Evaluations and Methods for Explanation through Robustness Analysis

Cheng-Yu Hsieh*[1], Chih-Kuan Yeh†[3], Xuanqing Liu‡[2], Pradeep Ravikumar§[3], Seungyeon Kim¶[4], Sanjiv Kumar‖[4], and Cho-Jui Hsieh**[2]

[1]Department of CSIE, National Taiwan University
[2]Department of Computer Science, UCLA
[3]Machine Learning Department, Carnegie Mellon University
[4]Google Research

Abstract

Among multiple ways of interpreting a machine learning model, measuring the importance of a set of features tied to a prediction is probably one of the most intuitive ways to explain a model. In this paper, we establish the link between a set of features to a prediction with a new evaluation criterion, robustness analysis, which measures the minimum distortion distance of adversarial perturbation. By measuring the tolerance level for an adversarial attack, we can extract a set of features that provides the most robust support for a prediction, and also can extract a set of features that contrasts the current prediction to a target class by setting a targeted adversarial attack. By applying this methodology to various prediction tasks across multiple domains, we observe the derived explanations are indeed capturing the significant feature set qualitatively and quantitatively.

1 Introduction

There is an increasing interest in machine learning models to be credible, fair, and more generally *interpretable* [13]. Researchers have explored various notions of model interpretability, ranging from trustability [30], fairness of a model [48], to characterizing the model's weak points [22, 42]. Even though

*chyu.hsieh@gmail.com
†cjyeh@cs.cmu.edu
‡xqliu@cs.ucla.edu
§pradeepr@cs.cmu.edu
¶seungyeonk@google.com
‖sanjivk@google.com
**chohsieh@cs.ucla.edu

the goals of these various model interpretability tasks vary, the vast majority of them use so called feature-based explanation, that assign importances to individual features.

There have also been a slew of recent *evaluation* measures for feature based explanations, such as completeness [36], sensitivity-n [2], infidelity [43], causal local explanation metric [29], and most relevant to the current paper, smallest sufficient region (SSR) and smallest destroying region (SDR) [33, 16, 10]. A common thread in all these evaluation measures is quantifying how close the sum of feature importances approximate the difference in function value after removing the set of features. Intuitively, for a good feature based explanation, removing the most salient features should lead to a large difference in prediction score.

One key caveat with the aforementioned evaluations of feature explanations is the bias that arises in the way they operationalize "removing features," which is typically by setting them to some arbitrary reference value. The choice of these reference values inherently introduces some bias. For example, if we set the feature value to 0 in RGB images, this introduces a bias favoring bright pixels: explanations that optimize such evaluations often omit important dark objects, which could constitute *pertinent negative* features in the image, that do not contain the object but where the absence of the object is crucial to the prediction [12]. An alternative approach to "remove features" is to sample from some predefined distribution or a generative model [7]. This in turn incurs the bias inherent to the generative model, and accurate generative models that approximate the data distribution well might not be available in all domains.

In this paper, we take a slightly different perspective, focusing on small but adversarial perturbations rather than removal of features or large perturbations to reference values. Such "minimum adversarial perturbation" is typically used in the context of test-time robustness [18, 39], but which we harness towards feature based explanations. The key idea behind doing so is that adversarial perturbations on irrelevant features should be ineffective, while only those on relevant features should be effective. Thus by quantifying the effectiveness of adversarial perturbations restricted to a feature subset, we can in turn evaluate any feature based explanations. While exactly computing such an effectiveness measure is NP-hard [21], we can leverage recent results from test-time robustness literature [6, 26] which show that perturbations computed by adversarial attacks can serve as reasonably tight upper bounds, leading to an efficient approximation for the proposed evaluation.

Given this adversarial effectiveness evaluation measure, we can also design feature based explanations that optimize this evaluation measure. Note that designing such optimal explanations can also be cast as a two-player min-max game between an explainer and adversarial attacker. The explainer aims to find a set of important features, while the adversarial attacker aims to find a perturbation over the irrelevant features that changes the model prediction,

with the dueling goals of the attacker aiming to find the smallest perturbation, and the explainer aiming to ensure the perturbation is as large as possible. As we show, the resulting explanations empirically perform much better than previous approaches both quantitatively, as well as with qualitatively convincing examples.

To summarize our contributions:

- We define new evaluation criteria for feature-based explanations based on robustness analysis involving small adversarial perturbations. These reduce the bias inherent in other recent evaluation measures that focus on "removing features" via large perturbations to some reference values.
- We design efficient algorithms to generate explanations that maximize the proposed criteria, which perform favorably against baseline methods on the proposed evaluation criteria.
- Experiments in computer vision and NLP models demonstrate that the proposed explanation can identify important features that are not captured by previous methods. An additional facet of our approach is that it is able to extract a "contrast important" set of features that specifically contrast why the model makes its current prediction instead of a target class.

2 Robustness Analysis for Evaluating Explanations

2.1 Problem Notation

Let us consider the following setting: a general K-way classification problem with input space $\mathcal{X} \subseteq \mathbb{R}^d$, output space $\mathcal{Y} = \{1, \ldots, K\}$, and a predictor function $f : \mathcal{X} \to \mathcal{Y}$ where $f(\boldsymbol{x})$ denotes the output class for some input example $\boldsymbol{x} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d] \in \mathcal{X}$. Then, for a particular prediction $f(\boldsymbol{x}) = y$, despite the different forms of existing feature-based explanations ranging from attributing an importance value to each feature, ranking the features by their importance, to simply identify a set of important features, a common goal of them is to extract a compact set of relevant features with respect to the prediction.

2.2 Evaluation through Robustness Analysis

A common thread underlying evaluations of feature based explanations, even ranging over axiomatic treatments [36, 25], is that the importance of a set of features corresponds to the change in prediction of the model when the features are removed from the original input. Nevertheless, as we discussed in the previous section, operationalizing such a removal of features, for instance, by setting them to some reference value, introduces biases. To finesse this, we leverage adversarial robustness, but to do so in this context, we rely on two key assumptions:

Assumption 1: When the values of the important features are anchored (fixed), perturbations restricted to the complementary set of features has a weaker influence on the model prediction.

Assumption 2: When perturbations are restricted to the set of important features, fixing the values of the rest of the features, even small perturbations could easily change the model prediction.

Based on these two assumptions, we propose a new framework based on adversarial robustness for evaluating feature based explanations.

Definition 2.1 *Given a set of features S, its minimum adversarial perturbation norm, which we will also term Robustness-S is defined as:*

$$\epsilon^*_S = g(\boldsymbol{x}, S) = \left\{ \min_{\boldsymbol{\delta}} \|\boldsymbol{\delta}\|_p \; s.t. \; f(\boldsymbol{x}+\boldsymbol{\delta}) \neq y, \; \boldsymbol{\delta}_{\overline{S}} = 0 \right\}, \tag{1}$$

where $\overline{S} = U \setminus S$ is the complementary set of features, and $\boldsymbol{\delta}_{\overline{S}} = 0$ means that the perturbation is constrained to be zero along features in $\overline{S}$.

Suppose that the feature based explanation partitions the input features into a relevant set S_r, and an irrelevant set $\overline{S_r}$, Assumption 1 implies that the quality of the relevant set can be measured by $\epsilon^*_{\overline{S_r}}$ – measuring adversarial robustness to perturbations that keep the relevant set unchanged, but perturb only the irrelevant set. Specifically, from Assumption 1, a larger coverage of pertinent features in set S_r entails a higher robustness value $\epsilon^*_{\overline{S_r}}$. On the other hand, from Assumption 2, such a coverage of pertinent features in set S_r would in turn entail a smaller robustness value $\epsilon^*_{S_r}$, which measures the magnitude of adversarial perturbations restricted to the relevant set. Therefore, Assumptions 1 and 2 together build up our twin proposed evaluation criteria: *Robustness-$\overline{S_r}$* and *Robustness-S_r*.

To summarize:

Robustness-$\overline{S_r}$ measures the minimum adversarial perturbation $\epsilon_{\overline{S_r}}$ when the set of important features S_r, typically represented by the high-weight features in a feature importance map, are anchored, and perturbations are only allowed in the low-weight regions. The higher the score the better the explanation.

Robustness-S_r measures the minimum adversarial perturbation ϵ_{S_r} when only the set of important features S_r are can be perturbed, and the rest of the feature values are anchored. Contrary to the above, lower scores on this metric indicates a better explanation.

Specifying the sets S_r. To measure Robustness-$\overline{S_r}$, as well as and Robustness-S_r, we would need to first determine the sets S_r. Given any feature attribute method that assigns weights to each feature, once we have the size K of the

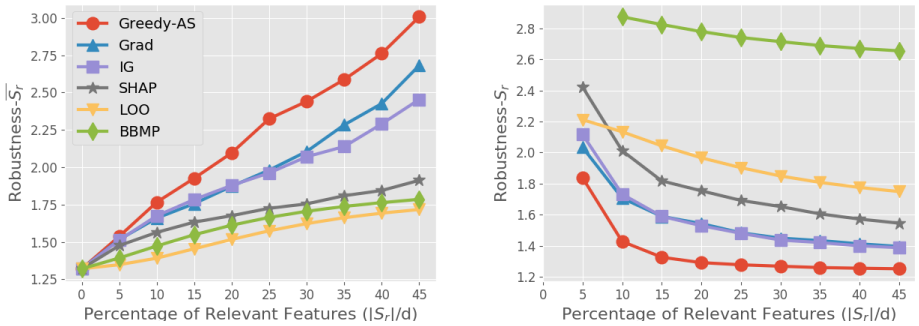

Figure 1: Evaluation curves for different methods under Robustness-$\overline{S_r}$ (left) and Robustness-S_r (right) with varying size of $|S_r|$. For Robustness-$\overline{S_r}$ (left), the higher the better; for Robustness-S_r (right), the lower the better. Note that we could calculate the area under the curves for each method to summarize its performance. We omit points in the plot with value that is too high to fit in the scale of y-axis.

set of important features, we can sort the features in descending order of important weights, and provide the top-K features. We thus largely need to specify the size of the sets $|S_r|$. We can set $|S_r|$ to the amount of anchors that an user is interested in or we may vary the size of $|S_r|$ and evaluate the corresponding values of Robustness-$\overline{S_r}$ and Robustness-S_r at different points. By varying the size of $|S_r|$, we could plot an evaluation curve for each explanation and in turn measure the area under curve (AUC), which corresponds to the average Robustness-$\overline{S_r}$ and Robustness-S_r at different sizes of relevant set. A larger (smaller) area under curve indicates a better feature attribution ranking. (See examples in Figure 1).

Untargeted v.s. Targeted Explanation. Definition 2.1 corresponds to the untargeted adversarial robustness – a perturbation that changes the predicted class to any label other than y is considered as a successful attack. Our formulation can also be extended to **targeted adversarial robustness**, where we replace (1) by

$$\epsilon^*_{S,t} = \left\{ \min_{\boldsymbol{\delta}} \|\boldsymbol{\delta}\|_p \text{ s.t. } f(\boldsymbol{x}+\boldsymbol{\delta}) = t; \boldsymbol{\delta}_{\overline{S}} = 0 \right\}, \tag{2}$$

where t is the targeted class. Using this definition, our approach will try to address the question "Why is this example classified as y instead of t", and the important features that optimize this criterion will highlight the contrast between class y and t. We will give examples of the "targeted explanations" in the experiment section.

Robustness Evaluation under Fixed Anchor Set. It is known that computing the exact minimum distortion distance in modern neural networks is intractable [21], so many different methods have been developed to estimate the value. Adversarial attacks, such as C&W [6] and PGD attack [26], aim to find a feasible solution of (1), which leads to an upper bound of ϵ_S^*. They are based on gradient based optimizers which are usually efficient. On the other hand, neural network verification methods aim to provide a lower bound of ϵ_S^* to ensure that the model prediction will not change within certain perturbation range [35, 40, 38, 17, 45, 37, 46]. However, these methods are usually time consuming (often > 50 times slower than a backpropagation).

The proposed framework can be combined with any method that aims to approximately compute (1), including attack, verification, and some other statistical estimations. However, for simplicity we only choose to evaluate (1) by the state-of-the-art projected gradient descent (PGD) attack [26], since the verification methods are too slow and often lead to much looser estimation as reported in some recent studies [32].

3 Extracting Model Supports through Robustness Analysis

Our adversarial robustness based evaluation allows us to evaluate any given feature based explanation. Here, we set out to design new explanations that explicitly optimize our evaluation measure. We focus on feature set based explanations, where we aim to provide an important subset of features S_r. Given our proposed evaluation measure, an optimal subset of feature S_r would aim to maximize (minimize) Robustness-$\bar{S}_r$ (Robustness-S_r), under a cardinality constraint on the feature set, leading to the following set of optimization problems:

$$\underset{S_r \in \{0,1\}^d}{\text{maximize}} \quad g(\boldsymbol{x}, \overline{S_r}) \quad \text{s.t.} \quad \|S_r\|_0 \leq K \tag{3}$$

$$\underset{S_r \in \{0,1\}^d}{\text{minimize}} \quad g(\boldsymbol{x}, S_r) \quad \text{s.t.} \quad \|S_r\|_0 \leq K \tag{4}$$

where K is a pre-defined size constraint on the set S_r, and $g(\boldsymbol{x}, S)$ computes the the minimum adversarial perturbation from Eqn. (1), with set-restricted perturbations.

It can be seen that this sets up an adversarial min-max game: the goal of the feature set explainer is to come up with a set S_r such that the minimal adversarial perturbation is as large as possible, while the adversarial attacker, given a set S_r, aims to design adversarial perturbations that are as small as possible. Directly solving these min-max problems in (3) and (4) is thus challenging, which is exacerbated by the discrete input constraint makes it intractable to find the optimal solution. As a result, in the next section, we propose a greedy algorithm, to estimate the optimal explanation sets.

3.1 Greedy Algorithm to Compute Optimal Explanations

We first consider a greedy algorithm where we iteratively add the most promising feature into S_r that optimizes the objective at each local step until S_r reaches the size constraint. In other words, we initialize the set S_r as empty, and sequentially solve the following subproblem at every step t:

$$\arg\max_i \; g(\boldsymbol{x}, \overline{S_r^t \cup i}), \text{ or } \arg\min_i \; g(\boldsymbol{x}, S_r^t \cup i), \; \forall i \in \overline{S_r^t} \tag{5}$$

where S_r^t is the anchor set at step t, and $S_r^0 = \emptyset$. We repeat this subprocedure until the size of set S_r^t reaches K. We name this method as *Greedy*. A straightforward way for solving (5) is to exhaustively search over every single feature.

3.2 Greedy by Set Aggregation Score

The main downside of using the greedy algorithm to optimize the objective function is that it ignores the interactions among features. Two features may perform bad when evaluated separately but become useful when added simultaneously. Therefore, in each greedy step, instead of considering how each individual feature will contribute to the objective, we propose to choose features based on its expected performance when evaluated with other unchosen features. To measure such aggregation score, we randomly choose sets of features and evaluate the performance of the objective function when the sets of features are added. Then we learn a regression function to distribute the performance of each set to each individual feature. Mathematically, let S_r^t and $\overline{S_r^t}$ be the ordered set of chosen and unchosen features at step t respectively, $\mathcal{P}^t(\overline{S_r^t})$ be all possible subsets of $\overline{S_r^t}$. We measure the expected contribution of including each unchosen feature to the relevant set would have on objective function by learning the following regression problem:

$$\boldsymbol{w}^t = \arg\min_{\boldsymbol{w}} \sum_{A \in \mathcal{P}^t(\overline{S_r^t})} ((\boldsymbol{w}^T B(A) + b) - g(\boldsymbol{x}, S_r^t \cup A))^2 \tag{6}$$

where B is a function that projects a set into its corresponding binary vector form: $B(A)[j] = \mathbb{I}(\overline{S_r^t}[j] \in A)$, i.e., ones in the vector indicate the inclusion of corresponding feature indices in the set and zeros otherwise. After the regression is learned, we can treat the coefficients $\boldsymbol{w}$ as each corresponding feature's approximated contribution to the objective value when they are included into the set S_r.

We note that $\boldsymbol{w}^t$ corresponds to the well-known Banzhaf value [4] when $S_r^t = \emptyset$, which is an axiomatic way to aggregate the importance of each player taking coalitions of players into account [14]. Hammer and Holzman [20] shows that Banzhaf value is equivalent to the optimal solution of a linear regression with pseudo-Boolean functions as targets, which corresponds to

Table 1: Area under curve of the proposed criteria for various explanations on MNIST and ImageNet. The higher the better for Robustness-$\overline{S_r}$; the lower the better for Robustness-S_r. Robustness measured with (1).

Datasets	Explanations	Grad	IG	SHAP	LOO	BBMP	Greedy-AS	Greedy	One-Step Banzhaf
MNIST	Robustness-$\overline{S_r}$	88.00	85.98	75.48	76.59	81.31	**98.01**	83.57	86.37
	Robustness-S_r	91.72	91.97	101.49	98.82	173.90	**82.81**	171.56	83.59
ImageNet	Robustness-$\overline{S_r}$	27.13	26.01	18.25	23.54	22.60	**31.62**	21.16	24.54
	Robustness-S_r	45.53	46.28	60.02	52.77	154.14	**43.97**	58.45	47.07

(6) with $S_r^t = \emptyset$. Banzhaf value can be interpreted as the importance of each player by taking coalitions of features into account. In each greedy step, we choose features with the highest aggregation score (Banzhaf value), which additionally considers the feature interactions between unchosen features compared to vanilla greedy. The chosen features each step are added to S_r^t and removed from $\overline{S_r^t}$. When S_r^t is not $\emptyset$, the solution of (6) can still be seen as Banzhaf value where the players are those features that are in $\overline{S_r^t}$, and the value function includes the features that are in S_r^t . We solve (6) by subsampling to lower the computational cost. We validate the effectiveness of greedy with aggregation score (Greedy-AS) in the experiment section. [1]

4 Experiments

In this section, we first evaluate different model interpretability methods on the proposed criteria. We justify the effectiveness of the proposed Greedy-AS. We then move onto further validating the benefits of the explanations extracted by Greedy-AS through comparisons to various existing methods both quantitatively and qualitatively. Finally, we demonstrate the flexibility of our method with the ability to provide targeted explanations as mentioned in Section 2.2. We perform the experiments on two image datasets, MNIST [23] and ImageNet [11], as well as a text classification dataset, Yahoo! Answers [47].

Setup. In the experiments, we consider $p = 2$ for $\|\cdot\|_p$ in (1) and (2), i.e., the ℓ_2 norm if not otherwise specified. For all quantitative results including the evaluation curves and the corresponding AUCs, we report the average over 100 random examples. For the baseline methods, we include vanilla gradient (Grad) [34] and integrated gradient (IG) [36] from gradient-based approaches; leave-one-out (LOO), or occlusion-1, [44, 24] and SHAP [25] from perturbation-based approaches [2]; and black-box meaningful perturbation (BBMP) [16] from SSR/SDR-based approaches. For the proposed Greedy

[1] We found that in parallel to our work, greedy with choosing the players with the highest restricted Banzhaf was used in Elkind et al. [15].

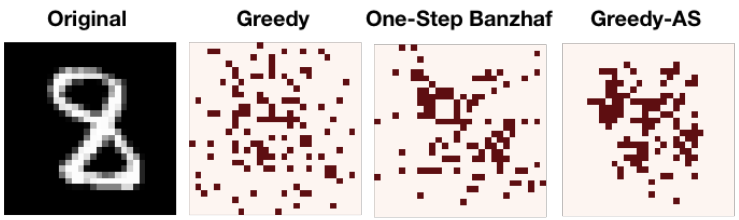

Figure 2: Visualization on our proposed methods. The top features selected by Greedy-AS are less noisy.

and Greedy-AS, at each greedy iteration, we include the top-5% features with highest scores into the relevant set to further speed up the selection process. We leave more implementation detail in Appendix A due to space limitation.

4.1 Robustness Analysis on Model Interpretability Methods

Here we analyze Greedy-AS as well as various existing explanation methods under both the proposed criteria Robustness-$\overline{S_r}$ and Robustness-S_r. For ease of comparison, we calculate the area under curve (AUC) for each corresponding evaluation curves. We list the results in Table 1, and leave the plots in Appendix B.

Ablation Study on Greedy-AS. As discussed in Section 3.2, Greedy-AS could be seen as a combination of the original greedy procedure with the approximated contribution of each feature computed by a regression. Here, we examine the importance of both components by comparing the Greedy-AS method to two baselines, where one selects important features based only on the pure Greedy method (Section 3.1) and the other utilizes only a single step of regression without the iterative greedy procedure. As the latter essentially corresponds to the Banzhaf value, we term this method as *One-Step Banzhaf*. First, as shown in Table 1, the pure Greedy method suffers degraded performances comparing to Greedy-AS under both criteria. The inferior performance could be explained by the ignorance of feature correlations which ultimately results in the introduction of noise, as shown in Figure 2. In addition, we also see that Greedy-AS performs better than One-Step Banzhaf. This could results from the fact that One-Step Banzhaf considers the feature interactions among all features with equal probability. However, in our objective, we only care about those interactions with the most important features. By iteratively selecting the features with highest Banzhaf value in Greedy-AS, we give more weight on the interactions among the most important features through iterations, and as a result lead to better performance.

Table 2: AUC of the Insertion and Deletion criteria for various explanations on MNIST. The higher the better for Insertion; the lower the better for Deletion.

Datasets	Explanations	Grad	IG	SHAP	LOO	BBMP	Greedy-AS
MNIST	Insertion	250.81	262.74	200.50	192.44	102.53	**379.15**
	Deletion	281.88	273.71	362.68	442.65	527.80	**159.77**
ImageNet	Insertion	54.66	78.64	1.87	32.21	103.51	**152.31**
	Deletion	267.37	245.67	**204.38**	280.40	603.89	247.20

Comparisons between Different Explanations. Furthermore from Table 1, we observe that the proposed Greedy-AS consistently outperforms other explanation methods on both criteria. On one hand, this suggests that the proposed algorithm indeed successfully optimizes towards the criteria; on the other hand, this might indicate the proposed criteria do capture different characteristics of explanations which most of the current explanations do not possess. Another somewhat interesting finding from the table is that while vanilla gradient has generally been viewed as a baseline method, it nonetheless performs competitively on the proposed criteria. We conjecture the phenomenon results from the fact that Grad does not assume any reference value as opposed to other baselines such as LOO which sets the reference value as zero to mask out the inputs. Indeed, it might not be surprising that Greedy-AS achieves the best performances on the proposed criteria since it is explicitly designed for so. To more objectively evaluate the usefulness of the proposed explanation, we demonstrate different advantages of our method by comparing Greedy-AS to other explanations quantitatively on existing commonly adopted measurements, and qualitatively through visualization in the following subsections.

4.2 Evaluating Greedy-AS

The Insertion and Deletion Metric. To further justify the proposed explanation not only performs well on the very metric it optimizes, we evaluate our method on a suite of existing popular quantitative measurements. In particular, we adopt the *Deletion* and *Insertion* criteria proposed by [28], which are generalized variants of the *region perturbation* criterion presented in [33]. The Deletion criterion measures the probability drop in the predicted class as top-relevant features, indicated by the given explanation, are progressively removed from the input. On the other hand, the Insertion criterion measures the increase in probability of the predicted class as top-relevant features are gradually revealed from the input whose features are originally all masked. Similar to our proposed criteria, a quick drop (and thus a small area under curve) or a sharp increase (that leads to a large area under curve)

Table 3: Rank correlation between explanations with respect to original and randomized model.

	Grad	IG	SHAP	LOO	BBMP	Greedy-AS
Corr.	0.30	0.30	0.11	0.49	0.17	0.18

in Deletion and Insertion respectively suggest a good explanation as the selected top-important features could indeed greatly influence the prediction. In the experiments, we follow [33] to remove features by setting their values to randomly sampled values. We plot the evaluation curves (in Appendix C) and report corresponding AUCs in Table 2. On these additional two criteria, we observe that Greedy-AS performs favorably against other explanations. The results further validate the benefits of the proposed explanation. We note that on ImageNet, SHAP obtains a better performance under the Deletion criterion. We however suspect such performance comes from the adversarial artifacts instead of meaningful explanation, since the explanation provided by SHAP seems to be rather noisy (as shown in Figure 4). [2] This also explains its relatively low performance under the Insertion criterion.

Sanity Check Metric. Recent literature has pointed out that an appropriate explanation should be related to the model being explained [1]. To ensure that our proposed explanation does indeed reflect the model behavior, we conduct the sanity check proposed by [1] to check if our explanations are adequately different when the model parameters are randomly re-initialized. In the experiment, we randomly re-initialize the last fully-connected layer of the neural network model. We then compute the rank correlation between explanation computed w.r.t. the original model and that w.r.t. the randomized model. From Table 3, we observe that Greedy-AS has a much lower rank correlation comparing to Grad, IG, and LOO, suggesting that Greedy-AS is indeed sensitive to model parameter change and is able to pass the sanity check.

4.3 Qualitative Results

Image Classification. To complement the quantitative measurements, we show several visualization results on MNIST and ImageNet in Figure 3 and Figure 4. More examples could be found in Appendix E and F. On MNIST, we observe that existing explanations tend to highlight mainly on the white pixels in the digits; among which SHAP and LOO show less noisy explanations comparing to Grad and IG. On the other hand, the proposed Greedy-AS focuses on both the "crucial positive" (important white pixels) as well as

[2] It has been observed that the Deletion criterion tends to favor adversarial artifacts in several previous work. [10, 7]

the "pertinent negative" (important black regions) that together support the prediction. For example, in the first row, a 7 might have been predicted as a 4 or 0 if the pixels highlighted by Greedy-AS are set to white. Similarly, a 1 may be turned to a 4 or a 7 given additional white pixels to its left, and a 9 may become a 7 if deleted the lower circular part of its head. From the results, we see that Greedy-AS focuses on *"the region where perturbation on its current value will lead to easier prediction change"*, which includes both the crucial positive and pertinent negative pixels. Such capability of Greedy-AS is also validated by its superior performance on the proposed robustness criteria, on which methods like LOO that highlights only the white strokes of digits show relatively low performance. From the visualized ImageNet examples shown in Figure 4, we observe that our method provides more compact explanations that focus mainly on the actual objects being classified. As opposed to methods that show noisy explanations, Greedy-AS could potentially provide more insights into the model prediction.

Text Classification. In addition to image datasets, here we demonstrate how our explanation method could be applied to text classification models. In the experiments, we represent a length-n sentence by n embedding vectors following the common setting. Thus, when applying our Greedy algorithm, at each iteration we will try to add an embedding vector to the relevant set S_r and choose the one with largest reward. Since there are only at most n choices, the Greedy algorithm doesn't suffer much from noise and has similar behavior to Greedy-AS.

We perform experiments on an LSTM network which learns to classify a given sentence into one of the ten classes (Society, Science, Health, ...). We showcase an example with explanations generated with different methods in Figure 5. We note that although the top-5 relevant keyword sets generated by the three methods do not vary much, the rankings within the highlighted keywords for each explanation are in fact quite different. We observe that our method Greedy tends to generate explanation that matches human intuition the most. Particularly, to predict the label of "sport", one might consider "cleats", "football", and "cut" as the strongest indications towards the concept "sport".

Targeted Explanation Analysis. Recall that in section 2.2, we discussed about the possibility of defining the robustness measurement by considering a targeted distortion distance as formulated in (2). Here, we provide examples, as shown in Figure 6, where we answer the question of *"why the input digit is an A but not a B"* by defining a targeted perturbation distance towards class B as our robustness measurement. In each row of the figure, we provide targeted explanation towards two different target classes for a same input image. Interestingly, as the target class changes, the generated explanation

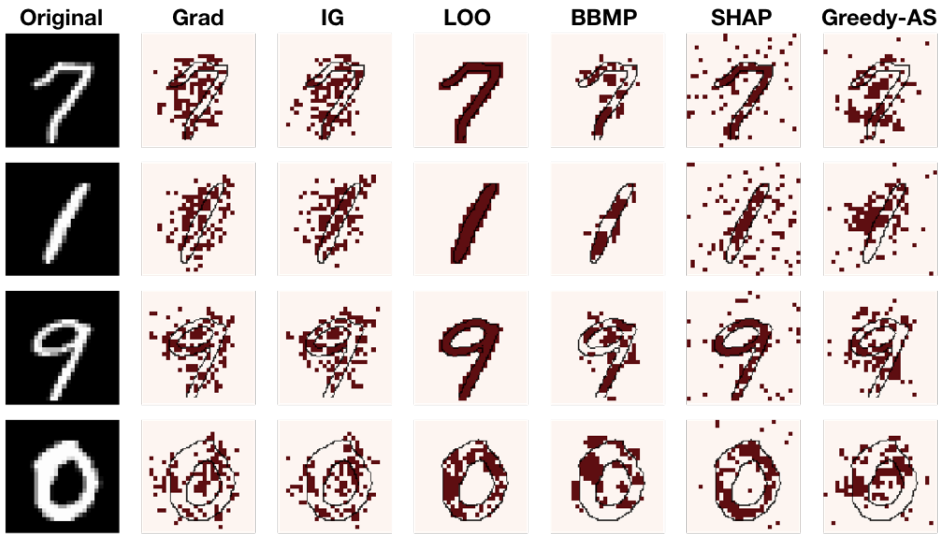

Figure 3: Visualization on top 20 percent relevant features provided by different explanations. We see Greedy-AS highlights both crucial positive and pertinent negative features supporting the prediction.

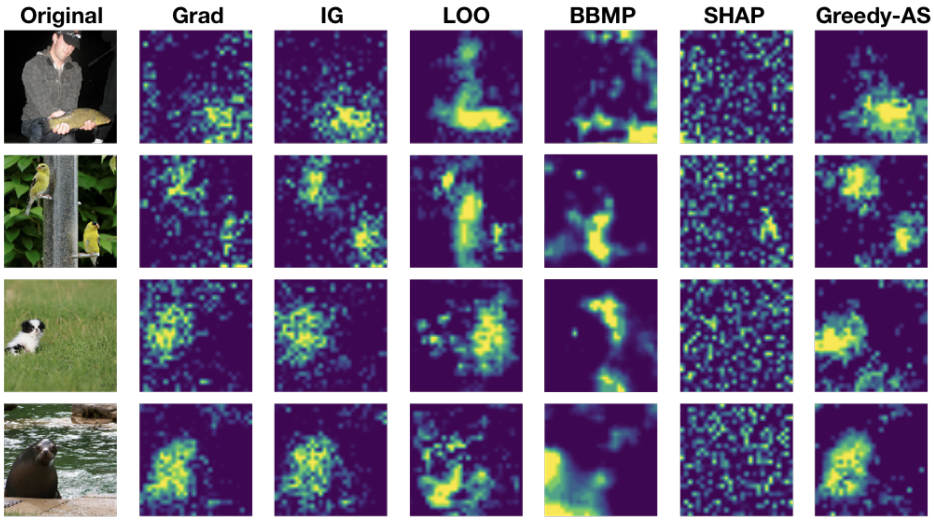

Figure 4: Visualization of different explanations on ImageNet, where the predicted class for each input is “fish", “bird", “dog", and “sea lion". Comparing to other methods, Greedy-AS focuses more on the areas that are essential to correctly classify the image.

varies in an interpretatble way. For example, in the first row, we explain why the input digit 7 is not classified as a 9 (middle column) or a 2 (rightmost column). The resulting explanation against 9 highlights the upper-left part

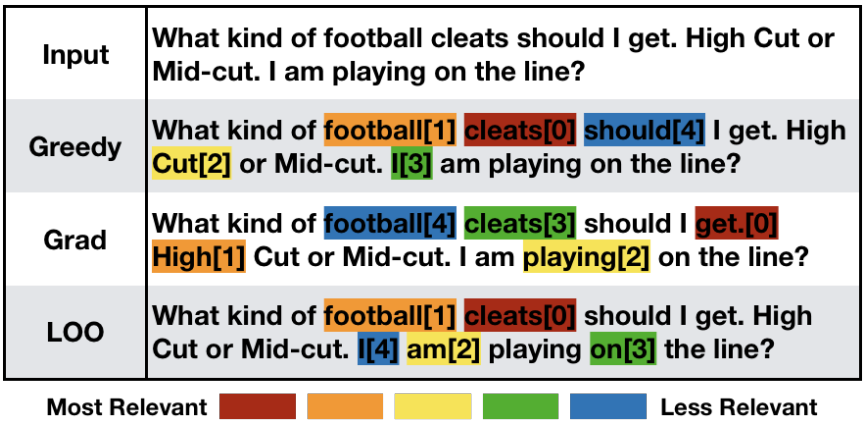

Figure 5: Explanations on a text classification model where the predicted label for this sentence is "sport". Unlike other methods, the top-3 relevant keywords highlighted by Greedy are all closely related to the concept "sport".

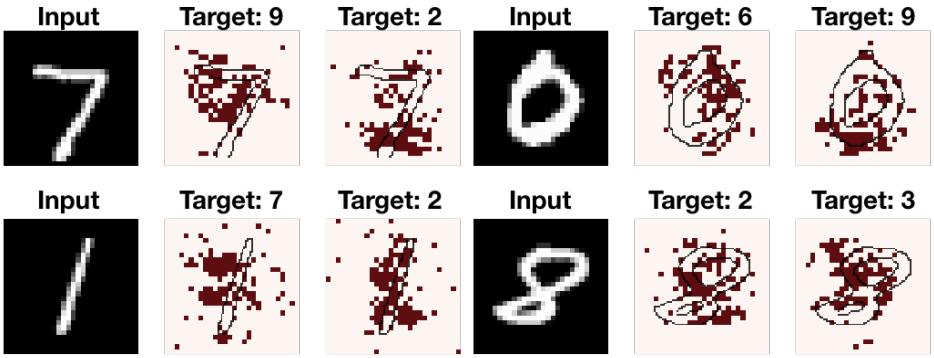

Figure 6: Visualization of targeted explanation. For each input, we highlight relevant regions explaining why the input is not predicted as the target class. We see the explanation changes in a semantically meaningful way as the target class changes.

of the 7. Semantically, this region is indeed pertinent to the classification between 7 and 9, since turning on the highlighted pixel values in the region (currently black in the original image) will then make the 7 resemble a 9. However, the targeted explanation against 2 highlights a very different but also meaningful region, which is the lower-right part of the 7; since adding a horizontal stroke on the area would turn a 7 into a 2.

The capability of capturing pertinent negative features has also been observed in explanations proposed in some recent work [12, 3, 27]. However, these methods are subject to different constraints. For example, [12] is designed to handle binary inputs which nonetheless limits it application; in

[3], the ability to capture pertinent negative features heavily depends on the input range; for [27], unlike our targeted explanation where we know exactly which targeted class the explanation is suggesting against. The pertinent negative features highlighted by their method by construction is not directly related to a specific target class, making it harder for users to interpret the result. We provide more detailed discussions and comparisons to these methods in Appendix G.

5 Related Work

Our work proposes an objective measurement of feature-based explanation by measuring the "minimum adversarial perturbation" in adversarial literature, which is estimated by adversarial attack. We provide a necessarily incomplete review on related works in objective measurement of explanations, adversarial robustness, as well as the intersection between the two.

Objective Measurements for Explanations. Evaluation of explanations has been a difficult problem mainly due to the absence of ground truth [2, 36]. Although one could rely on human intuitions to assess the quality of the generated explanations [25, 13], for example, judging whether the explanation focuses on the object of interest in an image classification task, these evaluations subject to human perceptions are prone to fall into the pitfall of favoring user-friendly explanations, such as attributions that visually aligns better with the input image, which might not reflect the model behavior [1]. As a result, in addition to subjective measurements, recent literature has also proposed objective measurements, which is also called functionally-grounded evaluations [13]. We roughly categorize existing objective measurements into two families.

This first family of explanation evaluation is called fidelity-based measurement. This includes that *Completeness* or *Sum to Delta* which requires the sum of attributions to equal the prediction difference of the original input and baseline [36, 34]; sensitivity-n which further generalizes completeness to any subset of the feature [2]; local accuracy [30, 25]; and *infidelity* which is a framework that encompasses several [43]. The general philosophy for this line of methods is to require the sum of attribution value faithfully reflect the change in prediction function value given the presence or absence of certain subset of features. The second family of explanation evaluation are removal-based and preservation-based measurements, which focus on identifying the most important set of features with respect to a particular prediction. The underlying assumption made is that by removing the most (least) salient feature, the resulting function value should drop (increase) the most. [33] proposed this idea as an evaluation to evaluate the ranking of feature-attribution score. Later on, [16] derive explanations by solving an

optimization problem to optimize the evaluation. And [10] proposed to learn the explanation generating process by training an auxiliary model.

Adversarial Robustness. Adversarial robustness has been extensively studied in the past few years. The adversarial robustness of a machine learning model on a given sample can be defined as the shortest distance from the sample to the decision boundary, which corresponds to our definition in 1. Algorithms have been proposed for finding adversarial examples (feasible solutions of 1), including [18, 6, 26]. However, those algorithms only work for neural networks, while for other models such as tree based models or nearest neighbor classifiers, adversarial examples can be found by decision based attacks [5, 9, 8]. Therefore the proposed framework can also be used in other decision based classifiers. On the other hand, several works aim to solve the neural network verification problem, which is equivalent to finding a lower bound of 1. Examples include [35, 40, 45]. In principal, our work can also apply these verification methods for getting an approximate solution of 1, but in practice they are very slow to run and often gives loose lower bounds on regular trained networks.

Interpretability and Adversarial Robustness Our work is closely related to recent studies that bridge the gap between model interpretability and adversarial robustness. Xu et al. [41] add group sparsity regularization to adversarial attack to enforce semantic structure for the perturbation, which is more interpretable. Ribeiro et al. [31] find a set of features that once fixed, probability of the prediction is high when perturbing other features. Several recent work has also considered the question "For situation A, why was the outcome B and not C", which we call counterfactual explanations. Goyal et al. [19] show how one could change the input feature such that the system would output a different class, where the change is limited to replacing a part of input feature by a part of an distractor image. Dhurandhar et al. [12] consider the pertinent negative in a binary setting by solving a carefully designed loss function.

6 Conclusion

In this paper, we establish the link between a set of features to a prediction with a new evaluation criteria, robustness analysis, which measures the minimum tolerance of adversarial perturbation. Furthermore, we develop a new explanation method to find important set of features to optimize this new criterion. Experimental results demonstrate that the proposed new explanations are indeed capturing significant feature sets across multiple domains.

References

[1] 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*, pages 9525–9536, 2018.

[2] Marco Ancona, Enea Ceolini, Cengiz Oztireli, and Markus Gross. A unified view of gradient-based attribution methods for deep neural networks. *International Conference on Learning Representations*, 2018.

[3] 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.

[4] John F Banzhaf III. Weighted voting doesn't work: A mathematical analysis. *Rutgers L. Rev.*, 19:317, 1964.

[5] Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *arXiv preprint arXiv:1712.04248*, 2017.

[6] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57. IEEE, 2017.

[7] Chun-Hao Chang, Elliot Creager, Anna Goldenberg, and David Duvenaud. Explaining image classifiers by counterfactual generation. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=B1MXz20cYQ.

[8] Hongge Chen, Huan Zhang, Duane Boning, and Cho-Jui Hsieh. Robust decision trees against adversarial examples. In *ICML*, 2019.

[9] Minhao Cheng, Thong Le, Pin-Yu Chen, Jinfeng Yi, Huan Zhang, and Cho-Jui Hsieh. Query-efficient hard-label black-box attack: An optimization-based approach. *arXiv preprint arXiv:1807.04457*, 2018.

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

[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[12] Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In *Advances in Neural Information Processing Systems*, pages 592–603, 2018.

[13] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[14] Pradeep Dubey and Lloyd S Shapley. Mathematical properties of the banzhaf power index. *Mathematics of Operations Research*, 4(2):99–131, 1979.

[15] Edith Elkind, Piotr Faliszewski, Martin Lackner, Dominik Peters, and Nimrod Talmon. Committee scoring rules, banzhaf values, and approximation algorithms. In *4th workshop on exploring beyond the worst case in computational social choice (EXPLORE'17)*, 2017.

[16] Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3449–3457, 2017.

[17] Timon Gehr, Matthew Mirman, Dana Drachsler-Cohen, Petar Tsankov, Swarat Chaudhuri, and Martin Vechev. Ai2: Safety and robustness certification of neural networks with abstract interpretation. In *2018 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE, 2018.

[18] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[19] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. In *International Conference on Machine Learning*, pages 2376–2384, 2019.

[20] Peter L Hammer and Ron Holzman. Approximations of pseudo-boolean functions; applications to game theory. *Zeitschrift für Operations Research*, 36(1):3–21, 1992.

[21] 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.

[22] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, pages 1885–1894, 2017.

[23] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann. lecun. com/exdb/mnist*, 2, 2010.

[24] Jiwei Li, Will Monroe, and Dan Jurafsky. Understanding neural networks through representation erasure. *CoRR*, abs/1612.08220, 2016. URL http://arxiv.org/abs/1612.08220.

[25] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774, 2017.

[26] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[27] Jose Oramas, Kaili Wang, and Tinne Tuytelaars. Visual explanation by interpretation: Improving visual feedback capabilities of deep neural networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=H1ziPjC5Fm.

[28] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018.

[29] Gregory Plumb, Denali Molitor, and Ameet S Talwalkar. Model agnostic supervised local explanations. In *Advances in Neural Information Processing Systems*, pages 2515–2524, 2018.

[30] 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*, pages 1135–1144. ACM, 2016.

[31] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[32] Hadi Salman, Greg Yang, Huan Zhang, Cho-Jui Hsieh, and Pengchuan Zhang. A convex relaxation barrier to tight robust verification of neural networks. *arXiv preprint arXiv:1902.08722*, 2019.

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

[34] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. *International Conference on Machine Learning*, 2017.

[35] Gagandeep Singh, Timon Gehr, Matthew Mirman, Markus Püschel, and Martin Vechev. Fast and effective robustness certification. In *Advances in Neural Information Processing Systems*, pages 10802–10813, 2018.

[36] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, 2017.

[37] Shiqi Wang, Kexin Pei, Justin Whitehouse, Junfeng Yang, and Suman Jana. Efficient formal safety analysis of neural networks. In *Advances in Neural Information Processing Systems*, pages 6367–6377, 2018.

[38] Tsui-Wei Weng, Huan Zhang, Hongge Chen, Zhao Song, Cho-Jui Hsieh, Luca Daniel, Duane Boning, and Inderjit Dhillon. Towards fast computation of certified robustness for relu networks. In *International Conference on Machine Learning*, pages 5273–5282, 2018.

[39] Tsui-Wei Weng, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, Dong Su, Yupeng Gao, Cho-Jui Hsieh, and Luca Daniel. Evaluating the robustness of neural networks: An extreme value theory approach. *arXiv preprint arXiv:1801.10578*, 2018.

[40] Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning*, pages 5283–5292, 2018.

[41] Kaidi Xu, Sijia Liu, Pu Zhao, Pin-Yu Chen, Huan Zhang, Quanfu Fan, Deniz Erdogmus, Yanzhi Wang, and Xue Lin. Structured adversarial attack: Towards general implementation and better interpretability. *arXiv preprint arXiv:1808.01664*, 2018.

[42] Chih-Kuan Yeh, Joon Kim, Ian En-Hsu Yen, and Pradeep K Ravikumar. Representer point selection for explaining deep neural networks. In *Advances in Neural Information Processing Systems*, pages 9291–9301, 2018.

[43] Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Sai Suggala, David I. Inouye, and Pradeep Ravikumar. On the (in)fidelity and sensitivity for explanations. *CoRR*, abs/1901.09392, 2019.

[44] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[45] Huan Zhang, Tsui-Wei Weng, Pin-Yu Chen, Cho-Jui Hsieh, and Luca Daniel. Efficient neural network robustness certification with general activation functions. In *Advances in neural information processing systems*, pages 4939–4948, 2018.

[46] Huan Zhang, Pengchuan Zhang, and Cho-Jui Hsieh. Recurjac: An efficient recursive algorithm for bounding jacobian matrix of neural

networks and its applications. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5757–5764, 2019.

[47] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657, 2015.

[48] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, 2017.