 100 runs on ResNet50 using an Nvidia Tesla P100 on Google Colab and a batch size of 64. Lower execution time can be reached with higher batch size.

the total Sobol indices with the sign of the occlusion. We compute the difference in score between the prediction on the original input $\boldsymbol{x}$ and a partial version $\boldsymbol{x}_{\setminus i}$ with the variable x_i occluded. Intuitively, this provides an estimate of the direction of the variations generated by the variables studied with respect to a reference state.

$$\hat{\mathcal{S}}^{\Delta}_{T_i} = \hat{\mathcal{S}}_{T_i} \times \text{sign}(\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x}_{\setminus i})) \tag{7}$$

4 Experiments

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

For our vision experiments, we compared the plausibility of the explanations produced on the Pointing Game [38] benchmark. We evaluate the fidelity of our explanations using the Deletion metric for 4 representative models commonly used in explainability studies: ResNet50V2 [39] , VGG16 [40], EfficientNet [41] and MobileNetV2 [42] trained on ILSVRC-2012 [43]. In addition, we also compared the speed of convergence of the proposed estimator with that of the leading approach, RISE [9], on the same models. For our NLP experiments, we fine-tuned a Bert model and trained a bi-LSTM on the IMDB sentiment analysis dataset [44] 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. In the supplementary material, we demonstrate the effectiveness of modeling higher-order interactions between image regions by comparing $\hat{\mathcal{S}}_{T_i}$ against $\hat{\mathcal{S}}_i$ which only models the main effects. 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 method before being applied with the *Inpainting* perturbation function. Finally, N was set to 32 which is equivalent to $3,936$ forward passes (see Section 3.3 for details). For $\hat{\mathcal{S}}^{\Delta}_{T_i}$, an occlusion using the same resolution as the masks was used to sign $\hat{\mathcal{S}}_{T_i}$, with zero as baseline. For RISE [9], we have followed the recommendations of the original paper with $8,000$ forward passes for all models.

4.1 Pointing game

Different evaluation methods have been proposed to compare attribution methods and their explanations [46–49]. 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 focused on the Pointing Game used in [38, 6, 24, 9]. For each attribution method, we compute a contribution score for each pixel of a given class of objects, e.g., bike or car. We then calculated the percentage of times the pixel with the highest score is included in the bounding box surrounding the

object of interest. In this benchmark, a good attribution method should point to the most important evidence of the object appearance in accordance with a human user.

In Table 1, a report results for the Pascal VOC [50] and MS COCO [51] datasets using VGG16 [40] and ResNet50[39]. In the last column we report the computation times for each method averaged over 100 MS COCO samples for the ResNet50 model. We subdivided explanation methods into two categories: white-box methods which require the use of backpropagation, such as Gradient [2] and Extremal Perturbation [24], and/or access to the internal states of the model, such as GradCAM [5] versus black-box methods such as Occlusion [2], RISE [9], or the proposed Sobol method $\mathcal{S}_{T_i}$ which only require the final model predictions. The proposed method outperforms RISE [9] on all of the tested cases, while reducing the number of forward passes by half. Surprisingly, white-box methods do not always lead to higher scores, and indeed $\hat{\mathcal{S}}^{\Delta}_{T_i}$ is the leading method for Pascal VOC / VGG16 and our two estimators $\hat{\mathcal{S}}^{\Delta}_{T_i}, \hat{\mathcal{S}}_{T_i}$ prevail on COCO / VGG16. Also note that our signed version of the estimator obtains higher scores overall. This might be due to the fact that images from VOC and COCO often feature several types of objects. Thus, the maximum variance in the output is not always induced by the object of interest but can be due to the masking of another object in the image. This result suggests that our signed version $\hat{\mathcal{S}}^{\Delta}_{T_i}$ should be used on multi-label datasets, while $\hat{\mathcal{S}}_{T_i}$ should be used on multi-class datasets. We indeed confirm this in the next set of experiments on a multi-class dataset.

4.2 Fidelity

There is a broad consensus that measuring the plausibility of an explanation alone is insufficient [52, 53]. 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 not just a consensus with humans. Failures to do so could have disastrous consequences. A first major limitation of current evaluation methods based on human-provided groundtruth such as the pointing game is that they do not work when a model prediction is wrong. In this case, an explanation method can be penalized for not pointing to the correct evidence even though explaining prediction errors is a critical use case for explanation methods. Another limitation of these evaluation methods is that they make the implicit assumption that the models should be relying on the same image regions than humans for recognition [54–56], which is likely to be an incorrect assumption. We thus use the fidelity metric as a complementary type of evaluation. This metric assumes that the more faithful an explanation is, the quicker the prediction score should drop when pixels that are considered important are reset to a baseline value (e.g., gray values).

In Table 2, we report results for the Deletion Metric [9] (or $1 - AOPC$ [46]) for 4 different pre-trained models: ResNet50 [39] , VGG16 [40], EfficientNet [41] and MobileNet [42] on 2,000 images sampled from the ImageNet validation set. TensorFlow [57] and the Keras [58] API were used to run the models. Several baseline values can be used [59], but we chose the standard approach with gray values. We observe that the proposed Sobol $\hat{\mathcal{S}}T_i$ is the most faithful black-box methods with the lowest deletion scores across all models. Overall $\hat{\mathcal{S}}_{T_i}$ is able to match the scores of the most faithful white-box method, namely Integrated Gradients [8], and gets the lowest score on ResNet50V2 with 0.121 against 0.123 (lower is better). We also report that our signed version $\hat{\mathcal{S}}^{\Delta}_{T_i}$ is less faithful that the standard Sobol $\hat{\mathcal{S}}T_i$. This can be explained by the fact that ImageNet images contains only one object and therefore the main variance area generally coincides with the class to be explained. This confirms our observation on the previous pointing game benchmark that $\hat{\mathcal{S}}T_i$ should be preferred in a multi-class setup and $\hat{\mathcal{S}}^{\Delta}_{T_i}$ in a multi-label setup.

Another metric called Insertion has been proposed by the authors of RISE [9]. Instead of deleting pixels in the original image like with Deletion, Insertion consists in adding pixels on a baseline image, e.g. one gray image, starting with pixels that are associated with the highest importance scores for a given explanation method. An issue with Insertion is that the score computed along the insertion path is highly influenced by the first inserted pixels which contributes disproportionately. A good score on this metric therefore requires exploring a region very far from the original image and closer to the baseline. For this reason, we rather preferred to focus our study on Deletion than Insertion. However, we also report results on Insertion in the supplementary material using the same hyperparameters as used in Deletion.

	Method	*ResNet50V2*	*VGG16*	*EfficientNet*	*MobileNetV2*
	Baseline Random (ours)	0.235	0.168	0.124	0.137
White box	Saliency [1]	0.174	0.134	0.105	0.125
	Guided-Backprop. [3]	0.142	0.138	0.105	0.102
	DeconvNet [2]	0.159	0.146	0.105	0.111
	Grad.-Input [45]	0.140	<u>0.096</u>	<u>0.093</u>	0.103
	Integ.-Grad. [8]	**0.123**	**0.095**	**0.091**	**0.093**
	SmoothGrad [34]	<u>0.130</u>	0.106	0.094	<u>0.098</u>
	GradCAM [5]	0.141	0.118	0.130	0.122
Black box	Occlusion [2]	0.350	0.357	0.252	0.357
	RISE [9]	<u>0.127</u>	0.121	<u>0.119</u>	<u>0.114</u>
	Sobol ($\hat{\mathcal{S}}_{T_i}$) (ours)	**0.121**	**0.109**	**0.104**	**0.107**
	Sobol signed ($\hat{\mathcal{S}}^{\Delta}_{T_i}$) (ours)	0.145	<u>0.114</u>	0.147	0.141

Table 2: **Deletion** scores obtained on 2,000 ImageNet validation set images. Lower is better. Random consists in removing pixels at each step at random. The first and second best results are **bolded** and <u>underlined</u>.

4.3 Efficiency

The black-box methods presented so far compete with white-box methods that do not require access to the internal representation of the model at the cost of a large number of forward passes, e.g., around $8,000$ for RISE [9]. This weakness leads us to take a more serious look at the performance of the proposed method. It seems critical for the deployment of black-box methods to lower the amount of compute required to produce correct explanations. We describe an experiment to show that beyond producing higher quality explanations, our estimator converges quickly. We first generate an explanation with a high number of forward passes that is large enough to reach convergence, e.g. $10,000$ forward passes. Then we compare this explanation that "converged" to other explanations obtained with lower numbers of forward passes. It allows us to measure the stability and rate of convergence towards this explanation that "converged", but more practically to find the proper trade-off between the amount of compute and the quality of explanations. 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 [60] to compare the similarity between explanations. Prior work has provided theoretical and experimental arguments in line with this choice [53, 52, 61, 49]. In Fig. 2, we compare the proposed Sobol attribution method $\hat{\mathcal{S}}_{T_i}$ against RISE [9], which is the current state-of-the-art for black-box methods. We use their respective gold explanation generated after $10,000$ forwards. To allow a fair comparison, both methods use masks generated in 7×7 dimensions (as recommended by RISE [9]). We report the average results and variance over 500 images from the ImageNet validation set using EfficientNet, a convolutional neural network optimized for fast forward computing times. We observe that our method exhibit higher convergence rate by getting higher Spearman's rank correlation of 0.8 after only 1,000 forwards against 0.65 for RISE, and consistently obtain higher scores until reaching 0.97 with 7,000 forwards against 0.73 for RISE. Additionally, we observe that Sobol has a more stable convergence by getting an overall lower variance than RISE. This implies that the number of forward passes used in Sobol can be greatly reduced to accommodate computational resources constraints compared to RISE. Indeed, RISE pro-

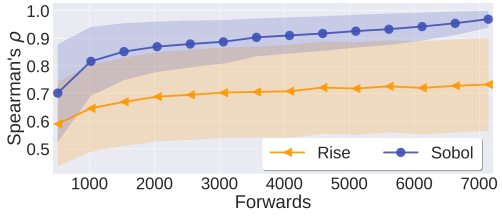

Figure 2: Comparison between the rate and stability of convergence of Sobol $\hat{\mathcal{S}}_{T_i}$ and RISE [9]. A high Spearman correlation rank corresponds to producing an explanation that is similar to the explanation that "converged", i.e. with $10,000$ forward passes. We report mean and variance computed over 500 images from the ImageNet dataset using EfficientNet.

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

Table 3: **Word deletion** scores, obtained on 1,000 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 **bolded** and <u>underlined</u>.

poses to use $8,000$ forwards, but our method is faster and reaches better results with half the number of passes. We report similar results for other neural network architectures in the supplementary material D. Finally, we perform an ablation study of Sobol to show the impact of lowering the number of forwards on the Deletion benchmark. We report competitive scores with 16 times fewer number of forwards than RISE by reaching 0.151 in Deletion score with 492 forwards. For reference, Sobol was reaching state-of-the-art results of 0.121 with 3,936 forwards. We report additional scores in the supplementary materials S2.

4.4 Word deletion

For NLP, black-box methods require the use of perturbations that can be applied to the space of characters, words or sentences. For instance, a common perturbation consists in simply removing one word of the sentence to be explained. Therefore, the *Inpainting* perturbation that we used with Sobol to reduce the intensity of pixels in a continuous manner cannot be directly applied in this context. Instead, we adapt it by binarizing the masks such that $\mathbf{\Phi}(\boldsymbol{x}, \mathbf{M}) = \boldsymbol{x} \odot \lceil \mathbf{M} - 0.5 \rceil$, i.e., if the value is greater than 0.5 the word is kept, otherwise it is removed. We then verify that our Sobol method can be used to identify words that support a specific decision of a text classifier. Inspired by previous work [12, 13, 62], we introduce an experimental benchmark on the IMDB Review dataset [63]. It is similar to the previous Deletion benchmark for images in that it focuses on assessing the faithfulness of the explanation and does not require specific human annotations. More precisely, we first trained two models: a bi-LSTM from scratch and a BERT model fine-tuned for the task (see appendix H for details). We generate explanations on $1,000$ sentences from the validation dataset. An explanation associates an importance score to each word. Similar to Deletion, we use these scores to successively remove the most relevant word of the sentence and measure the corresponding drop in the prediction score of the model.

In Table 3, we report results for explanation methods that are commonly used in NLP. Both our two Sobol methods have a better faithfulness than all the tested gradient-based white-box methods including Saliency, Grad-Input, SmoothGrad, and Integr-Grad. For instance, Sobol $\hat{\mathcal{S}}^{\Delta}_{T_i}$ even reaches a low Word deletion scores of 0.461 (lower is better) for the bi-LSTM compared to 0.529 for the best white-box approach. However, the proposed methods are only the second and third most faithful methods. Occlusion (often called Omit-1 in NLP) reaches the lowest score of 0.440 for LSTM and 0.531 for BERT, against 0.461 and 0.598 for Sobol $\hat{\mathcal{S}}^{\Delta}_{T_i}$. This is due to the fact that the default distribution of masks used in Sobol is centered around 0.5, which corresponds to removing on average half of the words as opposed to a single word for Occlusion. In IMDB, this causes the frequent removal of critical words that support the model decision. Indeed, we report comparable results with Occlusion (e.g., 0.527 for BERT) for a lower threshold of 0.05 to remove far fewer words. Since Sobol can model higher-order interactions between words, we believe that it could successfully be used for NLP tasks that are more complex than sentiment classification on IMDB.

5 Conclusion

We have presented a novel explainability method to study and understand the predictions of a black-box model. This approach is grounded within the theoretical framework of sensitivity analysis 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 the main effects of input variables but also higher-order interactions between them. We showed that our method can be efficiently used to explain the decisions of image classifiers. It reaches performance on par with or better than the current best black-box methods while being twice as fast. It even reaches

comparable results to the best white-box methods without requiring access to internal states. We also showed that our method could be applied to language models and reported initial competitive results. It is our 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.

Acknowledgments

This work was conducted as part the DEEL* project. It was funded by the Artificial and Natural Intelligence Toulouse Institute (ANITI) grant #ANR19-PI3A-0004. MC was funded by ANR grant VISADEEP (ANR-20-CHIA-0022). TS was funded by ONR (N00014-19-1-2029) and NSF (IIS-1912280). The computing hardware was supported in part by NIH Office of the Director grant #S10OD025181 via the Center for Computation and Visualization.

References

[1] 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, 9, 19

[2] 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, 8, 9, 15, 19

[3] 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. 3, 7, 9, 19

[4] 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

[5] 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, 7, 8, 9, 19

[6] 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, 7

[7] François Bachoc, Fabrice Gamboa, Jean-Michel Loubes, and Laurent Risser. Entropic variable boosting for explainability & interpretability in machine learning. 2018.

[8] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2017. 3, 4, 8, 9, 19

[9] 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, 7, 8, 9, 15, 16, 19

[10] Corentin Dancette, Remi Cadene, Damien Teney, and Matthieu Cord. Beyond question-based biases: Assessing multimodal shortcut learning in visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021. 1

[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, 10

[13] Leila Arras, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. Explaining recurrent neural network predictions in sentiment analysis. In *Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA) in ENMLP*, 2017. 2, 10

[14] I.M Sobol. Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. *Mathematics and Computers in Simulation*, 2001. 2, 15

[15] Andrea Saltelli. Making best use of model evaluations to compute sensitivity indices. *Computer physics communications*, 2002. 2

*https://www.deel.ai/

[16] 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*, 2010. 2, 4

[17] Bertrand Iooss and Paul Lemaître. A review on global sensitivity analysis methods. *Uncertainty management in Simulation-Optimization of Complex Systems: Algorithms and Applications*, 2015. 4, 6

[18] Thorsten Wagener and Francesca Pianosi. What has global sensitivity analysis ever done for us? a systematic review to support scientific advancement and to inform policy-making in earth system modelling. *Earth-science reviews*, 2019. 2

[19] Michiel JW Jansen, Walter AH Rossing, and Richard A Daamen. Monte carlo estimation of uncertainty contributions from several independent multivariate sources. In *Predictability and nonlinear modelling in natural sciences and economics*. Springer, 1994. 2

[20] Michiel J.W. Jansen. Analysis of variance designs for model output. *Computer Physics Communications*, 1999. 6

[21] Alexandre Janon, Thierry Klein, Agnes Lagnoux, Maëlle Nodet, and Clémentine Prieur. Asymptotic normality and efficiency of two sobol index estimators. *ESAIM: Probability and Statistics*, 2014. 4

[22] Arnald Puy, William Becker, Samuele Lo Piano, and Andrea Saltelli. A comprehensive comparison of total-order estimators for global sensitivity analysis. *International Journal for Uncertainty Quantification*, 2020. 2, 6

[23] 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

[24] 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, 7, 8

[25] 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

[26] 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. 4

[27] Ilya M Sobol. Sensitivity analysis for non-linear mathematical models. *Mathematical modelling and computational experiment*, 1:407–414, 1993. 4, 5

[28] Clément Bénesse, Fabrice Gamboa, Jean-Michel Loubes, and Thibaut Boissin. Fairness seen as global sensitivity analysis. *arXiv preprint arXiv:2103.04613*, 2021. 4

[29] Amandine Marrel, Bertrand Iooss, Beatrice Laurent, and Olivier Roustant. Calculations of sobol indices for the gaussian process metamodel. *Reliability Engineering & System Safety*, 2009. 4

[30] Art B Owen. Better estimation of small sobol'sensitivity indices. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 23(2):1–17, 2013. 4

[31] 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. 4

[32] Sébastien Da Veiga and Fabrice Gamboa. Efficient estimation of sensitivity indices. *Journal of Nonparametric Statistics*, 2013. 4

[33] Jean-Yves Tissot and Clémentine Prieur. Bias correction for the estimation of sensitivity indices based on random balance designs. *Reliability Engineering & System Safety*, 2012. 4

[34] 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, 9, 19

[35] Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics*, 1948. 5

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

[37] Mathieu Gerber. On integration methods based on scrambled nets of arbitrary size. *Journal of Complexity*, 2015. 6, 16

[38] 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. 7

[39] 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. 7, 8

[40] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015. 7, 8

[41] 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. 7, 8

[42] 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. 7, 8

[43] 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. 7

[44] 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*, 2011. 7

[45] 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. 7, 9, 17, 19

[46] 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*, 2016. 7, 8

[47] 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.

[48] Umang Bhatt, Adrian Weller, and José M. F. Moura. Evaluating and aggregating feature-based model explanations. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2020.

[49] Thomas Fel and David Vigouroux. Representativity and consistency measures for deep neural network explanations. *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2022. 7, 9

[50] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision (IJCV)*, 2010. 8

[51] 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. 8

[52] 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. 8, 9, 15, 19

[53] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2017. 8, 9, 15

[54] Shimon Ullman, Liav Assif, Ethan Fetaya, and Daniel Harari. Atoms of recognition in human and computer vision. *Proceedings of the National Academy of Sciences*, 2016. 8

[55] Drew Linsley, Sven Eberhardt, Tarun Sharma, Pankaj Gupta, and Thomas Serre. What are the visual features underlying human versus machine vision? In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017.

[56] Drew Linsley, Dan Shiebler, Sven Eberhardt, and Thomas Serre. Learning what and where to attend. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018. 8

[57] 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. 8

[58] François Chollet et al. Keras. `https://keras.io`, 2015. 8

[59] Pascal Sturmfels, Scott Lundberg, and Su-In Lee. Visualizing the impact of feature attribution baselines. *Distill*, 2020. 8

[60] Charles Spearman. The proof and measurement of association between two things. *American Journal of Psychology*, 1904. 9

[61] 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. 9

[62] 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*, 10(7):e0130140, 2015. 10

[63] 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*, 2011. 10

[64] 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. 15

[65] 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. 16

[66] Gunther Leobacher and Friedrich Pillichshammer. *Introduction to quasi-Monte Carlo integration and applications*. Springer, 2014. 16

[67] 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. 17

[68] 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. 17

[69] Matthew Sotoudeh and Aditya V. Thakur. Computing linear restrictions of neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2019. 18

[70] 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. 19, 20

[71] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 1997. 20

[72] 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. 20

[73] 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. 20

[74] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014. 20

A 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 [64, 53] which is still an open problem in the frame of attribution methods.

B Qualitative comparison

Regarding the visual consistency of our method, Fig. S1 shows a side-by-side comparison between our method and the other methods tested in the Fidelity benchmark. The images are not hand-picked but are the first images from the ImageNet validation set. To allow better visualization, the gradient-based methods were 2 percentile clipped. The only black box methods are Occlusion, Rise and $\mathcal{S}_{T_i}$. We found that $\mathcal{S}_{T_i}$ consistently provides a sparser map than RISE [9] while being equally consistent. On the other hand, we found that in general, the gradient-based method provides the sharpest map, but some are prone to failure (fourth row in the Fig. S1), which is a known problem [52].

C Effectiveness of modeling higher-order interactions

We introduced two approaches, Sobol ($\hat{\mathcal{S}}_{T_i}$) and Sobol signed ($\hat{\mathcal{S}}^{\Delta}_{T_i}$), that combine effects of first- and all higher-orders interactions between image regions. For comparison, Occlusion [2] only accounts for the first order as it removes one region at a time, while RISE [9] accounts for higher-order by removing around 50% of regions at a time. As seen in Table S1, RISE already surpasses Occlusion on ImageNet in term of Deletion scores, which may indicate that using higher-order information is effective.

To further demonstrates that it is critical to model the higher orders, we evaluate Sobol first-order ($\mathcal{S}_i$) on our Deletion benchmark. We report that Sobol ($\mathcal{S}_{T_i}$) reaches lower deletions scores (lower is better) than Sobol first-order ($\mathcal{S}_i$) with 0.121 against 0.170 respectively on ResNet50v2, and similar differences on VGG16, EfficientNet and MobileNetV2.

Method	*ResNet50V2*	*VGG16*	*EfficientNet*	*MobileNetV2*
Sobol first-order ($\hat{\mathcal{S}}_i$)	0.170	0.147	0.129	0.143
Sobol ($\hat{\mathcal{S}}_{T_i}$)	**0.121**	**0.109**	**0.104**	**0.107**

Table S1: **Deletion** scores obtained on 2,000 ImageNet validation set images. Lower is better.

D Efficiency of Sobol estimator

Regarding the estimation of the Sobol indices, we notice that we can derive a 'brute-force' (or often called double-loop method [14]) estimator from the definition 1:

$$\mathcal{S}_i = \frac{\int \left[\int \boldsymbol{f}(\boldsymbol{X}) d\boldsymbol{X}_{\sim i}\right]^2 d\boldsymbol{X}_i - (\int \boldsymbol{f}(\boldsymbol{X}) d\boldsymbol{X})^2}{\int \boldsymbol{f}(\boldsymbol{X})^2 d\boldsymbol{X} - (\int \boldsymbol{f}(\boldsymbol{X}) d\boldsymbol{X})^2} \tag{8}$$

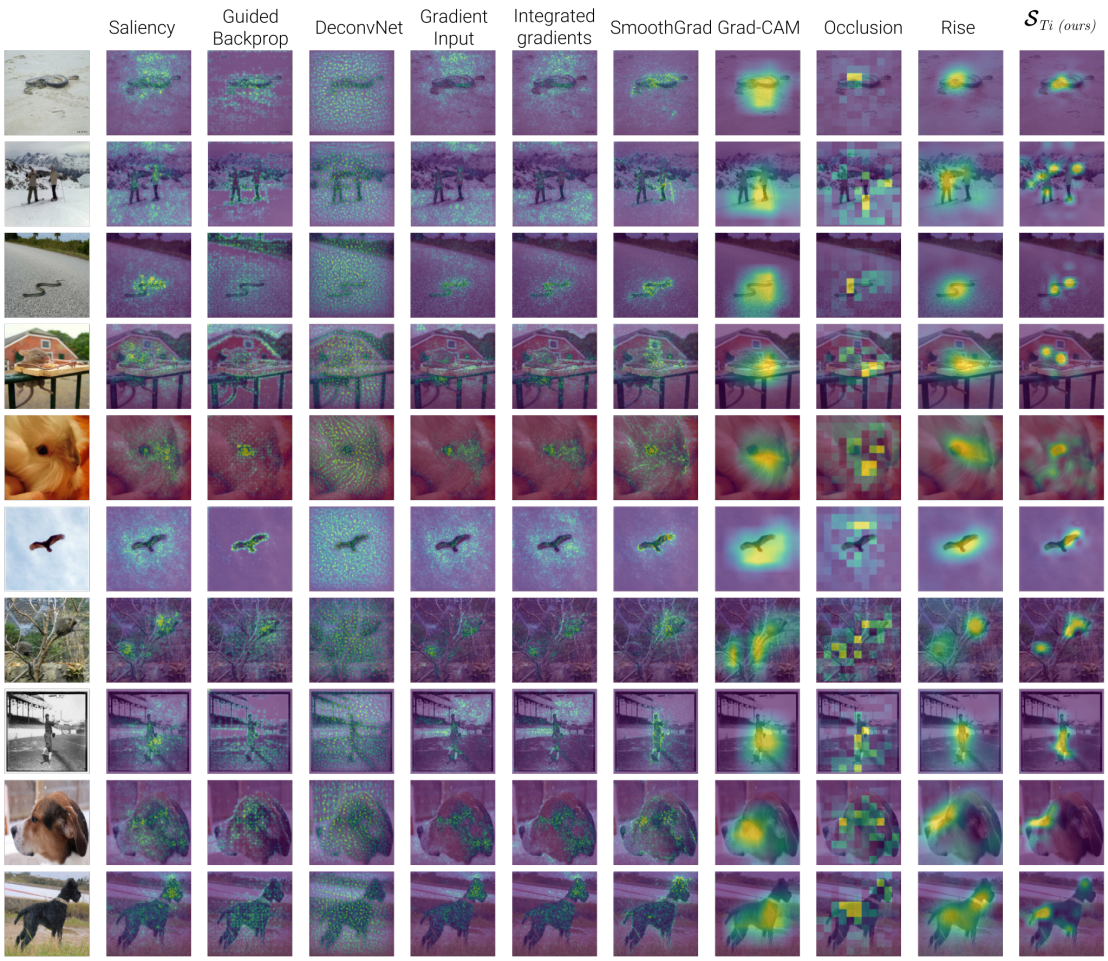

Figure S1: **Qualitative comparison** with other explainability methods. The heatmaps are normalized and clipped at 2 percentile for Saliency, Guided-Backprop, DeconvNet, Smoothgrad and Integrated-Gradients. Explanations are generated from a ResNet50V2.

However, one the main problems with this estimator is the cost of computation, which can be too heavy, especially with complex models such as large neural networks. This difficulty is particularly true for the calculation of total Sobol indices.

Since the perturbation masks are used to approximate these integrals, an efficient way to proceed is to generate those masks from a low discrepancy sequences, also called Quasi-random sequences. These sequences allow to efficiently integrate functions on the hypercube $[0, 1]^d$. In fact, they have a faster convergence rate compared to ordinary Monte Carlo methods [37] (with $\boldsymbol{f}$ sufficiently regular). This difference being due to the use of a deterministic sequence that covers $[0, 1]^d$ more uniformly. In our experiments we used Sobol sequences [65], we refer the readers to [66] for more informations. The efficiency of the estimator and the sampling is shown on Figures S2, S3 and S4 where our estimator consistently converges faster than RISE [9].

We also perform an ablation study of the number of forwards on the Deletion benchmark. In Table S2, we show that competitive scores can be obtained with lower number of forwards such as 0.151 in Deletion score with 492 forwards instead of 0.121 with 3936 forwards which is our default number of forwards.

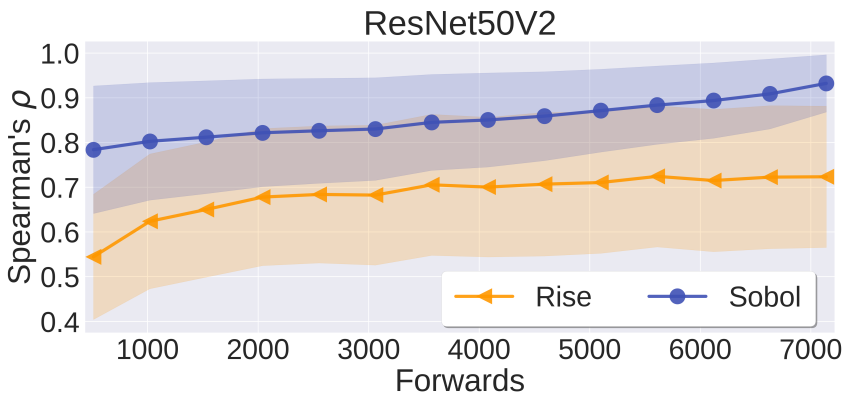

Figure S2: Spearman rank correlation of explanations as a function of the number of forwards, compared to an explanation generated with $10,000$ forwards. The model used is a ResNet50V2.

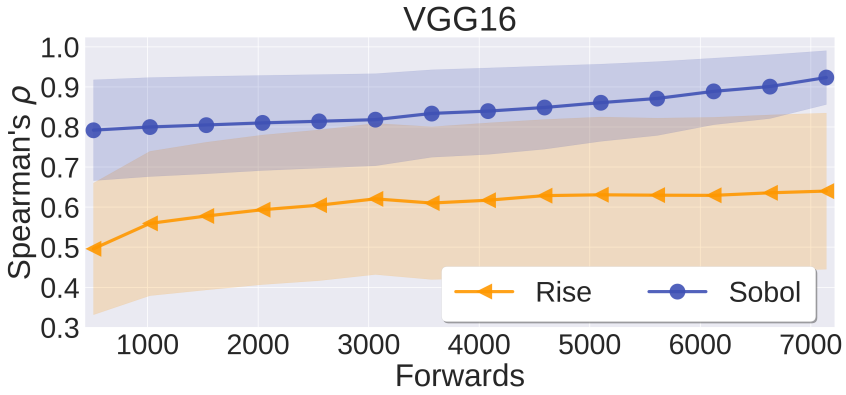

Figure S3: Spearman rank correlation of explanations as a function of the number of forwards, compared to an explanation generated with $1,0000$ forwards. The model used is a VGG16.

E Explanation methods

In the following section, the formulation of the different methods used in the experiment is given. We define $\boldsymbol{f}(\boldsymbol{x})$ the logit score (before softmax) for the class of interest. An explanation method provides an attribution score for each input variables. Each value then corresponds to the importance of this feature for the model results.

Saliency is a visualization techniques based on the gradient of a class score relative to the input, indicating in an infinitesimal neighborhood, which pixels must be modified to most affect the score of the class of interest.

$$g^{SA}(\boldsymbol{x}) = ||\nabla_{\boldsymbol{x}} \boldsymbol{f}(\boldsymbol{x})||$$

Gradient $\odot$ Input [45] is based on the gradient of a class score relative to the input, element-wise with the input, it was introduced to improve the sharpness of the attribution maps. A theoretical analysis conducted by [67] showed that Gradient $\odot$ Input is equivalent to ϵ-LRP and DeepLIFT [68] methods under certain conditions: using a baseline of zero, and with all biases to zero.

$$g^{GI}(\boldsymbol{x}) = \boldsymbol{x} \odot ||\nabla_{\boldsymbol{x}} \boldsymbol{f}(\boldsymbol{x})||$$

Integrated Gradients consists of summing the gradient values along the path from a baseline state to the current value. The baseline is defined by the user and often chosen to be zero. This integral can be approximated with a set of m points at regular intervals between the baseline and the point

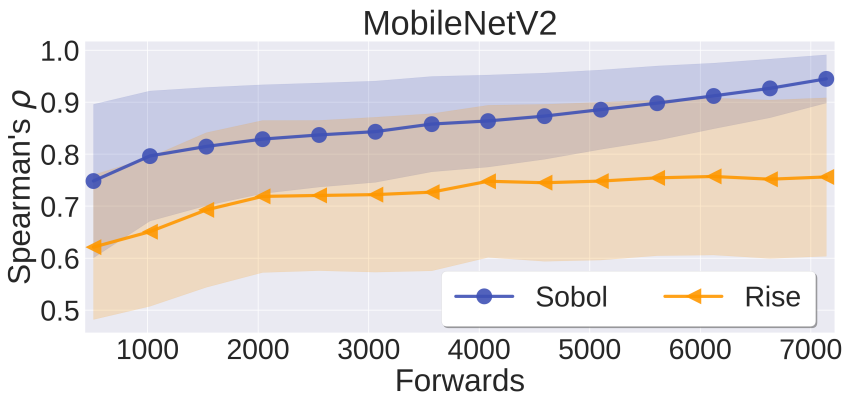

Figure S4: Spearman rank correlation of explanations as a function of the number of forwards, compared to an explanation generated with $1,0000$ forwards. The model used is a MobileNetV2.

Number of samples	Deletion scores
492	0.151
984	0.140
1476	0.132
1968	0.123
2460	0.121
2952	0.120
3444	0.120
3936	0.121

Table S2: **Deletion** scores averaged over 2,000 images of ImageNet validation set using ResNet50V2 and Sobol ($\hat{\mathcal{S}}_{T_i}$). Lower is better.

of interest. In order to approximate from a finite number of steps, we use a Trapezoidal rule and not a left-Riemann summation, which allows for more accurate results and improved performance (see [69] for a comparison). The final result depends on both the choice of the baseline $\mathbf{x}_0$ and the number of points to estimate the integral. In the context of these experiments, we use zero as the baseline and $m = 80$.

$$g^{IG}(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{x}_0) \int_0^1 \nabla_{\boldsymbol{x}} \boldsymbol{f}(\boldsymbol{x}_0 + \alpha(\boldsymbol{x} - \boldsymbol{x}_0)))d\alpha$$

SmoothGrad is also a gradient-based explanation method, which, as the name suggests, averages the gradient at several points corresponding to small perturbations (drawn i.i.d from a normal distribution of standard deviation σ) around the point of interest. The smoothing effect induced by the average help reducing the visual noise, and hence improve the explanations. In practice, Smoothgrad is obtained by averaging after sampling m points. In the context of these experiments, we took $m = 80$ and $\sigma = 0.2$ as suggested in the original paper.

$$g^{SG}(\boldsymbol{x}) = \mathbb{E}_{\varepsilon \sim \mathcal{N}(0, \boldsymbol{I}\sigma)}(\nabla_{\boldsymbol{x}} \boldsymbol{f}(\boldsymbol{x} + \varepsilon))$$

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

$$g^{GC} = max(0, \sum_k \alpha_k^c \mathbf{A}^k)$$

Notice that the size of the explanation depends on the size (width, height) of the last feature map, a bilinear interpolation is performed in order to find the same dimensions as the input.

Occlusion is a sensitivity method that sweep a patch that occludes pixels over the images, and use the variations of the model prediction to deduce critical areas. In the context of these experiments, we took a patch size and a patch stride of 20.

$$g_i^{OC} = \boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x}_{[\boldsymbol{x}_i=0]})$$

RISE is a black-box method that consist of probing the model with randomly masked versions of the input image to deduce the importance of each pixel using the corresponding outputs. The masks $\boldsymbol{m} \sim \mathcal{M}$ are generated randomly in a subspace of the input space, then upsampled with a bilinear interpolation (once upsampled the masks are no longer binary).

As recommended in the original paper, we used $N = 8,000$ and $\mathbb{E}(\mathcal{M}) = 0.5$ for all the experiments.

$$g^{RI}(\boldsymbol{x}) = \frac{1}{\mathbb{E}(\mathcal{M})N} \sum_{i=0}^{N} \boldsymbol{f}(\boldsymbol{x} \odot \boldsymbol{m}_i)\boldsymbol{m}_i$$

F Fidelity with Insertion

Insertion is an evaluation procedure introduced in [9] at the same time as Deletion. Deletion 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). Insertion is the opposite of Deletion in that it assumes that the prediction score should go up faster for the most faithful explanations when pixels from the original image that are considered important are added to a baseline image (e.g., gray image). Metrics similar to Insertion are less common than Deletion in the literature that is why we focus on Deletion in the main paper. Moreover, just like Deletion, the Insertion score is largely influenced by the first steps [9]: the first pixels removed from the original image for deletion, and the first pixels inserted in the insertion case. Maximizing Insertion means exploring in a space close to the baseline, while maximizing Deletion means exploring a space around the original image. We suggest that for this reason, the Insertion score is not as relevant as Deletion.

	Method	*ResNet50V2*	*VGG16*	*EfficientNet*	*MobileNetV2*
	Random Baseline (ours)	0.233	0.166	0.115	0.138
White box	Saliency [1]	0.363	0.303	0.229	0.253
	Guided-Backprop. [3]	0.377	0.242	0.229	<u>0.361</u>
	DeconvNet [2]	0.307	0.221	0.229	0.166
	Grad.-Input [45]	0.194	0.219	0.098	0.126
	Integ.-Grad. [8]	0.264	0.237	0.143	0.166
	SmoothGrad [34]	<u>0.445</u>	<u>0.374</u>	<u>0.299</u>	0.307
	GradCAM [5]	**0.524**	**0.438**	**0.393**	**0.419**
Black box	Occlusion [2]	0.154	0.115	0.152	0.135
	RISE [9]	**0.546**	**0.484**	**0.439**	**0.443**
	Sobol ($\hat{\mathcal{S}}_{T_i}$) (ours)	<u>0.370</u>	<u>0.313</u>	<u>0.309</u>	<u>0.331</u>
	Sobol signed ($\hat{\mathcal{S}}^{\Delta}_{T_i}$) (ours)	0.258	0.290	0.204	0.211

Table S3: **Insertion** scores, obtained on 2,000 images from ImageNet validation set. Higher is better. Random consists in inserting randomly among the pixels remaining at each step. The first and second best results are **bolded** and <u>underlined</u>.

G Sanity check

We followed the procedure used by [52], namely the progressive reset of the network weights. We used an Inception V3 [70] model, each images shows the $\mathcal{S}_{T_i}$ explanation for the network in which

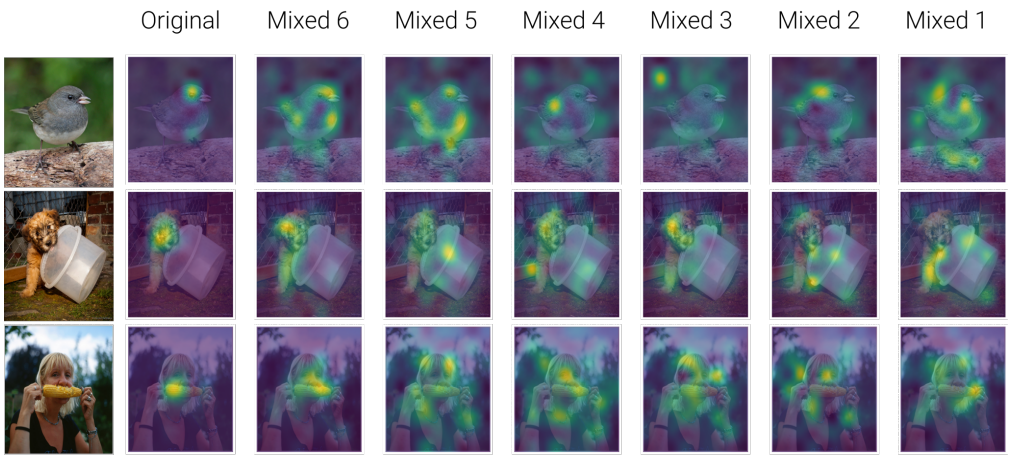

Figure S5: **Sanity Check** model weights are progressively reinitialized from Mixed 6 to Mixed 1 in InceptionV3 [70], demonstrating our method's sensitivity to model weights.

the upper layers (from logits) were reset. Fig. S5 shows that our method passes the sanity check: it turns out to be sensitive to the modification of the model weights.

H Word Deletion

For the bidirectional LSTM [71], the word embedding is in $\mathbb{R}^{300}$ and is initialized with the pre-trained GloVe embedding [72]. The layer has a hidden size of 64 (bidirectional architectures: 32 dimensions per direction). The resulting document representation is projected to 64 dimensions then 2 dimensions using fully connected layers, followed by a softmax and reached an accuracy of 89% on the test dataset.

For the BERT-based models, we use the Transformers library from HuggingFace [73] and more specifically the bert-base-uncased model. The final layer is tuned to minimize cross-entropy, with Adam optimizer [74] and initial learning rate of $1e^{-3}$ to reach an accuracy of 92% on the test dataset.

The observation that local perturbation: with the majority of words present, gets a better score is verified by playing on the threshold of the perturbation function. By decreasing the percentage of words removed on average we observe that a better deletion score is obtained.

	$\hat{\mathcal{S}}_{T_i}\Delta$ 50%	$\hat{\mathcal{S}}^{\Delta}_{T_i}$ 90%	$\hat{\mathcal{S}}_{T_i}\Delta$ 95%	Occlusion
Deletion	0.598	0.553	**0.527**	<u>0.531</u>

Table S4: **Word deletion scores** on the Bert based model when the perturbation threshold is modified to control the average presence of words in each generated perturbated input. Lower is better.