Don't Lie to Me! Robust and Efficient Explainability with Verified Perturbation Analysis

Thomas Fel[1,2,5†] **Melanie Ducoffe**[2,7†] **David Vigouroux**[2,4†] **Rémi Cadène**[1,3]
Mikael Capelle[6] **Claire Nicodème**[5] **Thomas Serre**[1,2]

[1]Carney Institute for Brain Science, Brown University, USA
[2]Artificial and Natural Intelligence Toulouse Institute
[3]Sorbonne Université, CNRS, France [4]IRT Saint-Exupery, France
[5]Innovation & Research Division, SNCF [6]Thales Alenia Space, France
[7]Airbus AI Research

Abstract

A plethora of attribution methods have recently been developed to explain deep neural networks. These methods use different classes of perturbations (e.g, occlusion, blurring, masking, etc) to estimate the importance of individual image pixels to drive a model's decision. Nevertheless, the space of possible perturbations is vast and current attribution methods typically require significant computation time to accurately sample the space in order to achieve high-quality explanations. In this work, we introduce EVA (Explaining using Verified Perturbation Analysis) – the first explainability method which comes with guarantees that an entire set of possible perturbations has been exhaustively searched. We leverage recent progress in verified perturbation analysis methods to directly propagate bounds through a neural network to exhaustively probe a – potentially infinite-size – set of perturbations in a single forward pass. Our approach takes advantage of the beneficial properties of verified perturbation analysis, i.e., time efficiency and guaranteed complete – sampling agnostic – coverage of the perturbation space – to identify image pixels that drive a model's decision. We evaluate EVA systematically and demonstrate state-of-the-art results on multiple benchmarks. Our code is freely available: `github.com/deel-ai/formal-explainability`

1. Introduction

Deep neural networks are now being widely deployed in many applications from medicine, transportation, and security to finance, with broad societal implications [40]. They are routinely used to making safety-critical decisions – often without an explanation as their decisions are notoriously hard to interpret.

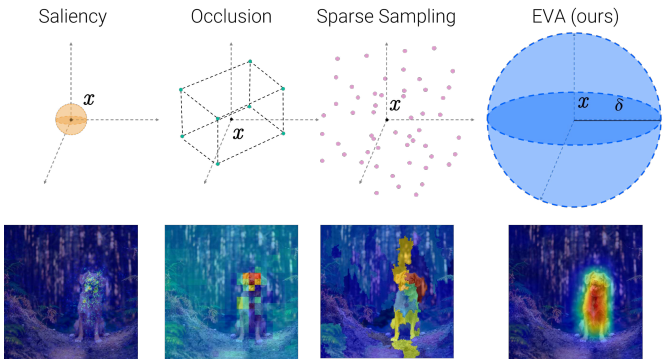

Figure 1. **Manifold exploration of current attribution methods.** Current methods assign an importance score to individual pixels using perturbations around a given input image $\boldsymbol{x}$. Saliency [56] uses infinitesimal perturbations around $\boldsymbol{x}$, Occlusion [71] switches individual pixel intensities on/off. More recent approaches [17,43,46,48,49] use (Quasi-) random sampling methods in specific perturbation spaces (occlusion of segments of pixels, blurring, ...). However, the choice of the perturbation space undoubtedly biases the results – potentially even introducing serious artifacts [26, 29, 38, 64]. We propose to use verified perturbation analysis to efficiently perform a complete coverage of a perturbation space around $\boldsymbol{x}$ to produce reliable and faithful explanations.

Many explainability methods have been proposed to gain insight into how network models arrive at a particular decision [17,24,43,46,48,49,53,55,61,65,71]. The applications of these methods are multiple – from helping to improve or debug their decisions to helping instill confidence in the reliability of their decisions [14]. Unfortunately, a severe limitation of these approaches is that they are subject to a confirmation bias: while they appear to offer useful expla-

nations to a human experimenter, they may produce incorrect explanations [2, 23, 59]. In other words, just because the explanations make sense to humans does not mean that they actually convey what is actually happening within the model. Therefore, the community is actively seeking for better benchmarks involving humans [12, 29, 37, 45].

In the meantime, it has been shown that some of our current and commonly used benchmarks are biased and that explainability methods reflect these biases – ultimately providing the wrong explanation for the behavior of the model [25, 29, 64]. For example, some of the current fidelity metrics [7, 18, 27, 34, 48] mask one or a few of the input variables (with a fixed value such as a gray mask) in order to assess how much they contribute to the output of the system. Trivially, if these variables are already set to the mask value in a given image (e.g., gray), masking these variables will not yield any effect on the model's output and the importance of these variables is poised to be underestimated. Finally, these methods rely on sampling a space of perturbations that is far too vast to be fully explored – e.g., LIME on a image divided in 64 segments image would need more than 10^{19} samples to test all possible perturbations. As a result, current attribution methods may be subject to bias and are potentially not entirely reliable.

To address the baseline issue, a growing body of work is starting to leverage adversarial methods [8, 29, 31, 42, 50] to derive explanations that reflect the robustness of the model to local adversarial perturbations. Specifically, a pixel or an image region is considered important if it allows the easy generation of an adversarial example. That is if a small perturbation of that pixel or image region yields a large change in the model's output. This idea has led to the design of several novel robustness metrics to evaluate the quality of explanations, such as Robustness-S_r [29]. For a better ranking on those robustness metrics, several methods have been proposed that make intensive use of adversarial attacks [29, 70], such as Greedy-AS for Robustness-S_r. However, these methods are computationally very costly – in some cases, requiring over 50 000 adversarial attacks per explanation – severely limiting the widespread adoption of these methods in real-world scenarios.

In this work, we propose to address this limitation by introducing EVA (Explaining using Verified perturbation Analysis), a new explainability method based on robustness analysis. Verified perturbation analysis is a rapidly growing toolkit of methods to derive bounds on the outputs of neural networks in the presence of input perturbations. In contrast to current attributions methods based on gradient estimation or sampling, verified perturbation analysis allows the full exploration of the perturbation space, see Fig. 1. We use a tractable certified upper bound of robustness confidence to derive a new estimator to help quantify the importance of input variables (i.e., those that matter the most). That is, the variables most likely to change the predictor's decision.

We can summarize our main contributions as follows:

- We introduce EVA, the first explainability method guaranteed to explore its entire set of perturbations using Verified Perturbation Analysis.
- We propose a method to scale EVA to large vision models and show that the exhaustive exploration of all possible perturbations can be done efficiently.
- We systematically evaluate our approach using several image datasets and show that it yields convincing results on a large range of explainability metrics
- Finally, we demonstrate that we can use the proposed method to generate class-specific explanations, and we study the effects of several verified perturbation analysis methods as a hyperparameter of the generated explanations.

2. Related Work

Attribution Methods. Our approach builds on prior attribution methods in order to explain the prediction of a deep neural network via the identification of input variables that support the prediction (typically pixels or image regions for images – which lead to importance maps shown in Fig. 1). "Saliency" was first introduced in [4] and consists in using the gradient of a classification score. It was later refined in [57, 61, 63, 65, 72] in the context of deep convolutional networks for classification. However, the image gradient only reflects the model's operation within an infinitesimal neighborhood around an input.Hence, it can yield misleading importance estimates [22] since gradients of the current large vision models are noisy [61]. Other methods rely on different image perturbations applied to images to produce importance maps that reflect the corresponding change in classification score resulting from the perturbation. Methods such as "Occlusion" [72], "LIME" [49], "RISE" [48], "Sobol" [17] or "HSIC" [46] leverage different sampling strategies to explore the space of perturbations around the image. For instance, Occlusion uses binary masks to occlude individual image regions, one at a time. RISE and HSIC combines these discrete masks to perturb multiple regions simultaneously. Sobol uses continuous masks for a finer exploration of the perturbation space.

Nevertheless, none of these methods are able to systematically cover the full space of perturbations. As a result, the corresponding explanations may not reliably reflect the true importance of pixels. In contrast, our approach comes with strong guarantees that can be derived from the verified perturbation analysis framework as it provides bounds on the perturbation space.

Robustness-based Explanation. To try to address the aforementioned limitations, several groups [8, 19, 29, 32, 33,

42, 60] have focused on the development of a new set of robustness-based evaluation metrics for trustworthy explanations. These new metrics are in contrast with the previous ones, which consisted in removing the pixels considered important in an explanation by substituting them with a fixed baseline – which inevitably introduces bias and artifacts [25, 26, 29, 38, 64]. Key to these new metrics is the assumption that when the important pixels are in their nominal (fixed) state, then perturbations applied to the complementary pixels – deemed unimportant – should not affect the model's decision to any great extent. The corollary that follows is that perturbations limited to the pixels considered important should easily influence the model's decision [29,42]. Going further along the path of robustness, abductive reasoning was used in [32] to compute optimal subsets with guarantees. The challenge consists in looking for the subset with the smallest possible cardinality – to guarantee the decision of the model. This work constituted one of the early successes of formal methods for explainability, but the approach was limited to low-dimensional problems and shallow neural networks. It was later extended to relax the subset minimum explanation by either providing multiple explanations, aggregating pixels in bundles [6] or by using local surrogates [9].

Some heuristics-oriented works also propose to optimize these new robustness based criteria and design new methods using a generative model [47] or adversarial attacks [29]. The latter approach requires searching for the existence or lack of an adversarial example for a multitude of ℓ_p balls around the input of interest. As a result, the induced computational cost is quite high as the authors used more than 50000 computations of adversarial examples to generate a single explanation.

More importantly, a failure to find an adversarial perturbation for a given radius does not guarantee that none exists. In fact, it is not uncommon for adversarial attacks to fail to converge – or fail to find an adversarial example – which will result in a failure to output an importance score. Our method addresses these issues while drastically reducing the computation cost. An added benefit of our approach is that verified perturbation analysis provides additional guarantees and hence opens the doors of certification which is a necessity for safety-critical applications.

Verified Perturbation Analysis. This growing field of research focuses on the development of methods that outer-approximate neural network outputs given some input perturbations. Simply put, for a given input $\boldsymbol{x}$ and a bounded perturbation $\boldsymbol{\delta}$, verification methods yield minimum $\underline{\boldsymbol{f}}(\boldsymbol{x})$ and maximum $\overline{\boldsymbol{f}}(\boldsymbol{x})$ bounds on the output of a model. Formally $\forall\ \boldsymbol{\delta}\ s.t\ ||\boldsymbol{\delta}||_p \leq \varepsilon$:

$$\underline{\boldsymbol{f}}(\boldsymbol{x}) \leq \boldsymbol{f}(\boldsymbol{x}+\boldsymbol{\delta}) \leq \overline{\boldsymbol{f}}(\boldsymbol{x}).$$

This allows us to explore the whole perturbation space without having to explicitly sample points in that space.

Early works focused on computing reachable lower and upper bounds based on satisfiability modulo theories [16, 36], and mixed-integer linear programming problems [66]. While these early results were encouraging, the proposed methods struggled even for small networks and image datasets. More recent work has led to the independent development of methods for computing looser certified lower and upper bounds more efficiently thanks to convex linear relaxations either in the primal or dual space [51]. While looser, those bounds remain tight enough to yield non-ubiquitous robustness properties on medium size neural networks. CROWN (hereafter called Backward) uses Linear Relaxation-based Perturbation Analysis (LiRPA) and achieves the tightest bound for efficient single neuron linear relaxation [58, 67, 73]. In addition, linear relaxation methods offer a wide range of possibilities with a vast trade-off between "tigthness" of the bounds and efficiency. These methods form two broad classes: 'forward' methods which propagate constant bounds (more generally affine relaxations from the input to the output of the network) also called Interval Bound Propagation (IBP, Forward, IBP+Forward) vs. 'backward' methods which bound the output of the network by affine relaxations given the internal layers of the network, starting from the output to the input. Note that these methods can be combined, e.g. (CROWN + IBP + Forward). For a thorough description of the LiRPA framework and theoretical analysis of the worst-case complexities of each variant, see [68]. In this work, we remain purposefully agnostic to the verification method used and opt for the most accurate LiRPA method applicable to the predictor. Our approach is based on the formal verification framework DecoMon, based on Keras [15].

3. Explainability with Verified Perturbation Analysis

Notation. We consider a standard supervised machine-learning classification setting with input space $\mathcal{X} \subseteq \mathbb{R}^d$, an output space $\mathcal{Y} \subseteq \mathbb{R}^c$, and a predictor function $\boldsymbol{f} : \mathcal{X} \to \mathcal{Y}$ that maps an input vector $\boldsymbol{x} = (x_1, \ldots, x_d)$ to an output $\boldsymbol{f}(\boldsymbol{x}) = (\boldsymbol{f}_1(\boldsymbol{x}), \ldots, \boldsymbol{f}_c(\boldsymbol{x}))$. We denote $\mathcal{B} = \{\boldsymbol{\delta} \in \mathbb{R}^d : ||\boldsymbol{\delta}||_p \leq \varepsilon\}$ the perturbation ball with radius $\varepsilon > 0$, with $p \in \{1, 2, \infty\}$. For any subset of indices $\boldsymbol{u} \subseteq \{1, \ldots, d\}$, we denote $\mathcal{B}_{\boldsymbol{u}}$ the ball without perturbation on the variables in $\boldsymbol{u}$: $\mathcal{B}_{\boldsymbol{u}} = \{\boldsymbol{\delta} : \boldsymbol{\delta} \in \mathcal{B},\ \boldsymbol{\delta}_{\boldsymbol{u}} = 0\}$ and $\mathcal{B}(\boldsymbol{x})$ the perturbation ball centered on $\boldsymbol{x}$. We denote the lower (resp. upper) bounds obtained with verification perturbation analysis as $\underline{\boldsymbol{f}}(\boldsymbol{x}, \mathcal{B}) = \Big(\underline{\boldsymbol{f}}_1(\boldsymbol{x}, \mathcal{B}), \ldots, \underline{\boldsymbol{f}}_c(\boldsymbol{x}, \mathcal{B})\Big)$, and $\overline{\boldsymbol{f}}(\boldsymbol{x}, \mathcal{B}) = \big(\overline{\boldsymbol{f}}_1(\boldsymbol{x}, \mathcal{B}), \ldots, \overline{\boldsymbol{f}}_c(\boldsymbol{x}, \mathcal{B})\big)$. Intuitively, these bounds delimit the output prediction for any perturbed sample in $\mathcal{B}(\boldsymbol{x})$.

3.1. The importance of setting the importance right

Different attribution methods implicitly assume different definitions of the notion of importance for input variables based either on game theory [43], the notion of conditional expectation of the score logits [48], their variance [17] or on some measure of statistical dependency between different areas of an input image and the output of the model [46]. In this work, we build on robustness-based explainability methods [29] which assume that a variable is important if small perturbations of this variable lead to large changes in the model decision. Conversely, a variable is said to be unimportant if changes to this variable only yield small changes in the model decision. From this intuitive assertion, we construct an estimator that we call *Adversarial overlap*.

3.2. Adversarial overlap

We go one step beyond previous work and propose to compute importance by taking into account not only the ability of individual variables to change the network's decision but also its confidence in the prediction. *Adversarial overlap* measures the extent to which a modification on a group of pixels can generate overlap between classes, i.e. generate a point close to $\boldsymbol{x}$ such that the attainable maximum of an unfavorable class c' can match the minimum of the initially predicted class c.

Indeed, if a modification of a pixel – or group of pixels – allows generating a new image that changes the decision of $\boldsymbol{f}$, this variable must be considered important. Conversely, if the decision does not change regardless of the value of the pixel, then the pixel can be left at its nominal value and should be considered unimportant.

Among the set of possible variable perturbations $\boldsymbol{\delta}$ around a point $\boldsymbol{x}$, we, therefore, look for points that can modify the decision with the most confidence. Hence our scoring criterion can be formulated as follows:

$$AO_c(\boldsymbol{x}, \mathcal{B}) = \max_{\substack{\boldsymbol{\delta} \in \mathcal{B} \\ c' \neq c}} \boldsymbol{f}_{c'}(\boldsymbol{x} + \boldsymbol{\delta}) - \boldsymbol{f}_c(\boldsymbol{x} + \boldsymbol{\delta}). \quad (1)$$

Intuitively, this score represents the confidence of the "best" adversarial perturbation that can be found in the perturbation ball $\mathcal{B}$ around $\boldsymbol{x}$. Throughout the article, when c is not specified, it is assumed that $c = \arg\max \boldsymbol{f}(\boldsymbol{x})$.

In order to estimate this criterion, a naive strategy could be to use adversarial attacks to search within $\mathcal{B}$. However, when they converge - which is not ensured, such methods only explore certain points of the considered space, thus giving no guarantee regarding the optimality of the solution. Moreover, adversarial methods have no guarantee of success and therefore cannot ensure a valid score under every circumstance. Finally, the large dimensions of the current datasets make exhaustive searches impossible.

To overcome these issues, we take advantage of one of the main results from verified perturbation analysis to derive a guaranteed upper bound on the criterion introduced in Eq. 1. We can upper bound the *adversarial overlap* criterion as follows:

$$AO(\boldsymbol{x}, \mathcal{B}) \leq \overline{AO}(\boldsymbol{x}, \mathcal{B}) = \max_{c' \neq c} \overline{\boldsymbol{f}}_{c'}(\boldsymbol{x}, \mathcal{B}) - \underline{\boldsymbol{f}}_c(\boldsymbol{x}, \mathcal{B}).$$

The computation of this upper bound becomes tractable using any verified perturbation analysis method.

For example, $\overline{AO}(\boldsymbol{x}, \mathcal{B}) \leq 0$ guarantees that no adversarial perturbation is possible in the perturbation space.[1] Our upper bound $\overline{AO}(\boldsymbol{x}, \mathcal{B})$ corresponds to the difference between the verified lower bound of the class of interest c and the maximum over the verified upper bounds among the other classes. Thus, when important variables are modified (e.g the head of the dog in Fig. 2, using $\mathcal{B}$), the lower bound for the class of interest will get smaller than the upper bound of the adversary class. On the other hand, this overlap is not possible when important variables are fixed (e.g in Fig. 2 when the head of the dog is fixed, using $\mathcal{B}_{\boldsymbol{u}}$). We now demonstrate how to leverage this score to derive an efficient estimator of variable importance.

3.3. EVA

We are willing to assign a higher importance score for a variable allowing (1) a change in a decision, (2) a greater adversarial – thus a solid change of decision. Modifying all variables gives us an idea of the robustness of the model. In the same way, the modification of all variables without the subset $\boldsymbol{u}$ allows quantifying the change of the strongest adversarial perturbation and thus quantifies the importance of the variables $\boldsymbol{u}$. Intuitively, if an important variable $\boldsymbol{u}$ is discarded, then it will be more difficult, if not impossible, to succeed in finding any adversarial perturbation. Specifically, removing the possibility to modify $\boldsymbol{x}_{\boldsymbol{u}}$ allows us to reveal its importance by taking into account its possible interactions.

The complexity of current models means that the variables are not only treated individually in neural network models but collectively. In order to capture these higher-order interactions, our method consists in measuring the *adversarial overlap* allowed by all the variables together $\overline{AO}(\boldsymbol{x}, \mathcal{B})$ – thus taking into account their interactions – and then forbidding to play on a group of variables $\overline{AO}(\boldsymbol{x}, \mathcal{B}_{\boldsymbol{u}})$ to estimate the importance of $\boldsymbol{u}$. Making the interactions of $\boldsymbol{u}$ disappear reveals their importance. Note that several works have mentioned the importance of taking into account the interactions of the variables when calculating the importance [17, 20, 30, 48]. Formally, we introduce EVA (Explainability using Verified perturbation Analysis) that measure the drop in *adversarial overlap* when we fixed the variables $\boldsymbol{u}$:

$$\text{EVA}(\boldsymbol{x}, \boldsymbol{u}, \mathcal{B}) \equiv \overline{AO}(\boldsymbol{x}, \mathcal{B}) - \overline{AO}(\boldsymbol{x}, \mathcal{B}_{\boldsymbol{u}}). \quad (2)$$

[1] Note that with adversarial attacks, failure to find an adversarial example does not guarantee that it does not exist.

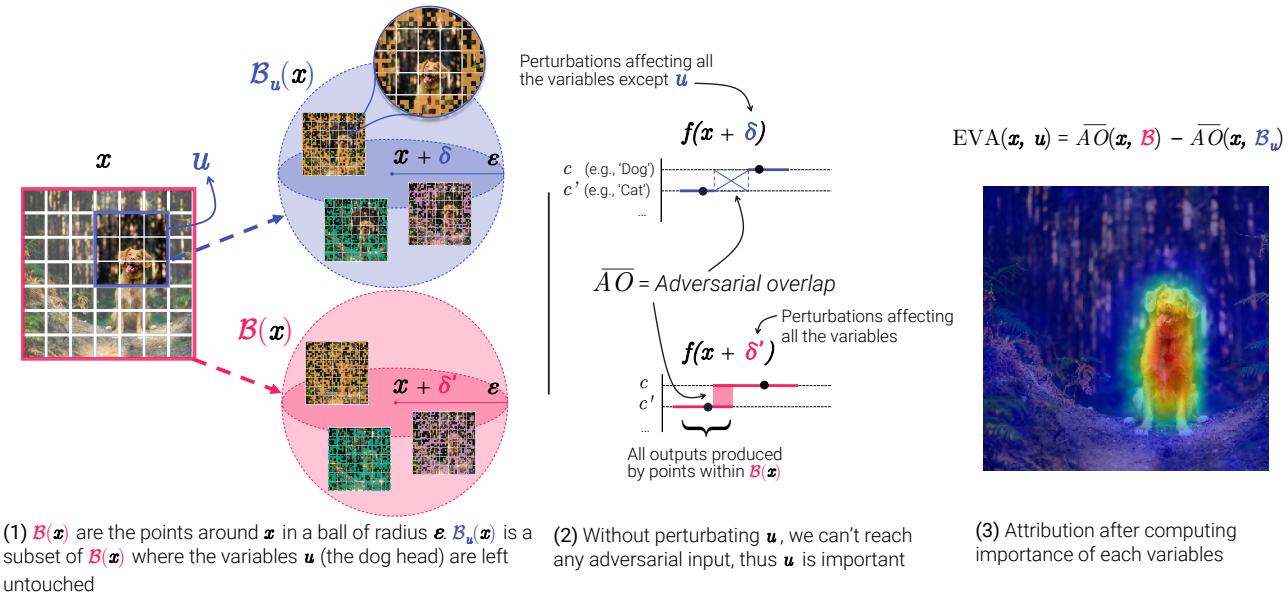

Figure 2. **EVA attribution method.** In order to compute the importance for a group of variables u – for instance the dog's head – the first step (1) consists in designing the perturbation ball $\mathcal{B}_u(x)$. This ball is centered in x and contain all the possible images perturbed by δ $s.t$ $||\delta||_p \leq \varepsilon, ||\delta_u||_p = 0$ which do not perturb the variables u. Using verified perturbation analysis, we then compute the *adversarial overlap* $\overline{AO}(x, \mathcal{B}_u)$ which corresponds to the overlapping between the class c – here dog – and c', the maximum among the other classes. Finally, the importance score for the variable u corresponds to the drop in *adversarial overlap* when u cannot be perturbed, thus the difference between $\overline{AO}(x, \mathcal{B})$ and $\overline{AO}(x, \mathcal{B}_u)$. Specifically, this measures how important the variables u are for changing the model's decision.

As explained in Fig. 2, the estimator requires two passes of the perturbation analysis method; one for $\overline{AO}(\mathcal{B})$, and the other for $\overline{AO}(\mathcal{B}_u)$: the first term consists in measuring the *adversarial overlap* by modifying all the variables, the second term measures the adversarial surface when fixing the variables of interest u. In other words, EVA measures the *adversarial overlap* that would be left if the variables u were to be fixed.

From a theoretical point of view, we notice that EVA - under reasonable assumptions - yield the optimal subset of variables to minimize the theoretical Robustness-S_r metric (see Theorem C.6). From a computational point of view, we can note that the first term of the *adversarial overlap* $\overline{AO}(x, \mathcal{B})$ – as it does not depend on u – can be calculated once and re-used to evaluate the importance of any other variables considered. Moreover, contrary to an iterative process method [21, 29, 32], each importance can be evaluated independently and thus benefit from the parallelization of modern neural networks. Finally, the experiments in Section 4 show that even with two calls to $\overline{AO}$ per variables, our method remains much faster than the one based on sampling or on adversarial attacks (such as Greedy-AS or Greedy-AO, see appendix B).

In this work, the verified perturbation-based analysis considered is not always adapted to high dimensional models, especially those running on ImageNet [13]. We are confident that the verification methods will progress towards more scalability in the near future, enabling the original version of EVA on deeper models.

In the meantime, we introduce an empirical method that allows to scale EVA to high dimensional models. This method sacrifices theoretical guarantees, but the results section reveals that it may be a good compromise.

3.4. Scaling to larger models

We propose a second version of EVA, which is a combination of sampling and verification perturbation analysis. The aim of this hybrid method is twofold: (***i***) take advantage of sampling to approach the bounds of an intermediate layer in a potentially very large model, (***ii***) then complete only the rest of the propagations with verified perturbation analysis and thus move towards the native EVA method which benefits from theoretical guarantees. Note that, combining verification methods with empirical methods (a.k.a adversarial training) has notably been proposed in [5] for robust training.

Specifically, our technique consists of splitting the model into two parts, and (***i***) estimating the bounds of an intermediate layer using sampling, (***ii***) propagating these empirical intermediate bounds onto the second part of the model with verified perturbation analysis methods.

For the first step (***i***) we consider the original predictor f

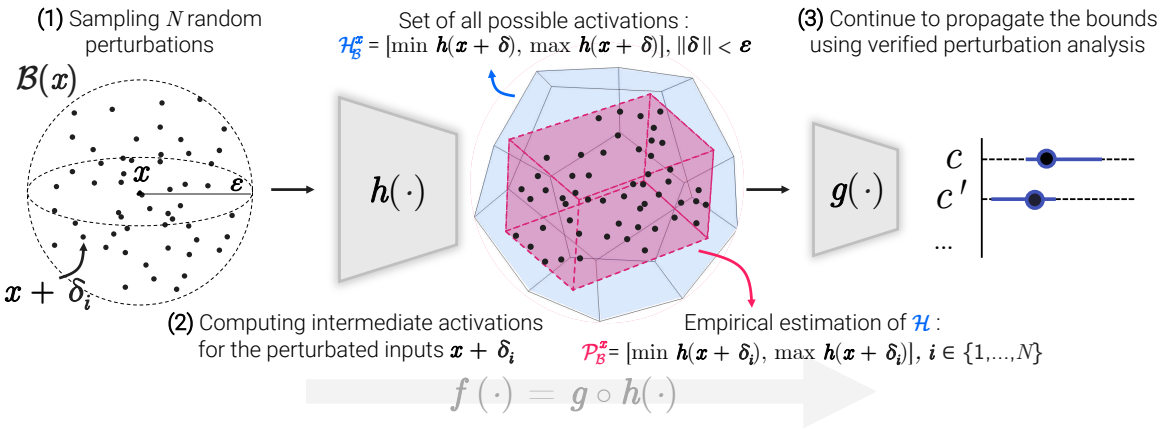

Figure 3. **Scaling strategy.** In order to scale to very large models, we propose to estimate the bounds of an intermediate layer's activations empirically by (1) Sampling N input perturbations and (2) calculating empirical bounds on the resulting activations for the layer $\boldsymbol{h}(\cdot)$. We can then form the set $\mathcal{P}_{\mathcal{B}}^{x}$ which is a subset of the true bounds $\mathcal{H}_{\mathcal{B}}^{x}$ since the sampling is never exhaustive. We can then plug this set into a verified perturbation analysis method (3) and continue the forward propagation of the inputs through the rest of the network.

as a composition of functions $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{g} \circ \boldsymbol{h}(\boldsymbol{x})$. For deep neural networks, $\boldsymbol{h}(\cdot)$ is a function that maps input to an intermediate feature space and $\boldsymbol{g}(\cdot)$ is a function that maps this same feature space to the classification.

We propose to empirically estimate bounds $(\underline{\boldsymbol{h}}_{\mathcal{B}}^{\boldsymbol{x}}, \overline{\boldsymbol{h}}_{\mathcal{B}}^{\boldsymbol{x}})$ for the intermediate activations $\boldsymbol{h}(\cdot) \in \mathbb{R}^{d'}$ using Monte-Carlo sampling on the perturbation $\boldsymbol{\delta} \in \mathcal{B}$. Formally:

$$\forall j \in [0, \dots, d'], \ \underline{\boldsymbol{h}}_{\mathcal{B}}^{\boldsymbol{x}}[j] = \min_{\boldsymbol{\delta}_1, \dots \boldsymbol{\delta}_i, \dots \boldsymbol{\delta}_n \overset{\text{iid}}{\sim} U(\mathcal{B})} \boldsymbol{h}(\boldsymbol{x} + \boldsymbol{\delta}_i)[j]$$
$$\overline{\boldsymbol{h}}_{\mathcal{B}}^{\boldsymbol{x}}[j] = \max_{\boldsymbol{\delta}_1, \dots \boldsymbol{\delta}_i, \dots \boldsymbol{\delta}_n \overset{\text{iid}}{\sim} U(\mathcal{B})} \boldsymbol{h}(\boldsymbol{x} + \boldsymbol{\delta}_i)[j].$$

Obviously, since the sampling is never exhaustive, the bounds obtained underestimate the true maximum $\overline{\boldsymbol{h}}_{\mathcal{B}}^{\boldsymbol{x}} \leq \max \boldsymbol{h}(\boldsymbol{x} + \boldsymbol{\delta})$ and overestimates the true minimum $\underline{\boldsymbol{h}}_{\mathcal{B}}^{\boldsymbol{x}} \geq \min \boldsymbol{h}(\boldsymbol{x} + \boldsymbol{\delta})$ as illustrated in the Fig. 3. In a similar way, we define $\underline{\boldsymbol{h}}_{\mathcal{B}_{\boldsymbol{u}}}^{\boldsymbol{x}}$ and $\overline{\boldsymbol{h}}_{\mathcal{B}_{\boldsymbol{u}}}^{\boldsymbol{x}}$ when $\boldsymbol{\delta} \in \mathcal{B}_{\boldsymbol{u}}$. Once the empirical bounds are estimated, we may proceed to the second step and use the obtained bounds to form the new perturbation set

$$\mathcal{P}_{\mathcal{B}}^{\boldsymbol{x}} = [\underline{\boldsymbol{h}}_{\mathcal{B}}^{\boldsymbol{x}} - \boldsymbol{h}(\boldsymbol{x}), \overline{\boldsymbol{h}}_{\mathcal{B}}^{\boldsymbol{x}} - \boldsymbol{h}(\boldsymbol{x})].$$

Intuitively, this set bounds the intermediate activations obtained empirically and can then be fed to a verified perturbation verification method.

We then carry out the end of the bounds propagation in the usual way, using verified perturbation analysis. This amounts to computing bounds for the outputs of the network for all possible activations contained in our empirical bounds. The only change is that we no longer operate in the pixel space $\boldsymbol{x}$ with the ball $\mathcal{B}$, but in the activation space $\boldsymbol{h}(\cdot)$ with the perturbations set $\mathcal{P}_{\mathcal{B}}^{\boldsymbol{x}}$. The importance score of a set of variables $\boldsymbol{u}$ is then :

$$\text{EVA}_{\text{hybrid}}(\boldsymbol{x}, \boldsymbol{u}, \mathcal{B}) \equiv \text{EVA}(\boldsymbol{h}(\boldsymbol{x}), \boldsymbol{u}, \mathcal{P}_{\mathcal{B}}^{\boldsymbol{x}}).$$

This hybrid approach allows us to use EVA on state-of-the-art models and thus to benefit from our method while remaining tractable. We believe this extension to be a promising step towards robust explanations on deeper networks.

4. Experiments

To evaluate the benefits and reliability of our explainability method, we performed several experiments on a standard dataset, using a set of common explainability metrics against EVA. In order to test the fidelity of the explanations produced by our method, we compare them to that of 10 other explainability methods using the (1) Deletion, (2) Insertion, and (3) MuFidelity metrics. As it has been shown that these metrics can exhibit biases, we completed the benchmark by adding the (4) Robustness-S_r metric. Each score is averaged over 500 samples.

We evaluated these 4 metrics on 3 image classification datasets, namely MNIST [41], CIFAR-10 [39] and ImageNet [13]. Through these experiments, the explanations were generated using EVA estimator introduced in Equation 2. The importance scores were not evaluated pixel-wise but on each cell of the image after having cut it into a grid of 12 sides (see Fig. 2). For MNIST and Cifar-10, we used $\varepsilon = 0.5$, whereas for ImageNet $\varepsilon = 5$. Concerning the verified perturbation analysis method, we used (IBP+Forward+Backward) for MNIST, and (IBP+Forward) on Cifar-10 and $p = \infty$. For computational purposes, we used the hybrid approach introduce in Section 3.4 for ImageNet using the penultimate layer (FC-4096) as the intermediate layer $\boldsymbol{h}(\cdot)$. We give in Appendix the complete set of hyperparameters used for the other explainability methods, metrics considered as well as the architecture of the models used on MNIST and Cifar-10.

	MNIST					Cifar-10					ImageNet				
	Del.↓	Ins.↑	Fid.↑	Rob.↓	Time	Del.↓	Ins.↑	Fid.↑	Rob.↓	Time	Del.↓	Ins.↑	Fid.↑	Rob.↓	Time
Saliency [56]	.193	.633	.378	.071	0.04	.171	.172	-.021	.026	0.16	.057	.126	.035	.769	0.36
GradInput [3]	.222	.611	.107	.074	0.04	.200	.143	-.018	.095	0.17	.057	.050	.023	.814	0.36
SmoothGrad [61]	.185	.621	.331	.070	1.91	.174	.181	.092	.048	9.07	**.051**	.069	.019	.809	9.63
VarGrad [54]	.207	.555	.216	.077	1.76	.183	.211	-.012	.193	9.07	.098	.201	.021	.787	9.62
InteGrad [65]	.209	.615	.108	.074	1.77	.194	.171	-.016	.154	7.19	.058	.052	.023	.813	8.39
Occlusion [3]	.247	.545	.137	.082	0.04	.217	.290	.105	.232	1.13	.100	.266	.026	.821	4.97
GradCAM [53]	n/a	n/a	n/a	n/a	n/a	.297	.282	.056	.195	0.39	.073	.232	.036	.817	0.18
GradCAM++ [10]	n/a	n/a	n/a	n/a	n/a	.270	**.326**	.102	.094	0.39	.074	.285	.054	.800	0.19
RISE [48]	.248	.558	.133	.093	2.26	.196	.273	.157	.385	20.5	.074	.276	**.154**	.818	1215
Greedy-AS [29]	.260	.497	.110	**.061**	335	.205	.264	-.003	**.013**	4618	.088	.047	.023	**.612**	180056
EVA (ours)	**.089**	**.736**	**.428**	.069	1.29	**.164**	.290	**.352**	.025	12.7	.070	**.289**	.048	.758	6454

Table 1. Results on Deletion (Del.), Insertion (Ins.), μFidelity (Fid.) and Robustness-S_r (Rob.) metrics. Time in seconds corresponds to the generation of 500 (MNIST/CIFAR-10) and 100 (ImageNet) explanations on an Nvidia P100. Note that EVA is the only method with guarantees that the entire set of possible perturbations has been exhaustively searched. Verified perturbation analysis with IBP + Forward + Backward is used for MNIST, with Forward only for CIFAR-10 and with our hybrid strategy described in Section.3.4 for ImageNet. Grad-CAM and Grad-CAM++ are not calculated on the MNIST dataset since the network only has dense layers. The first and second best results are in **bold** and underlined, respectively.

4.1. Comparison with the state of the art

There is a general consensus that fidelity is a crucial criterion for an explanation method. That is, if an explanation is used to make a critical decision, then users are expecting it to reflect the true decision-making process underlying the model and not just a consensus with humans. Failure to do so could have disastrous consequences. Pragmatically, these metrics assume that the more faithful an explanation is, the faster the prediction score should drop when pixels considered important are changed. In Table 1, we present the results of the Deletion [48] (or $1 - AOPC$ [52]) metric for the MNIST and Cifar-10 datasets on 500 images sampled from the test set. TensorFlow [1] and the Keras API [11] were used to run the models and Xplique [18] for the explainability methods. In order to evaluate the methods, the metrics require a baseline and several were proposed [29,64], but we chose to keep the choice of [29] using their random baseline.

We observe that EVA is the explainability method getting the best Deletion, Insertion, and μFidelity scores on MNIST, and is just behind Greedy-AS on Robustness-S_r. This can be explained by the fact that the Robustness metric uses the adversarial attack PGD [44], which is the same one used to generate Greedy-AS, thus biasing the adversarial search. Indeed, if PGD does not find an adversarial perturbation using a subset $\boldsymbol{u}$ does not give a guarantee of the robustness of the model, just that the adversarial perturbation could be difficult to reach with PGD.

For Cifar-10, EVA remains overall the most faithful method according to Deletion and μFidelity, and obtains the second score in Insertion behind Grad-Cam++ [10]. Finally, we notice that if Greedy-AS [29] allows us to obtain a good Robustness-S_r score, but this comes with a considerable computation time, which is not the case of EVA which is much more efficient. Eventually, EVA is a very good compromise for its relevance to commonly accepted explainability metrics and more recent robustness metrics.

ImageNet After having demonstrated the potential of the method on vision datasets of limited size, we consider the case of ImageNet which has a significantly higher level of dimension. The use of verified perturbation analysis methods other than IBP is not easily scalable on these datasets. We, therefore, used the hybrid method introduced in Section 3.4 in order to estimate the bounds in a latent space and then plug those bounds into the perturbation analysis to get the final *adversarial overlap* score.

Table 1 shows the results obtained with the empirical method proposed in Section 3.4. We observe that even with this relaxed estimation, EVA is able to score high on all the metrics. Indeed, EVA obtains the best score on the Insertion metric and ranks second on μFidelity and Robustness-S_r. Greedy-AS ranks first on Robustness-S_r at the expense of the other scores where it performs poorly. Finally, both RISE and SmoothGrad perform well on all the fidelity metrics but collapse on the robustness metric. Extending results with ablations of EVA, including Greedy-AO, are available in Table 3.

Qualitatively, Fig. 5 shows examples of explanations produced on the ImageNet VGG-16 model. The explanations produced by EVA are more localized than Grad-CAM or RISE, while being less noisy than the gradient-based or Greedy-AS methods.

In addition, as the literature on verified perturbation analysis is evolving rapidly we can conjecture that the advances

will benefit the proposed explainability method. Indeed, EVA proved to be the most effective on the benchmark when an accurate formal method was used. After demonstrating the performance of the proposed method, we study its ability to generate class explanations specific.

4.2. Tighter bounds lead to improved explanations

	Tightness↓	Del.↓	Ins.↑	Fid.↑	Rob.↓
IBP	4.58	.148	.588	.222	.077
Forward	2.66	.150	.580	.209	.078
Backward	<u>2.36</u>	<u>.115</u>	<u>.607</u>	<u>.274</u>	<u>.074</u>
IBP + Fo. + Ba.	**1.55**	**.089**	**.736**	**.428**	**.069**

Table 2. **Impact of the verified perturbation analysis method on EVA.** Results of EVAon Tightness, Deletion (Del.), Insertion (Ins.), Fidelity (Fid.) and *Robustness-*S_{r} (Rob.) metrics obtained on MNIST. The Tightness score corresponds to the average adversarial surface. A lower Tightness score indicates that the method is more precise: it reaches tighter bound, resulting in better explanations and superior scores on the other metrics. The first and second best results are respectively in **bold** and <u>underlined</u>.

The choice of the verified perturbation analysis method is a hyperparameter of EVA. Hence, it is interesting to see the effect of the choice of this hyperparameter on the previous benchmark. We recall that only the MNIST dataset could benefit from the (IBP+Forward+Backward) combo. Table 2 reports the results of the fidelity metrics using other verified perturbation analysis methods. We also report a tightness score which corresponds to the average of the *adversarial overlap* : $\mathbb{E}_{\boldsymbol{x}\sim\mathcal{X}}(\overline{AO}(\boldsymbol{x},\mathcal{B}))$. Specifically, a low score indicates that the verification method is precise, meaning that the over-approximation is closer to the actual value. It should be noted that the true value is intractable, but remains the same across all three tested cases. We observe that the tighter the bounds, the higher the scores. This allows us to conjecture that the more scalable the formal methods will become, the better the quality of the generated explanations will be. We perform additional experiments to ensure that the certified component of EVA score is significant by comparing EVA to a sampling-based version of EVA. The details of these experiments are available in Appendix B.

4.3. Targeted Explanations

In some cases, it is instructive to look at the explanations for unpredicted classes in order to get information about the internal mechanisms of the models studied. Such explanations allow us to highlight contrastive features: elements that should be changed or whose absence is critical. Our method allows us to obtain such explanations: for a given input, we are then exclusively interested in the class we are trying to explain, without looking at the other decisions. Formally, for a given targeted class c' the *adversarial overlap* (Equation 1) become $AO(\boldsymbol{x},\mathcal{B}) = \max_{\boldsymbol{\delta}\in\mathcal{B}} \boldsymbol{f}_{c'}(\boldsymbol{x}+\boldsymbol{\delta}) - \boldsymbol{f}_{c}(\boldsymbol{x}+\boldsymbol{\delta})$. Moreover, by splitting the perturbation ball into a positive one $\mathcal{B}^{(+)} = \{\boldsymbol{\delta}\in\mathcal{B} \,:\, \boldsymbol{\delta}_i \geq 0,\ \forall i \in \{1,...,d\}\}$ and a negative one $\mathcal{B}^{(-)} = \{\boldsymbol{\delta}\in\mathcal{B} \,:\, \boldsymbol{\delta}_i \leq 0,\ \forall i \in \{1,...,d\}\}$, one can deduce which direction – adding or removing the black line in the case of gray-scaled images – will impact the most the model decision.

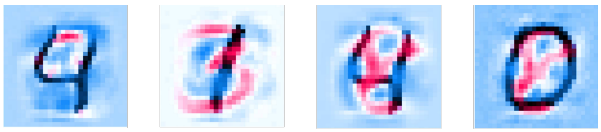

Figure 4. **Targeted explanations.** Generated explanations for a decision other than the one predicted by the model. The class explained is indicated at the bottom of each sample, e.g., the first sample is a '4' and the explanation is for the class '9'. As indicated in section 4.3, the red areas indicate that a black line should be added and the blue areas that it should be removed. More examples are available in the Appendix.

We generate targeted explanations on the MNIST dataset using (IBP+Forward+Backward). For several inputs, we generate the explanation for the 10 classes. Fig. 7 shows 4 examples of targeted explanations, the target class c' is indicated at the bottom. The red areas indicate that adding a black line increases the *adversarial overlap* with the target class. Conversely, the blue areas indicate where the increase of the score requires removing black lines. All other results can be found in the Appendix. In addition to favorable results on the fidelity metrics and guarantees provided by the verification methods, EVA can provide targeted explanations that are easily understandable by humans, which are two qualities that make them a candidate of choice to meet the recent General Data Protection Regulation (GDPR) adopted in Europe [35]. More examples are available in the Appendix H.

5. Conclusion

In this work, we presented the first explainability method that uses verification perturbation analysis that exhaustively explores the perturbation space to generate explanations. We presented an efficient estimator that yields explanations that are state-of-the-art on current metrics. We also described a simple strategy to scale up perturbation verification methods to complex models. Finally, we showed that this estimator can be used to form easily interpretable targeted explanations.

We hope that this work will for searching for safer and more efficient explanation methods for neural networks – and that it will inspire further synergies with the field of formal verification.

References

[1] 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, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. 7, 16

[2] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In Advances in Neural Information Processing Systems (NIPS), 2018. 2

[3] 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. 7, 15

[4] 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. 2

[5] Mislav Balunovic and Martin Vechev. Adversarial training and provable defenses: Bridging the gap. In Proceedings of the International Conference on Learning Representations (ICLR), 2019. 5

[6] Shahaf Bassan and Guy Katz. Towards formal approximated minimal explanations of neural networks. arXiv preprint arXiv:2210.13915, 2022. 3

[7] 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. 2, 14, 16

[8] Akhilan Boopathy, Sijia Liu, Gaoyuan Zhang, Cynthia Liu, Pin-Yu Chen, Shiyu Chang, and Luca Daniel. Proper network interpretability helps adversarial robustness in classification. In Proceedings of the International Conference on Machine Learning (ICML), 2020. 2

[9] Ryma Boumazouza, Fahima Cheikh-Alili, Bertrand Mazure, and Karim Tabia. Asteryx: A model-agnostic sat-based approach for symbolic and score-based explanations. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management, pages 120–129, 2021. 3

[10] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2018. 7, 15

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

[12] Julien Colin, Thomas Fel, Rémi Cadène, and Thomas Serre. What i cannot predict, i do not understand: A human-centered evaluation framework for explainability methods. Advances in Neural Information Processing Systems (NeurIPS), 2021. 2

[13] 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. 5, 6

[14] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. ArXiv e-print, 2017. 1

[15] Ducoffe, Melanie. Decomon: Automatic certified perturbation analysis of neural networks, 2021. 3

[16] Ruediger Ehlers. Formal verification of piece-wise linear feed-forward neural networks. In International Symposium on Automated Technology for Verification and Analysis, pages 269–286. Springer, 2017. 3

[17] Thomas Fel, Remi Cadene, Mathieu Chalvidal, Matthieu Cord, David Vigouroux, and Thomas Serre. Look at the variance! efficient black-box explanations with sobol-based sensitivity analysis. In Advances in Neural Information Processing Systems (NeurIPS), 2021. 1, 2, 4

[18] Thomas Fel, Lucas Hervier, David Vigouroux, Antonin Poche, Justin Plakoo, Remi Cadene, Mathieu Chalvidal, Julien Colin, Thibaut Boissin, Louis Béthune, Agustin Picard, Claire Nicodeme, Laurent Gardes, Gregory Flandin, and Thomas Serre. Xplique: A deep learning explainability toolbox. Workshop, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2022. 2, 7, 14

[19] Thomas Fel and David Vigouroux. Representativity and consistency measures for deep neural network explanations. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022. 2

[20] Gabriel Ferrettini, Elodie Escriva, Julien Aligon, Jean-Baptiste Excoffier, and Chantal Soulé-Dupuy. Coalitional strategies for efficient individual prediction explanation. Information Systems Frontiers, pages 1–27, 2021. 4

[21] 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. 5

[22] Sahra Ghalebikesabi, Lucile Ter-Minassian, Karla DiazOrdaz, and Chris C Holmes. On locality of local explanation models. Advances in Neural Information Processing Systems (NeurIPS), 2021. 2

[23] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2017. 2

[24] Mara Graziani, Iam Palatnik de Sousa, Marley MBR Vellasco, Eduardo Costa da Silva, Henning Müller, and Vincent Andrearczyk. Sharpening local interpretable model-agnostic explanations for histopathology: improved understandability and reliability. In Medical Image Computing and Computer Assisted Intervention (MICCAI). Springer, 2021. 1

[25] Peter Hase, Harry Xie, and Mohit Bansal. The out-of-distribution problem in explainability and search methods for feature importance explanations. Advances in Neural Information Processing Systems (NeurIPS), 2021. 2, 3

[26] Johannes Haug, Stefan Zürn, Peter El-Jiz, and Gjergji Kasneci. On baselines for local feature attributions. arXiv preprint arXiv:2101.00905, 2021. 1, 3

[27] Anna Hedström, Leander Weber, Dilyara Bareeva, Franz Motzkus, Wojciech Samek, Sebastian Lapuschkin, and Marina M-C Höhne. Quantus: an explainable ai toolkit for responsible evaluation of neural network explanations. The Journal of Machine Learning Research (JMLR), 2022. 2

[28] 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 (NeurIPS), 2019. 15

[29] Cheng-Yu Hsieh, Chih-Kuan Yeh, Xuanqing Liu, Pradeep Ravikumar, Seungyeon Kim, Sanjiv Kumar, and Cho-Jui Hsieh. Evaluations and methods for explanation through robustness analysis. In Proceedings of the International Conference on Learning Representations (ICLR), 2021. 1, 2, 3, 4, 5, 7, 12, 13, 15, 16

[30] Marouane Il Idrissi, Nicolas Bousquet, Fabrice Gamboa, Bertrand Iooss, and Jean-Michel Loubes. On the coalitional decomposition of parameters of interest, 2023. 4

[31] Marouane Il Idrissi, Vincent Chabridon, and Bertrand Iooss. Developments and applications of shapley effects to reliability-oriented sensitivity analysis with correlated inputs. Environmental Modelling & Software, 2021. 2

[32] Alexey Ignatiev, Nina Narodytska, and Joao Marques-Silva. Abduction-based explanations for machine learning models. In Advances in Neural Information Processing Systems (NeurIPS), 2019. 2, 3, 5, 13

[33] Alexey Ignatiev, Nina Narodytska, and Joao Marques-Silva. On relating explanations and adversarial examples. In Advances in Neural Information Processing Systems (NeurIPS), 2019. 2, 13

[34] Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL Short Papers), 2020. 2

[35] Margot E Kaminski. The right to explanation, explained. In Research Handbook on Information Law and Governance. Edward Elgar Publishing, 2021. 8

[36] Guy Katz, Clark Barrett, David L Dill, Kyle Julian, and Mykel J Kochenderfer. Reluplex: An efficient smt solver for verifying deep neural networks. In International Conference on Computer Aided Verification, pages 97–117. Springer, 2017. 3

[37] Sunnie S. Y. Kim, Nicole Meister, Vikram V. Ramaswamy, Ruth Fong, and Olga Russakovsky. HIVE: Evaluating the human interpretability of visual explanations. In Proceedings of the IEEE European Conference on Computer Vision (ECCV), 2022. 2

[38] Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (un) reliability of saliency methods. 2019. 1, 3

[39] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images, 2009. 6

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

[41] Yann LeCun and Corinna Cortes. MNIST handwritten digit database, 2010. 6

[42] Zhong Qiu Lin, Mohammad Javad Shafiee, Stanislav Bochkarev, Michael St Jules, Xiao Yu Wang, and Alexander Wong. Do explanations reflect decisions? a machine-centric strategy to quantify the performance of explainability algorithms. In Advances in Neural Information Processing Systems (NIPS), 2019. 2, 3

[43] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (NIPS), 2017. 1, 4

[44] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. Proceedings of the International Conference on Learning Representations (ICLR), 2018. 7, 16

[45] Giang Nguyen, Daeyoung Kim, and Anh Nguyen. The effectiveness of feature attribution methods and its correlation with automatic evaluation scores. Advances in Neural Information Processing Systems (NeurIPS), 2021. 2

[46] Paul Novello, Thomas Fel, and David Vigouroux. Making sense of dependence: Efficient black-box explanations using dependence measure. In Advances in Neural Information Processing Systems (NeurIPS), 2022. 1, 2, 4

[47] Matthew O'Shaughnessy, Gregory Canal, Marissa Connor, Mark Davenport, and Christopher Rozell. Generative causal explanations of black-box classifiers. In Advances in Neural Information Processing Systems (NeurIPS), 2020. 3

[48] 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, 4, 7, 15, 16

[49] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In Knowledge Discovery and Data Mining (KDD), 2016. 1, 2

[50] Alexis Ross, Himabindu Lakkaraju, and Osbert Bastani. Learning models for actionable recourse. Advances in Neural Information Processing Systems (NeurIPS), 2021. 2

[51] Hadi Salman, Greg Yang, Huan Zhang, Cho-Jui Hsieh, and Pengchuan Zhang. A convex relaxation barrier to tight robustness verification of neural networks. Advances in Neural Information Processing Systems (NeurIPS), 2019. 3

[52] 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

[53] 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, 7, 15

[54] Junghoon Seo, Jeongyeol Choe, Jamyoung Koo, Seunghyeon Jeon, Beomsu Kim, and Taegyun Jeon. Noise-adding methods of saliency map as series of higher order partial derivative. In Workshop on Human Interpretability in Machine Learning, Proceedings of the International Conference on Machine Learning (ICML), 2018. 7

[55] 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. 1, 15

[56] 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), 2013. 1, 7, 14

[57] 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. 2, 16

[58] Gagandeep Singh, Timon Gehr, Markus Püschel, and Martin Vechev. An abstract domain for certifying neural networks. Proceedings of the ACM on Programming Languages, 2019. 3

[59] Dylan Slack, Anna Hilgard, Himabindu Lakkaraju, and Sameer Singh. Counterfactual explanations can be manipulated. Advances in Neural Information Processing Systems (NeurIPS), 2021. 2

[60] Dylan Slack, Anna Hilgard, Sameer Singh, and Himabindu Lakkaraju. Reliable post hoc explanations: Modeling uncertainty in explainability. Advances in Neural Information Processing Systems (NeurIPS), 34, 2021. 2

[61] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. In Workshop on Visualization for Deep Learning, Proceedings of the International Conference on Machine Learning (ICML), 2017. 1, 2, 7, 15

[62] Matthew Sotoudeh and Aditya V. Thakur. Computing linear restrictions of neural networks. In Advances in Neural Information Processing Systems (NeurIPS), 2019. 15

[63] 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. 2

[64] Pascal Sturmfels, Scott Lundberg, and Su-In Lee. Visualizing the impact of feature attribution baselines. Distill, 2020. 1, 2, 3, 7

[65] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In Proceedings of the International Conference on Machine Learning (ICML), 2017. 1, 2, 7, 15

[66] Vincent Tjeng and Russ Tedrake. Verifying neural networks with mixed integer programming. Proceedings of the International Conference on Learning Representations (ICLR), 15, 2019. 3

[67] Shiqi Wang, Huan Zhang, Kaidi Xu, Xue Lin, Suman Jana, Cho-Jui Hsieh, and J Zico Kolter. Beta-crown: Efficient bound propagation with per-neuron split constraints for neural network robustness verification. Advances in Neural Information Processing Systems (NeurIPS), 2021. 3

[68] Kaidi Xu, Zhouxing Shi, Huan Zhang, Yihan Wang, Kai-Wei Chang, Minlie Huang, Bhavya Kailkhura, Xue Lin, and Cho-Jui Hsieh. Automatic perturbation analysis for scalable certified robustness and beyond. Advances in Neural Information Processing Systems (NeurIPS), 2020. 3

[69] Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Sai Suggala, David I. Inouye, and Pradeep Ravikumar. On the (in)fidelity and sensitivity for explanations. In Advances in Neural Information Processing Systems (NeurIPS), 2019. 14

[70] Fan Yin, Zhouxing Shi, Cho-Jui Hsieh, and Kai-Wei Chang. On the sensitivity and stability of model interpretations in nlp. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2631–2647, 2022. 2

[71] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In Proceedings of the IEEE European Conference on Computer Vision (ECCV), 2014. 1, 15

[72] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In Proceedings of the IEEE European Conference on Computer Vision (ECCV), 2014. 2

[73] Huan Zhang, Tsui-Wei Weng, Pin-Yu Chen, Cho-Jui Hsieh, and Luca Daniel. Efficient neural network robustness certification with general activation functions. Advances in Neural Information Processing Systems (NeurIPS), 2018. 3