Building Reliable Explanations of Unreliable Neural Networks: Locally Smoothing Perspective of Model Interpretation

Dohun Lim Hyeonseok Lee Sungchan Kim*
Division of Computer Science and Engineering, Jeonbuk National University, Korea

Abstract

We present a novel method for reliably explaining the predictions of neural networks. We consider an explanation reliable if it identifies input features relevant to the model output by considering the input and the neighboring data points. Our method is built on top of the assumption of smooth landscape in a loss function of the model prediction: locally consistent loss and gradient profile. A theoretical analysis established in this study suggests that those locally smooth model explanations are learned using a batch of noisy copies of the input with the L1 regularization for a saliency map. Extensive experiments support the analysis results, revealing that the proposed saliency maps retrieve the original classes of adversarial examples crafted against both naturally and adversarially trained models, significantly outperforming previous methods. We further demonstrated that such good performance results from the learning capability of this method to identify input features that are truly relevant to the model output of the input and the neighboring data points, fulfilling the requirements of a reliable explanation.

1. Introduction test

The recent progress of deep neural networks has led to their adoption in various decision-critical applications, including medical, finance, and legal fields and autonomous vehicles. However, the high modeling capacity of deep models renders the inner operations of the models mostly uninterpretable and demands human-understandable explanations for model predictions. For this purpose, the model output attribution for input features is a popular idea. The attribution aims to identify the importance of input features using end-to-end relationships between inputs and model predictions. We use a *saliency map* as the visual form of an *explanation*. A saliency map is a common approach to visual tasks to implement pixel-level or regional attributions for a given image [32, 25, 9, 37, 16, 22].

*Correspondence to: Sungchan Kim (s.kim@jbnu.ac.kr).

The susceptibility of neural networks can cause false predictions for a given imperceptibly modified image [12], which has emerged as a new challenge in explaining model predictions [9, 3, 37, 10, 5, 14, 29, 33]. For instance, recent work has demonstrated that input can be manipulated, resulting in different saliency maps without damaging the classification accuracy [10, 5, 14, 33]. Such a false explanation is primarily due to the fragility of learned models that have highly nonsmooth decision boundaries rather than due to the explanation methods. Although adversarial training has addressed these concerns [17, 39, 24, 7, 11, 36, 19], it is not always applicable and still incomplete.

This discussion indicates the need to build a reliable model explanation with two requirements if the goal is *to recover input features that are important in a local neighborhood*: first, a reliable explanation method should be robust so that it generates consistent explanations along with neighboring (and thus similar) data points; second, explanations generated by the method must have high fidelity for model predictions.

We propose RelEx, a novel method to *reliably* explain predictions of neural network-based classifiers. RelEx aims to generate *robust* and yet *accurate* saliency maps of pixel-level importance for a given image. Inspired by recent work on adversarial training [18, 17, 19], we built RelEx on top of an assumption on a *locally smooth explanation* for the vicinity of the input.

Although substantial work has proposed creating saliency maps based on gradient [28, 34, 30, 1, 27, 25] or perturbation [20, 9, 37, 16, 22], researchers have hardly addressed both robustness and accuracy in a single method. Existing methods often fail to find out essential features of the input and neighborhood for the model predictions despite being visually plausible, as shown in this study. Moreover, using either a random or adversarial perturbation of data points can manipulate their explanations [10, 5]. Recent work has addressed such an issue partially [3, 9, 37, 5]. In contrast to the existing methods, we construct an analysis to characterize the behavior of the proposed explanation method based on the assumption of the local explanation. Specifically, the contributions of this paper are as follows.

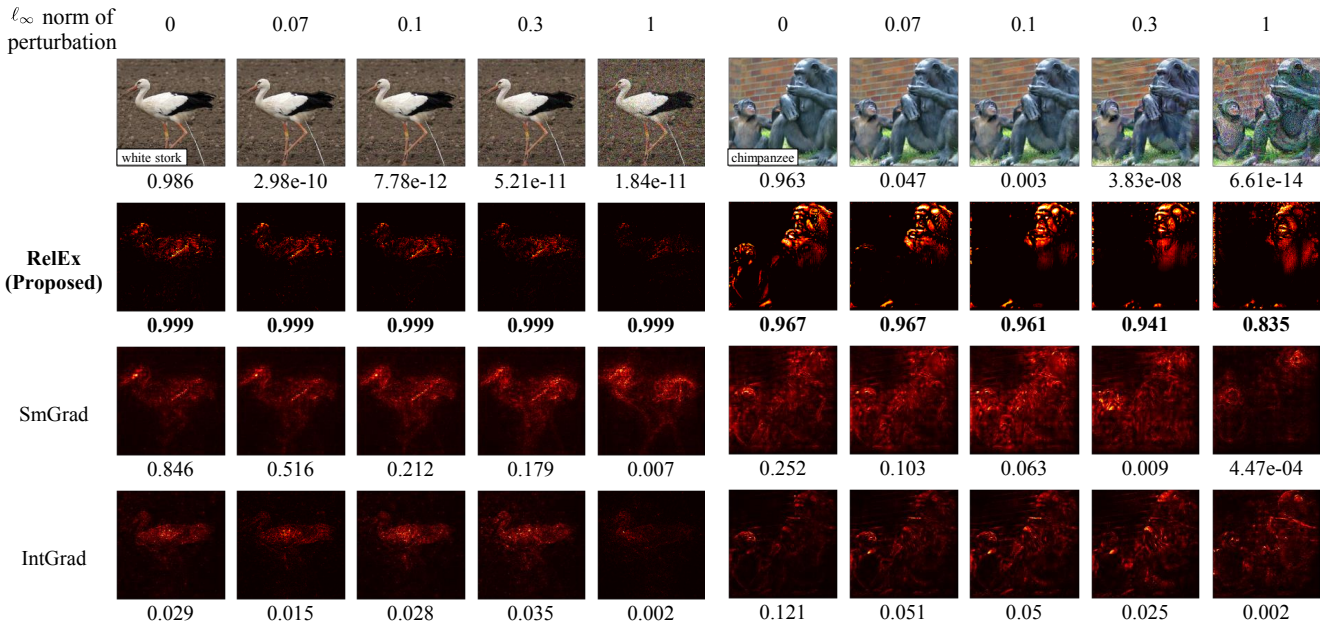


Figure 1. **Comparison of saliency maps.** Images on the top row depict adversarial examples created by a PGD attack [17] for given perturbation distance in the ℓ_1-norm on (a) naturally trained and (b) adversarially trained ResNet-50. Numbers below images represent its softmax scores while ones below saliency maps indicate the scores of corresponding explanations. Our method generates saliency maps that leads to the scores close to 1 consistently in the presence of perturbations. The saliency maps of other methods, SmGrad [30] and IntGrad [34], are visually plausible but irrelevant to the score. (See the supp. S3.7 for more results)

- We establish a quadratic approximation on a locally smooth landscape from the model explanation perspective to identify a trade-off between the accuracy and robustness of saliency maps. Our analysis reveals that the robustness of an explanation method is better achieved at the cost of the reduced accuracy of the model explanation and that the curvature of a loss function for learning a saliency map is inverse-proportional to the ℓ_1-norm of saliency maps to which, however, the explanation accuracy is proportional. A similar trace-off was investigated in adversarial training [36, 39]; but it was hardly addressed in the context of an explanation method. Although the use of the ℓ_1-norm is not new this work is the first attempt to address its effects on building a reliable explanation method to the best of our knowledge.

- Our analysis leads to an easy-to-implement objective function to learn a saliency map by using backpropagation. We need only noisy copies of an input image as a batch for the optimization that is regularized by the ℓ_1-norm of a saliency map.

- RelEx identifies input features relevant to a decision for all points in a neighborhood over the existing methods when applying it to naturally and adversarially trained models. We demonstrate that explanations by the proposed method achieved a remarkably robust retrieval of the target classes from adversarial examples created via strong white-box attacks (Figure 1). Extensive evaluations indicate that such an advantage our method is due to learning appropriate saliency maps even with severe perturbations.

2. Related Work

We briefly review previous work related to the explanation of neural networks to highlight the benefit of this approach. We first focus on methods to generate saliency maps and then on the literature on robust model explanations.

Gradient-based methods. Existing explanation methods generate the importance of input features primarily based on the gradient or perturbation of an input image. Gradient-based approaches measure individual feature importance as the sensitivity of input features concerning changes in the model prediction by using standard back-propagation [28, 30, 34, 1, 25]. The pixel-level gradient has inherent limitations, such as being noisy and saturation, to explain the model output [1, 30, 34, 27]. A method, called SmoothGrad, generates explanations by averaging gradient-based saliency maps of noisy copies of an input image [30], partially addressing adversarial attacks [14, 5]. An approach in [34] takes integrated gradients along with a straight path from the baseline value to each of the input features as an attribution of the particular feature. Layer-wise relevance propagation and DeepLIFT back-propagate the model out-

put by distributing it through a neural network according to neuronal activation [1, 27]. Although these principled approaches provide improved saliency maps, one can manipulate input images for the model to classify the images correctly but result in their saliency maps differently from original ones [10, 5, 14, 29, 33].

To summarize, because these methods merely react to the interactions of the model with the input and thus are unsupervised processes in nature, they are restricted to presenting the reaction of the model to the change.

Perturbation- and activation-based methods. Another approach generates saliency maps as a change in the model output caused by perturbing the input image [3, 9, 20, 37, 16, 22]. These methods learn feature importance of an input at the pixel-level [37] like the proposed or at the regional basis [3, 9, 20, 16, 22]. Perturbation, such as occlusion and masks, is queried to an image repeatedly to learn an optimal saliency map. Some incorporated regularizers for adversarial defense, unlike gradient-based methods [3, 9, 37]. We demonstrate that these defenses are insufficient and can be deceived by carefully crafted adversaries.

Activation-based approaches combine activations of convolutional layers linearly to translate the spatial information of features maps in different layers to saliency maps [25, 3, 22]. However, their saliency maps are diffused and are unreliable against the perturbation of input images.

Robust explanation. Recent work has indicated that the susceptibility of neural networks is caused by a high modeling capacity with numerous parameters and a kinky landscape of the gradient of the model output due to the nonlinearity of the models [10, 5]. The adversarial training of models incorporating robustness into the model during the training process is an active research area [23, 8, 21, 19]. Their main idea is to encourage robustness by enforcing a locally linear landscape of the model prediction for neighboring examples [26, 23, 19].

We regard the notion of a locally smoothing behavior as a requirement of a reliable explanation. Thus, explanations with saliency maps for the vicinity of an input should be similar and lead to the same class. The authors in [5] demonstrated that piece-wise nonlinearity due to a rectified linear unit (ReLU) of neural networks might mislead to a wrong explanation, even with a correct prediction. They proposed to use SoftPlus instead of the ReLU in the model, which is identical to using SmoothGrad [30], providing partial tolerance to an adversarial attack [5, 14]. An analysis established in [37] indicated that the robustness of the model is degraded by numerous parameters and the largest singular value of the Hessian regarding the parameters that represents the curvature of the gradient landscape. Despite previous efforts, no tangible realization of a robust explanation method has been addressed. Some effort has been made to improve the quality of saliency maps, considering adversarial defense [3, 9, 37, 5] by reducing a total variation [3, 9] or modifying the activation function of neurons [37, 5]. They, however, result in insufficient defense capability. RelEx exploits the analysis established in this work to address the limitations above efficiently and effectively, and non-intrusively through a simple optimization framework.

3. Proposed Approach

This section describes the details of RelEx. We begin by stating two assumptions that a reliable explanation method should satisfy. Then, we analyze the bounds on the robustness of the proposed method in response to these requirements in Section 3.1. A cost function of RelEx for learning a saliency map is formulated in Section 3.2.

Notations. A neural network-based classification model maps an input image $x_0 \in \mathbb{R}^d$ to an output $y \in [0,1]^{|C|}$, where y is a vector representing the softmax scores of a specified set of classes C. We define $f_c(x_0)$ as the probability of x_0 being classified as $c \in C$. A *saliency map* $m_c \in [0,1]^d$ represents the importance of individual features (i.e., pixels of x_0) corresponding to the model prediction $f_c(x_0)$. For simplicity, we use m and $f(x_0)$ instead of m_{c_T} and $f_{c_T}(x_0)$, respectively, when referring to the target class c_T of x_0.

3.1. Local Smoothness for Robust Explanation

Suppose a local interpretation of the model prediction refers to identify features relevant to x_0 and its local neighborhood. In that case, the interpretation implies that explanations of data samples neighboring to x_0 should vary slowly, rendering a corresponding smooth landscape although it holds locally. According to the notion of the local interpretation, we expect data points in the vicinity of x_0 to be in the same class with similar saliency maps. We claim that, in particular, two conditions should be met in the local data points for the model explanation to be reliable by ensuring the local smoothness.

Assumption 1 (Locally consistent model prediction). *For a given saliency map m of input x_0 with a target class c_T, we consider data samples $\mathcal{D} = \{x_i\}$ where $\|x_i - x_0\|_p \leqq \epsilon$ and ϵ is a small positive number, which we regard as the perturbation of x_0. Thus, $m \odot x_i$ for data points $x_i \in \mathcal{D}$ concerning m would be correctly classified as c_T (i.e., $f(m \odot x_i) \approx f(m \odot x_0)$), where $\odot$ is an element-wise product of vectors. We use the ℓ_2-norm or ℓ_∞-norm as the perturbation distance.*[1]

Assumption 2 (Locally consistent saliency maps). *A saliency map $m^{(x_i)}$ for data point $x_i \in \mathcal{D}$ is similar to that of x_0, m, which leads to $\|m - m^{(x_i)}\| \leqq \delta$ for a small positive number δ.*

[1] We represent the ℓ_2-norm of vector a as $\|a\|$.

Local smoothness for label consistency. We consider a simple analysis to elucidate how a saliency map is related to the label consistency of an explanation method. As the distance between the data points depends on the task, we use cross-entropy in this study. Thus, given an input x and a saliency map m, we denote a loss function of classifying $m \odot x$ as its target class c_T as $L(x, m) = -\log f(m \odot x)$. Inspired by the setting of the analysis established in [19], we assume that $L(\cdot)$ is well approximated as a quadratic form at a sufficiently small distance γ, which is given by the following:

$$L(x_0+\gamma, m) \approx L(x_0, m)+\nabla L(x_0, m)^T\gamma+\frac{1}{2}\gamma^T H\gamma \quad (1)$$

where $\nabla L(x_0, m)$ and H denote the gradient and Hessian of $L(\cdot)$ at x_0, respectively. The robustness of the explanation method is represented by γ if we ensure that all samples in the L_2 ball of radius γ centered at x_0 are classified correctly. The second-order derivative term in Eq. (1) enables the investigation of the divergent curvature of the loss function from the perspective of the optimization landscape.

If all data points in the L_2 ball are classified correctly, it holds that $L(x_0 + \gamma, m) \leqq \tau$ and $L(x_0, m) \leqq \tau$ where τ is the threshold for the input to be classified correctly. For example, in the case of binary classification, $\tau = -\log \frac{1}{2}$. We measure the robustness γ by evaluating its maximum γ^+ as follows:

$$\gamma^+ = \arg\max_{\gamma} \|\gamma\| \text{ s.t. } L(x_0 + \gamma, m) \leqq \tau. \quad (2)$$

For the ease of calculation, we convert Eq. (2) to an equivalent minimization problem as follows:

$$\gamma^+ = \arg\min_{\gamma} \|\gamma\| \text{ s.t. } L(x_0 + \gamma, m) \geqq \tau. \quad (3)$$

Let $c = \tau - L(x_0, m) \geqq 0$. Substituting c into Eq. (3) yields

$$\gamma^+ = \arg\min_{\gamma} \|\gamma\| \text{ s.t. } \nabla L(x_0, m)^T\gamma + \frac{1}{2}\gamma^T H\gamma \geqq c. \quad (4)$$

The introduction of Eq. (4) is not to calculate γ^+ but to derive the lower bound of the robustness and the upper bound of the loss function as a function of m. This analysis also holds when we use the ℓ_∞-norm for the perturbation distance metric.

Theorem 1 (Local explanations with respect to label consistency). *Let $\gamma = \alpha \cdot v$ where $\|\gamma\| = \alpha$ and $\|v\| = 1$. Let $\mathcal{D} = \{x_i\}$ be a set of data samples where $\|x_i - x_0\| \leqq \epsilon$. For a given saliency map m calculated from x_0, it holds that*

$$\alpha \geqq \frac{c}{\|m\|_1} \cdot \frac{2}{\| - g(x_0 + \alpha v) + g(x_0)\| + 2\|g(x_0)\|} \quad (5)$$

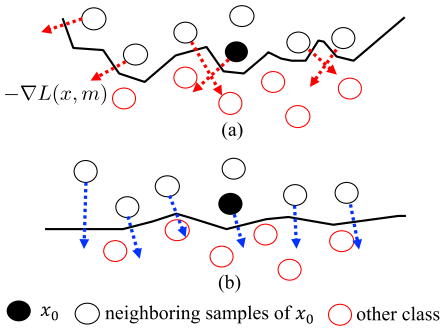


Figure 2. **Kinky gradient landscape.** Two landscapes on the classification loss lead to the identical classification accuracy. Given that a saliency map is dependent on the negative gradient of the loss function, the saliency map varies according to the landscape geometry. (a) Landscape is kinky. This results in divergent gradient profile of the neighboring data points and, thus, their inconsistent saliency maps, violating Assumption 2. (b) Ensuring a smooth landscape provides similar saliency maps for the data points.

where

$$g(x) = -\nabla L(x, m) = \nabla \log f(m \odot x). \quad (6)$$

It also holds that the loss function in Eq. (1) with respect to $x_0 + \gamma$ is upper-bounded as follows:

$$L(x_0 + \gamma, m) \leqq \alpha\|m\|_1 \left(\frac{\| - g(x_0 + \alpha v) + g(x_0)\|}{2} + \|g(x_0)\| \right). \quad (7)$$

Theorem 1 has two implications. First, the label consistency of the proposed explanation method is inversely proportional to the complexity of the saliency map that is represented as $\|m\|_1$ in Eq. (5). Intuitively, images far from x_0 can be viewed as significantly perturbed copies of x_0. Thus, a robust explanation applicable along with these data points is likely to contain a small number of features that are invariant to the perturbation. Previous work on adversarial training has addressed this issue to explain the trade-off between robustness and accuracy, arguing that the high prediction performance of models is due to subtle features that are susceptible to perturbation [36]. Our analysis reveals similar concerns in the context of the model explanation with the explanation complexity notion. Figure 1 illustrates two examples where saliency maps become more sparse (i.e., decreasing $\|m\|_1$) to maintain the classification performance of the explanations as the images undergo further perturbations. We further validate this observation using extensive experiments in Section 4.1.

Second, in contrast to the behavior of the robustness in the explanation, the classification accuracy for the neighboring data points is proportional to $||m||_1$ as shown in Eq. (7). This equation is another representation of Eq. (5) that agrees with the aforementioned trade-off of adversarial training.

Local smoothness for consistency of saliency maps. We demonstrated the effects of the explanation complexity given by $\|m\|_1$ on the robustness in terms of the classification accuracy. However, it is still unclear how the explanation complexity affects the saliency map consistency among the adjacent data points. Figure 2 motivates this investigation, where two landscapes in the classification loss have identical accuracy but the explanations of the data points at the landscapes are quite different.

Given the assumption of the local consistency in model prediction, it holds that $f(m^{(x_i)} \odot x_i) \approx f(m \odot x_0)$, and, thus, $L(x_i, m^{(x_i)}) \approx L(x_0, m)$, where $m^{(x_i)}$ and m are saliency maps for x_i and x_0, respectively. Then, it is possible to approximate $L(x_i, m^{(x_i)})$ using $L(x_0, m)$. The following result reveals the relationship between $\|m\|_1$ and the saliency map consistency along with the data points around the input data.

Theorem 2 (Local explanations with respect to saliency map consistency). *Let $\mathcal{D} = \{x_i\}$ be the vicinity of the input data x_0 such that $\|x_i - x_0\| \leqq \epsilon$ where ϵ being a small positive number. Then, the distance between the gradients of explanations of x_i and x_0 is lower-bounded as follows:*

$$\|\nabla L(x_i, m) - \nabla L(x_0, m)\| \leqq \|m\|_1 \cdot \|-g(x_0+\alpha v)+g(x_0)\|. \tag{8}$$

The distance between corresponding gradients represents similarity of saliency maps by referring to Eq. (6). Therefore, Theorem 2 indicates that the proposed method prefers a smaller value of $\|m\|_1$ for the saliency maps to be consistent, as is the classification robustness case.

The results of Theorem 1 suggest formulating RelEx as an optimization problem to consider the trade-off represented by Eq. (5) and (7). We denote an objective function $\mathcal{J}(\cdot)$ for the data points in the ℓ_p ball centered at the input image x_0 as follows:

$$\mathcal{J}(\mathcal{D}, m) = -\frac{1}{|\mathcal{D}|} \sum_i \log f(m \odot x_i). \tag{9}$$

RelEx learns a saliency map m^+ given $\mathcal{D} = \{x_i\}$,

$$m^+ = \arg\min_m \mathcal{J}(\mathcal{D}, m) + \lambda_1 \|m\|_1 \tag{10}$$

where λ_1 a regularization strength of $\|m\|_1$ following the results of Eq. (5) and (8). The use of $\|m\|_1$ coincides with previous work that generated perceptually improved images [15, 40, 38], which also applies to the proposed method. The previous explanation methods have also learned a saliency map using $\|m\|_1$ like our method [9, 37, 3]. However, we demonstrated that using a batch $\mathcal{D}$ in conjunction with the regularizer increases the robustness considerably. See the supp. S1 for proofs of the theorems.

3.2. Optimization of the Proposed Method

Fidelity of saliency maps and faithful explanations. Another critical aspect of a reliable explanation method is that a saliency map should represent *essential regions* of the input [9, 18, 3, 37]. While it is difficult to quantify the fidelity of a saliency map in general, we consider two definitions: *the smallest susceptive region* and *smallest evidential region* [9, 3, 2, 6]. The smallest susceptive region is the minimum area of an image that changes the model prediction when the region is altered. The smallest evidential region refers to the minimum area to be preserved to maintain the model prediction. Although these concepts appear similar, they are different, for example, in an image classified as "dog" containing two dogs. From the smallest evidential region viewpoint, as long as the model explanation covers any of the dogs, it may lead to correct classification. However, the explanation is not the smallest susceptive region because the part untapped by the explanation still has information concerning the target class. Determining the smallest susceptive region can be viewed as identifying a background.

The objective function in Eq. (10) is likely to determine the smallest evidential region because it is advantageous in terms of reducing $||m||_1$. Inspired by the above discussion, we incorporated an additional regularization term into the objective function to improve our explanation in the smallest susceptive region. Given the batch of $\mathcal{D}$, we considered an additional loss, $\mathcal{B}(\mathcal{D}, m)$, a classification loss for the counterpart region in x_0 with respect to m. Thus, the objective function is given by the following:

$$\begin{aligned} \mathcal{B}(\mathcal{D}, m) &= -\frac{1}{|D|} \sum_{x_i \in \mathcal{D}} \log\left(1 - f((1-m) \odot x_i)\right), \\ m^+ &= \arg\min_m \mathcal{J}(\mathcal{D}, m) + \lambda_1 ||m||_1 + \lambda_2 \mathcal{B}(\mathcal{D}, m) \end{aligned} \tag{11}$$

where λ_2 is the coefficient of $\mathcal{B}(\mathcal{D}, m)$.

4. Experimental Results

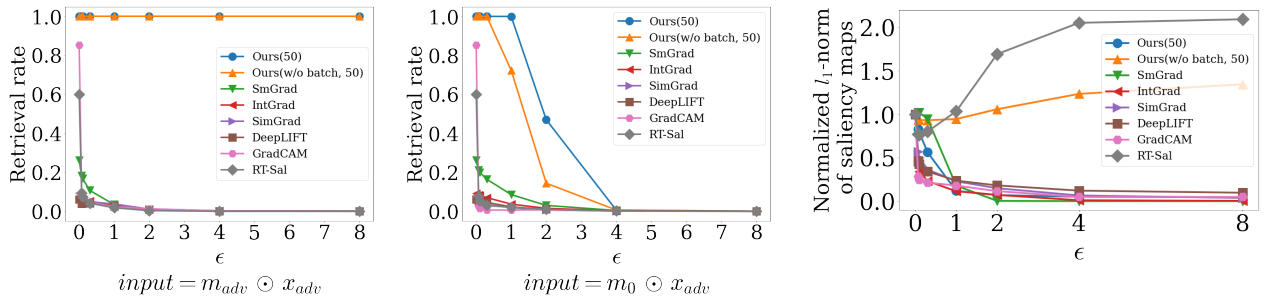
We validate the proposed method through extensive experiments to answer the following questions:

- *Are explanations using RelEx robust to retrieve the original target classes of input images against various adversarial attacks?*
- *If it is the case, how do saliency maps learned by RelEx represent relevant evidence for model predictions compared with previous work?*

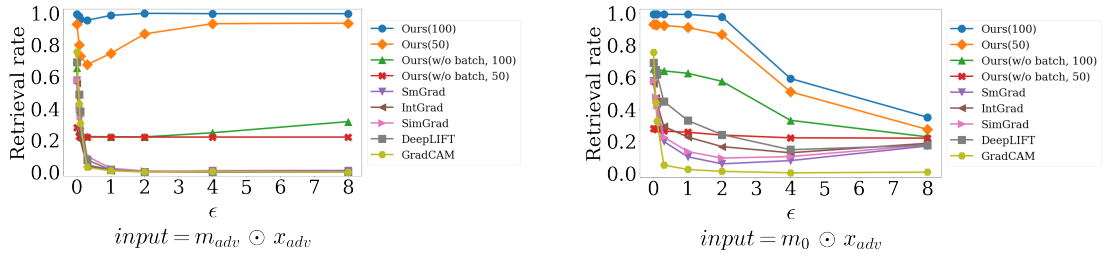
The supp. S2.1 presents the implementation details. The code of RelEx is available at https://github.com/JBNU-VL/RelEx.

4.1. Robustness Evaluations with Class Retrieval

We first demonstrate the robustness of this approach by evaluating the retrieval of the original target classes for



(a) Natural ResNet-50



(b) Robust ResNet-50

Figure 3. **Target class retrieval against the untargeted attacks on (a) the natural and (b) robust ResNet-50.** The inputs are presented below each of plots. Our method has three variants without using the batch (*w/o batch*) or different epochs to iterate (*50* and *100*) when learning saliency maps. (Top-right) Plots present the average ℓ_1-norms of saliency maps of the adversarial examples by the methods on the natural model. The norms of the saliency maps are normalized to those of the clean images for each method.

given images against adversarial attacks [35, 12] as performed in previous work [9]. Because the adversarial attacks fool models with only small perturbations to images, we use the attacks to validate the advantage of the proposed method. We applied *untargeted* and *targeted attacks* based on a white-box threat model to the classifiers to create adversarial copies of the images.

We denote sampled clean images by X_0 and their adversarial counterparts by X_{adv} for a given adversarial attack, respectively. Also, m_{adv} denotes a saliency map of $x_{adv} \in X_{adv}$. Then, we measure the rate of the successful retrieval of the target class of x_0 as the most likely class of x_{adv} for a given input, $m_{adv} \odot x_{adv}$, which is $\mathbb{E}\left[\mathbb{I}\{\arg\max_c f_c(m_{adv} \odot x_{adv}) = c_{x_0}\}\right]$ where c_{x_0} is the target class of x_0, and $\mathbb{I}(\cdot)$ is an indication function.

We evaluate two types of inputs: one is what is described above (i.e., $m_{adv} \odot x_{adv}$) and the other using a saliency map of a clean image (i.e., $m_0 \odot x_{adv}$). The evaluation of the former reveals whether a method extracts a saliency map accurately in the presence of perturbation. The latter evaluates the robustness of saliency maps of clean images.

Adversarial attack methods. The untargeted attack manipulates an input image *to lead model predictions to arbitrary false labels*. We applied RelEx to pretrained ResNet-50 [13] using ImageNet [4]. We also considered a robust counterpart of the model that was adversarially trained [36]. We denote them as *natural* and *robust* models, respectively. We expect that adversarial examples crafted against the robust model are more difficult to defend than those from the natural model and, therefore, can better evaluate explanation methods.

The targeted attack aims to change an image to mislead to a specified false class or saliency map. We chose to *create false saliency maps of given images against a given explanation method with their classes kept* comparing with previous work [5, 10]. We used two strategies for manipulating saliency maps: *unstructured* and *structured* attacks. The former aims to create uninformative saliency maps whereas the latter misleads to the saliency map of a specified class.

Setups for evaluation using the untargeted attack. For the natural ResNet-50, we randomly sampled 4000 images from the validation set of ImageNet. The robust ResNet was trained on the Restricted ImageNet, a customized subset of ImageNet [36]. We additionally sampled 1000 random images from the validation set of the dataset for the robust model. Then, we created corresponding adversarial images using the projected gradient descent (PGD), one of the best universal first-order adversarial attacks [17] with the configuration for the MNIST dataset. We varied the ℓ_∞-norm of the perturbation distance to $\{0.07, 0.1, 0.3, 1, 2, 4, 8\}$. We compared our approach with the followings: *SimGrad* [28], *SmGrad* [30], *IntGrad* [34], *DeepLIFT* [27], *RT-Sal* [3], and *GradCam* [25]. See the supp. S2.2 for the details on the adversarial example generations.

Results of the untargeted attack. We present the results in Figure 3. First, our explanations successfully retrieve the

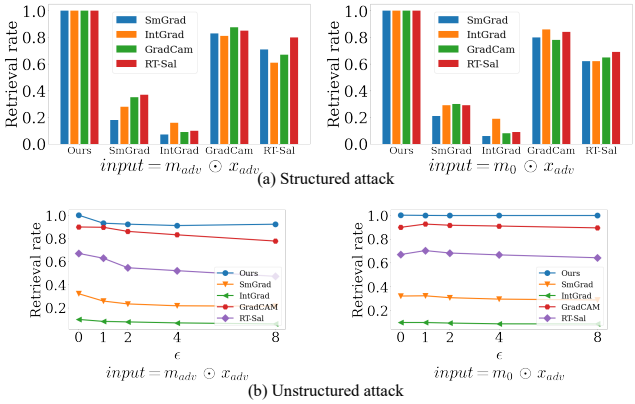


Figure 4. **Target class retrieval performance against the targeted attacks on the natural ResNet-50.** The explanation types are presented below each plot. (a) The horizontal axis represents the explanation methods to compare with each other. Each colored bar in the plots indicates a method to which the attack is applied. (b) Plots correspond to the adversarial images against RT-Sal [3].

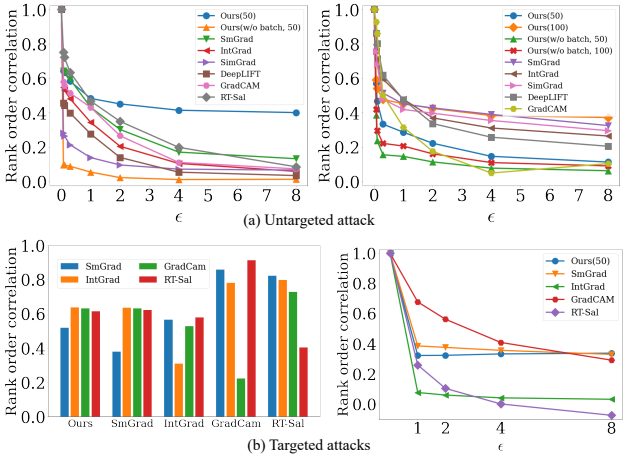


Figure 5. **Similarity of the saliency maps in the rank order correlation for (a) untargeted and (b) structured attacks.** (Bottom-left) See Figure 4(a) for the axis labels. (Bottom-right) The adversarial images were created against RT-Sal [3].

original target classes along with the entire perturbation distance. This result shows that our approach extracts meaningful evidence robustly in the presence of severe perturbations. Other methods encountered significant performance decrease even at the smallest perturbation, $\epsilon = 0.07$, which is visually imperceptible. This is probably because the subtle features in the clean images were mostly perturbed even with such a small value of ϵ; thus, the methods failed to determine robust features invariant to the perturbations. Second, our explanation learned from the clean images contains robust features such that it applies to the adversaries up to $\epsilon = 1.0$. This allows a model to be tolerant against such an attack without adversarial training when combined with RelEx. In contrast, other approaches performed poorly, similar to the previous case.

For the ablation study, we considered four variants of our method without the batch or varying the number of epochs to iterate for solving Eq. (11). The benefit of the batch and iterating more epochs increased the robustness of explanations on clean images against the attack, which is more significant on the robust model.

Figure 3 also depicts that the higher perturbation results in the smaller ℓ_1-norms of our saliency maps. Although we observe a similar behavior with the gradient-based approaches, they failed to determine relevant features, performing poorly in the class retrieval. These results demonstrate the benefit of smoothing the local explanation as discussed in Section 3.1.

Setups for evaluation using the targeted attacks. We created 1000 adversarial examples of the sampled input images by applying the targeted attacks to each method on the natural ResNet-50 as proposed in [5] and [10] for the structured and unstructured attacks, respectively. However, we observed that no adversarial examples were generated against our method. Although a correct analysis of this observation is beyond the scope of this study, we propose that it is because of the insufficient perturbations of the targeted attacks to mislead our method. We empirically validated the assumption that the ℓ_2-norm of saliency maps due to the targeted attacks belongs to the region where RelEx is reliable in the untargeted attack. (See the supp. S3.1) Instead, we evaluated our method using the adversarial examples against other methods by assuming that the adversarial attack is transferable [14].

Results of the targeted attacks. Overall, the results of the targeted attacks are similar to those of the untargeted attack, as depicted in Figure 4. As expected, we observed that the attacks are transferable. RelEx achieved an outstanding retrieval rate of close to 1 over all the settings. Unlike the results of the untargeted attacks, the retrieval rates with GradCAM and RT-Sal, which are about 0.83 and 0.60, respectively, are comparable to the rate of the proposed method. This suggests that the PGD-based untargeted attack is more effective than the targeted attacks. Although Figure 4 provides the results against RT-Sal in the unstructured attack, we found similar observations in the attacks against other methods. See the supp. S3.4 for more results.

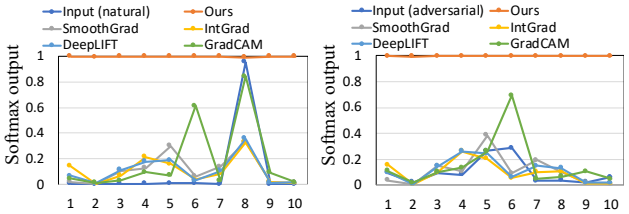
4.2. Evaluations of the Fidelity of Saliency Maps

Metrics for the similarity of saliency maps. The evaluations of the target class retrieval demonstrated the robustness of the learned explanations by RelEx. To understand why, we delved into the fidelity of the saliency maps. In particular, we quantified the quality of the saliency maps of adversarial examples by measuring 1) their spatial similarity to those of their clean counterparts and 2) the relevance of features identified by an explanation to a class score.

In the similarity evaluation, we use a metric, *Spearman's*

Table 1. **Comparison of the feature relevance, R, of saliency maps for given ℓ_∞-norm of perturbation, ϵ, against the untargeted attack.**

	Natural ResNet-50				Robust ResNet-50			
ϵ	0.07	0.1	0.3	1.0	0.07	0.1	0.3	1.0
SimGrad	0.03	0.03	0.04	0.03	0.40	0.37	0.22	0.14
IntGrad	0.07	0.07	0.06	0.03	0.35	0.30	0.13	0.07
SmoothGrad	0.18	0.17	0.12	0.04	0.38	0.33	0.15	0.10
DeepLIFT	0.04	0.04	0.04	0.04	0.48	0.43	0.20	0.12
GradCAM	0.14	0.11	0.08	0.03	0.58	0.45	0.39	0.16
RT-Sal	0.12	0.10	0.06	0.02	N/A	N/A	N/A	N/A
MASK	0.05	0.04	0.03	0.01	0.40	0.32	0.15	0.14
RelEx	**0.94**	**0.95**	**0.96**	**0.97**	**0.78**	**0.66**	**0.60**	**0.56**



(a) Natural images (b) Adversarial images

Figure 6. **Comparison of explanations for arbitrarily chosen classes on CIFAR-10.** (a) Plots show the softmax scores of explanations corresponding to all 10 classes for given natural images of *class 8*. (b) Plots correspond to adversarial counterparts.

rank-order correlation [31], to evaluate the similarity of saliency maps as in [10, 5]. The Spearman's rank-order correlation inherently ranks the importance of input features according to a saliency map and enables us to naturally correlate feature ranks between saliency maps.

Results of the similarity of saliency maps. Figure 5 indicates that the spatial similarity of RelEx is analogous to those of other methods, unlike the case of the class retrieval. Worse, all other methods in the case of the untargeted attack outperform our method without the batch. The results suggest that the similarity metric is inadequate to explain the outstanding performance of RelEx in the class retrieval. The performance mismatch between the class retrieval and the similarity of saliency maps by existing methods is due to the incorrect attribution of input features. We observed that RelEx learned a saliency map adapting to the degree of perturbations. In contrast, the existing methods failed to attribute relevant input features to the neighboring data points, and their saliency maps remain fixed, as illustrated in Figure 1(b). See the supp. S3.2 for results of another metric.

Metrics for feature relevance evaluation. We use two metrics to evaluate the pixel-level relevancy for a given saliency map: *deletion* and *preservation*. Deletion quantifies the accuracy of finding the smallest susceptive region, whereas preservation corresponds to the smallest evidential region, discussed in Section 3.2. See the supp. S3.3 for the details on the metrics. The sole use of deletion is discouraged because, for instance, two extreme cases of the accurate and completely wrong smallest susceptive regions may have an identical deletion score. Therefore, similar to the F_1 score, we use the harmonic mean of the deletion and preservation, R, as a metric of feature relevance; in particular, $\frac{1}{R} = \frac{1}{2}\left(\frac{1}{P} + \frac{1}{1-D}\right)$, where P and D are the preservation and deletion scores. We used $1 - D$ because a lower score results in better deletion.

Results of the feature relevancy. Table 1 indicates that RelEx outperforms other methods in terms of the pixel-level relevancy, R. The results confirm the evaluations of the target class retrieval, and visually plausible saliency maps do not necessarily present true evidence for explaining model predictions. See the supp. S3.5 for more results.

Extracting explanations conditioned on arbitrary classes. Finally, we investigate whether our method can extract explanations conditioned on arbitrary non-target classes for a given input. We sampled 400 images annotated as *class 8* from the test set of CIFAR-10. Then, we drew explanations of classes from the samples and their adversaries on natural ResNet-18 with the ℓ_∞-norm of perturbation set to 8 out of 255 [36], and compared the softmax scores of our explanations to those of the existing methods. Figure 6 illustrates that the scores of our explanations are close to 1 consistently through all classes on both the natural samples and their adversarial examples, outperforming others. The results indicate that RelEx can extract explanations as long as relevant evidence exists in the input. The explanations represent the specified classes faithfully, effectively excluding information on irrelevant classes. We provide more results in the supp. S3.6.

5. Conclusion

We introduced a reliable explanation of neural networks that requires consistency on model outputs and the corresponding saliency maps along with neighboring data points. The proposed method, RelEx, addresses the concern by interpreting the model explanations via a locally smooth landscape with respect to the loss function of the model output. Our analysis demonstrated that the smoothness in the landscape improves as we reduce the ℓ_1-norm of a saliency map. The experimental results demonstrated that the proposed method based on the analysis identifies features relevant to the target class retrieval against the strong white-box attacks. We also demonstrated that causal evidence for model predictions does not always coincide with visually appealing saliency maps as in previous methods.

Acknowledgement

This work was supported by a National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (2019R1F1A1061941).

References

[1] 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. 1, 2, 3

[2] Chun-Hao Chang, Elliot Creager, Anna Goldenberg, and David Duvenaud. Explaining image classifiers by counterfactual generation. In *International Conference on Learning Representations*, 2018. 5

[3] Piotr Dabkowski and Yarin Gal. Real time image saliency for black box classifiers. In *Advances in Neural Information Processing Systems*, pages 6967–6976, 2017. 1, 3, 5, 6, 7

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[5] Ann-Kathrin Dombrowski, Maximillian Alber, Christopher Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. Explanations can be manipulated and geometry is to blame. In *Advances in Neural Information Processing Systems*, pages 13589–13600, 2019. 1, 2, 3, 6, 7, 8

[6] Mengnan Du, Ninghao Liu, Qingquan Song, and Xia Hu. Towards explanation of dnn-based prediction with guided feature inversion. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1358–1367, 2018. 5

[7] Alhussein Fawzi, Hamza Fawzi, and Omar Fawzi. Adversarial vulnerability for any classifier. In *Advances in neural information processing systems*, pages 1178–1187, 2018. 1

[8] Alhussein Fawzi, Seyed-Mohsen Moosavi-Dezfooli, Pascal Frossard, and Stefano Soatto. Empirical study of the topology and geometry of deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3762–3770, 2018. 3

[9] Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3429–3437, 2017. 1, 3, 5, 6

[10] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3681–3688, 2019. 1, 3, 6, 7, 8

[11] Justin Gilmer, Luke Metz, Fartash Faghri, Samuel S Schoenholz, Maithra Raghu, Martin Wattenberg, and Ian Goodfellow. Adversarial spheres. *arXiv preprint arXiv:1801.02774*, 2018. 1

[12] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 1, 6

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[14] Juyeon Heo, Sunghwan Joo, and Taesup Moon. Fooling neural network interpretations via adversarial model manipulation. In *Advances in Neural Information Processing Systems*, pages 2925–2936, 2019. 1, 2, 3, 7

[15] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 5

[16] Andrei Kapishnikov, Tolga Bolukbasi, Fernanda Viégas, and Michael Terry. Xrai: Better attributions through regions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4948–4957, 2019. 1, 3

[17] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. 1, 2, 6

[18] David Alvarez Melis and Tommi Jaakkola. Towards robust interpretability with self-explaining neural networks. In *Advances in Neural Information Processing Systems*, pages 7775–7784, 2018. 1, 5

[19] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Jonathan Uesato, and Pascal Frossard. Robustness via curvature regularization, and vice versa. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9078–9086, 2019. 1, 3, 4

[20] 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, 3

[21] Chongli Qin, James Martens, Sven Gowal, Dilip Krishnan, Krishnamurthy Dvijotham, Alhussein Fawzi, Soham De, Robert Stanforth, and Pushmeet Kohli. Adversarial robustness through local linearization. In *Advances in Neural Information Processing Systems*, pages 13847–13856, 2019. 3

[22] Sylvestre-Alvise Rebuffi, Ruth Fong, Xu Ji, and Andrea Vedaldi. There and back again: Revisiting backpropagation saliency methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8839–8848, 2020. 1, 3

[23] Andrew Slavin Ross and Finale Doshi-Velez. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. *arXiv preprint arXiv:1711.09404*, 2017. 3

[24] Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. In *Advances in Neural Information Processing Systems*, pages 5014–5026, 2018. 1

[25] 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*, pages 618–626, 2017. 1, 2, 3, 6

[26] Uri Shaham, Yutaro Yamada, and Sahand Negahban. Understanding adversarial training: Increasing local stability of neural nets through robust optimization. *arXiv preprint arXiv:1511.05432*, 2015. 3

[27] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation

differences. In *International Conference on Machine Learning*, pages 3145–3153, 2017. 1, 2, 3, 6

[28] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 1, 2, 6

[29] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 180–186, 2020. 1, 3

[30] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017. 1, 2, 3, 6

[31] Charles Spearman. The proof and measurement of association between two things. *The American journal of psychology*, 100(3/4):441–471, 1987. 8

[32] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014. 1

[33] Akshayvarun Subramanya, Vipin Pillai, and Hamed Pirsiavash. Fooling network interpretation in image classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2020–2029, 2019. 1, 3

[34] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3319–3328, 2017. 1, 2, 6

[35] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 6

[36] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2018. 1, 2, 4, 6, 8

[37] Jorg Wagner, Jan Mathias Kohler, Tobias Gindele, Leon Hetzel, Jakob Thaddaus Wiedemer, and Sven Behnke. Interpretable and fine-grained visual explanations for convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9097–9107, 2019. 1, 3, 5

[38] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018. 5

[39] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, pages 7472–7482, 2019. 1, 2

[40] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 5