Model-Agnostic Adversarial Detection by Random Perturbations

Bo Huang[1,2] , **Yi Wang**[1] * and **Wei Wang**[3]
[1]Dongguan University of Technology, Dongguan, China
[2]Shenzhen University, Shenzhen, China
[3]The University of New South Wales, Sydney, Australia
huangbo1024@gmail.com, wangyi@dgut.edu.cn, weiw@unsw.edu.au

Abstract

Adversarial examples induce model classification errors on purpose, which has raised concerns on the security aspect of machine learning techniques. Many existing countermeasures are compromised by adaptive adversaries and transferred examples. We propose a model-agnostic approach to resolve the problem by analysing the model responses to an input under random perturbations, and study the robustness of detecting norm-bounded adversarial distortions in a theoretical framework. Extensive evaluations are performed on the MNIST, CIFAR-10 and ImageNet datasets. The results demonstrate that our detection method is effective and resilient against various attacks including black-box attacks and the powerful CW attack with four adversarial adaptations.

1 Introduction

Machine learning techniques are core to many real-world systems. With their prevalent application and wide deployments, there are increasing concerns of machine learning in adversarial settings where an intelligent attacker may compromise a learning-based decision maker and disfunction the depending system. For instance, machine learning models are found to be susceptible to adversarial examples – those inputs crafted with non-random *adversarial perturbations* to intentionally cause model misclassification, known as *evasion attacks* at test time [Biggio *et al.*, 2013]. Surprisingly, the amount of adversarial perturbation required to fool complex models like deep neural networks (DNNs) is small and often imperceptible to human eyes [Szegedy *et al.*, 2014].

Existing approaches to alleviate the adversarial problem can be roughly categorized into 1) *defenses* that aim at making the underlying model more robust to adversarial attacks, and 2) *detections* that attempt to distinguish adversarial examples from normal inputs. In this paper, we focus on devising an effective detection method that maximizes the chance of allowing only the legitimate input to the intended model.

By directly dealing with the learning model or not, adversarial detection methods may be further classified into *model-dependent* and *model-agnostic* approaches. The model-dependent schemes often leverage the underlying model properties or internal states to detect the adversarial class such as by adding detection layers/subnetworks [Lu *et al.*, 2017] or changing the loss/activation functions [Madry *et al.*, 2018]. The model-agnositic detectors are mainly built based on analysing the input and/or output feature characteristics without requiring access to the model under protection [Grosse *et al.*, 2017; Xu *et al.*, 2017; Guo *et al.*, 2017].

However, many existing defenses are defeated by *adversarial adaptations* when the attack is no longer assumed oblivious of defenses being deployed [Carlini and Wagner, 2017a]. In *adaptive white-box* attacks, for example, it was shown possible to incorporate specific model information into constructing more powerful adversarial examples to evade both the classifier and the adversarial detector at the same time. Adversarial perturbations are also transferrable across models and transformations [Szegedy *et al.*, 2014]. The resulting *black-box attacks* often perform better than white-box attacks against defenses that are based on breaking gradient descents [Athalye *et al.*, 2018; Tramèr *et al.*, 2018].

In this paper, we propose a simple yet effective method for detecting adversarial image examples, which can be easily deployed into all off-the-shelf deep learning models. Our intuition stems from the observations that decision boundaries of adversarial subspaces tend to lie closely to the submanifold of legitimate data in adversarial directions [Ma *et al.*, 2018]. Thus, expanding the adversarial subspace by additive random perturbation can result in a certain probability of landing an adversarial example *back* to the data manifold. On the other hand, small random noise does not usually cause misclassification due to the robustness of deep learning models [Fawzi *et al.*, 2016].

The difference of classifiers' robustness to "noisy" inputs motivated us to extract statistical features from the model responses under multiple random perturbations. We note that there are recent work that also use random noise in improving robust training against adversarial examples such as [Lecuyer *et al.*, 2018]. Our approach has the following novelty in terms of the use of random noise:

- We take relative variation of confidence as the discriminative feature for adversarial detection, instead of taking the expectation of the prediction scores for robustness evaluation of the target classifier.

*Corresponding Author

- Our method is *Model-agnostic property*, which does not require knowledge of the protected model details and hence can be applied blindly to all models.
- We develop our own theoretical analysis by relating the tail bound of random perturbation to the norm-bounded adversarial distortions.

2 Related Work

The literature is seeing a fast growing number of attacks and countermeasures for adversarial machine learning based on some understanding of adversarial examples and characterization of the surrounding subspaces [Papernot *et al.*, 2018; Carlini and Wagner, 2017a]. For instance, [Goodfellow *et al.*, 2015] attributed adversarial perturbations to the linearity hypothesis of neural networks and proposed the fast gradient sign method (FGSM). It was then extended to generate more powerful attacks such as the iterative version of BIM [Kurakin *et al.*, 2017], Jacobian-based Saliency Map Attack (JSMA) [Papernot *et al.*, 2016]. [Madry *et al.*, 2018] considered the whole class of gradient-based methods as the *first-order adversary* and proposed an optimization view towards adversarial robustness. In fact, it was shown that iterative optimization-based attacks seem to produce near-optimal adversarial examples in terms of *minimal distortions* [Carlini and Wagner, 2017a]. Attack examples of this strategy include the CW attack [Carlini and Wagner, 2017b] and DeepFool [Moosavi-Dezfooli *et al.*, 2016].

Most of the existing attacks are constructed with explicit model information, e.g., the loss gradients, in generic *white-box* settings. A natural remedy is to conceal such knowledge or mislead the attacker. [Athalye *et al.*, 2018] identified three ways to obfuscate the gradients: 1) *shattered gradients* caused by a non-differentiable defense function, 2) *stochastic gradients* due to randomization – either randomized network or randomized input, and 3) *vanishing/exploiding gradients* with extremely deep or cascaded networks. Accordingly, the authors proposed alternative gradient estimation techniques for each of these obfuscation strategies and demonstrated that it is possible to defeat most of the existing defenses.

On the other hand, the intrinsic property of adversarial transferrability is an obstacle for building robust countermeasures [Tramèr *et al.*, 2018]. Without knowing details of the model under protection, model-agnostic approaches have inherent advantages in dealing with transferred examples. Such detectors are often built on features extracted from the input sample statistic [Grosse *et al.*, 2017] or the prediction difference w.r.t. an input transformation [Akhtar *et al.*, 2018; Xu *et al.*, 2017]. For example, feature squeezing [Xu *et al.*, 2017] compares the model output before and after "squeezing" the input features by some operations to differentiate the normal and adversarial examples.

Unfortunately, many existing detectors are still not immune to adversarial adaptations when the attacker is *not* assumed oblivious of the defense [Carlini and Wagner, 2017a]. In the so-called *adaptive white-box* attacks, it was shown possible to construct more powerful adversarial examples to evade the classifier and the detector at the same time. The key to success in such adaptive attacks is to incorporate *differentiable* loss functions of both models into the objective function of an attack. Therefore, it was highly recommended that detector evaluations should involve the threat model of adaptive attacks and demonstrate the detector robustness against attack transferrability [Carlini and Wagner, 2017b].

3 Proposed Approach

Adversarial examples generated by all attack methods are of the form $\mathbf{x}^{\text{adv}} = \mathbf{x} + \boldsymbol{\delta}$, inherently constrained by the fact that $\boldsymbol{\delta}$ is bounded by some small constant. We propose to apply some "disturbance" signal of appropriate magnitude to the input example. If the input is indeed an adversarial example, we may have a non-negligible probability of pushing the resulting input back to the manifold of the original class. If this happens, there will be a clear change of class labels and their associated scores. Therefore, we can gauge the model responses and summarize certain statistics to distinguish $\mathbf{x}^{\text{adv}}$ from normal $\mathbf{x}$. In the following, we introduce the main steps of our approach followed by a theoretical uncertainty analysis of the model responses to random perturbations over adversarial examples.

3.1 Main Steps

Given an input $\mathbf{x}$, let $\hat{c}$ be its predicted class by model F, i.e., $\hat{c} = \arg\max_i F(\mathbf{x})[i]$. We apply a random perturbation $\boldsymbol{\eta}$ drawn i.i.d. from the Gaussian distribution $N(\mathbf{0}, \text{diag}(\sigma))$, and measure the *relative* score difference for $\hat{c}$ as

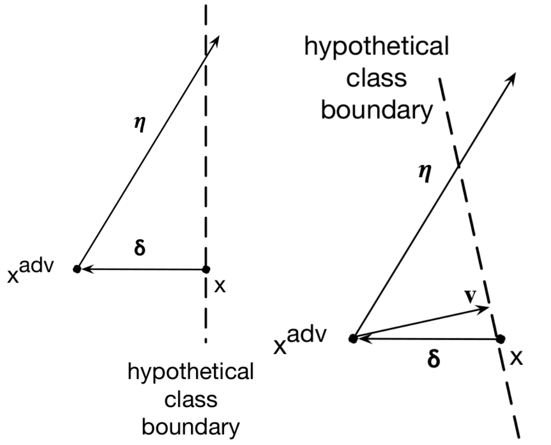
$$r_{\hat{c}} = \frac{F(\mathbf{x})[\hat{c}] - F(\mathbf{x} + \boldsymbol{\eta})[\hat{c}]}{F(\mathbf{x})[\hat{c}]}$$

To account for the stochastic nature of such raw signals, we decide to repeat the process m times and extract statistically robust feature from such sampled distribution. For example, we extract an 17-dimensional feature vector by taking the $10\%, 15\%, 20\%, \ldots, 90\%$ quantiles of m samples so that it can be more robust to noise and outliers.

We then train a binary classifier [1] for the adversarial example detection. The training data of detector classifier consists of the original training data, labelled as the normal examples, and those generated from the normal examples using a specific attack algorithm, labelled as adversarial examples. Once the detector classifier has been trained, it can be viewed as a probability generator and output a confidence score for a given sample. We apply a simple thresholding to the detector confidence score such that, the greater the probability is, the more likely the sample is considered as adversarial.

While the above detection algorithm alone can work fairly well, as will be shown in Section 4.2, here we also propose a complementary step that can significantly enhance the adversarial detection performance. We encourage the target model to classify a randomly perturbed normal example to be in the same class as the non-perturbed one. This can be done by simply injecting similar noise perturbations into model training. In this way, the retrained model will be more robust to a legitimate input with benign noise, which helps to enlarge the difference of patterns between model responses to the normal and adversarial inputs. We refer to the complementary step as *noise augmentation* in the following sections.

[1] We use an SVM (with RBF kernel) classifier in our experiments.



(a) Perpendicular Case (b) General Case

Figure 1: Illustration of calculating cross-boundary probability for our random perturbation $\boldsymbol{\eta}$

3.2 Cross-Boundary Probability

By the definition of adversarial examples, we know that the predicted class labels of $\mathbf{x}$ and $\mathbf{x}^{\text{adv}}$ are different. Assuming we are fed with an adversarial example $\mathbf{x}^{\text{adv}}$, we would like to compute the probability that the predicted class label of our randomly perturbed example $\mathbf{x}^{\text{adv}} + \boldsymbol{\eta}$ is the same as that of $\mathbf{x}$ (i.e., successfully recover the clean label by crossing the class boundary). This probability strongly correlates with the changes in the scores, and serves as a means both to determine the parameter σ and to understand the underlying principle of our detection methods.

Unfortunately, the probability cannot be computed without additional information or assumption, as we do not know the location and the shape of the class boundary. We make a mild assumption that the class boundary is a $d-1$ dimension hyperplane incident on $\mathbf{x}$. The local linearity assumption is not unusual in other studies such as [Moosavi-Dezfooli *et al.*, 2016]. We also make an additional assumption that the hyperplane is perpendicular to $\boldsymbol{\delta}$. See Fig. 1(a) for an illustration. We will relax this assumption later.

Let $\boldsymbol{\eta} = [\eta_1, \ldots, \eta_d]$, where $\eta_i \sim N(0, \sigma^2)$ is sampled i.i.d.. Consider the projection of X onto a fixed unit vector $\mathbf{u} \overset{\text{def}}{=} \frac{-\boldsymbol{\delta}}{||\boldsymbol{\delta}||}$. As Gaussian variables 2-stable, let $t \overset{\text{def}}{=} \mathbf{u}^\top \boldsymbol{\eta}$, we know t follows $\sigma \cdot v$ where $v \sim N(0,1)$. In other words, $\frac{t}{\sigma} \sim N(0,1)$.

If we repeat this process m times, the probability that at least one such $\boldsymbol{\eta}$ reaches the other class is

$$p_{\text{cross}} \overset{\text{def}}{=} 1 - (1 - \Pr\{t \geq \delta\})^m \ .$$

We only need to let $\Pr\{t \geq \delta\} \geq \frac{1}{m}$, such that p_{cross} is at least $1 - \frac{1}{e} \approx 0.6321$.

Remarks. Since $\Phi^{-1}(x) = \sqrt{2}\text{erf}^{-1}(2x-1)$, to ensure

$$\Pr\{t \geq \delta\} = 1 - \Phi\left(\frac{\delta}{\sigma}\right) \geq \frac{1}{m},$$

one of the following conditions should be met:

Dataset	Model	Top-1 Accuracy	Top-5 Accuracy
MNIST	LeNet	99.2%	-
CIFAR-10	ResNet-20	91.2%	-
ImageNet	ResNet-101	76.4%	92.9%

Table 1: Summary of target models

- If δ and σ are fixed, we only need $m \geq \frac{1}{1-\Phi(\delta/\sigma)}$.
- If σ and m are fixed, the maximum $\delta_{\max}$ that at least one of our m perturbations can cross over with a probability of at least $\kappa \cdot \sigma$, where $\kappa = \sqrt{2}\text{erf}^{-1}(1 - \frac{2}{m})$. When $m = 50$, $\kappa = 2.0537$.

General Case

If the class boundary is not perpendicular to $\boldsymbol{\delta}$, we denote the vector starting from $\mathbf{x}^{\text{adv}}$ and perpendicular to the class boundary as $\mathbf{v}$. Obviously $||\mathbf{v}|| \leq ||\boldsymbol{\delta}||$ as shown in Fig. 1(b). Following the same argument as in the perpendicular case, we only need

$$\Pr\left\{\frac{\mathbf{v}}{||\mathbf{v}||}^\top \boldsymbol{\eta} \geq ||\mathbf{v}||\right\} \geq \frac{1}{m}$$

to ensure $p_{\text{cross}} \geq 1 - \frac{1}{e}$. Since $||\mathbf{v}|| \leq ||\boldsymbol{\delta}||$, hence

$$\Pr\left\{\frac{\mathbf{v}}{||\mathbf{v}||}^\top \boldsymbol{\eta} \geq ||\mathbf{v}||\right\} \geq \Pr\left\{\mathbf{u}^\top \boldsymbol{\eta} \geq ||\boldsymbol{\eta}||\right\} \ .$$

The perpendicular turns out to be the worst case, so our above conditions/conclusions still apply in the general case.

The above calculation, especially $\delta_{\max}$, gives us a rule of thumb to set our random perturbation parameter σ. Our empirical evaluation shows that this theoretical analysis is highly accurate (see Fig. 3). The uncertainty analysis can also be extended to other norm objectives such as l_∞ in a similar way.

Here we only analyze the case when the input is indeed an adversarial example. Thus, we rely on the robustness of deep learning models to random noises [Fawzi *et al.*, 2016]. Empirically, we found most models are reasonably robust and, with additional noise augmentation training, the resulting models demonstrate high robustness.

4 Evaluations

We evaluate the performance of our approach on detecting adversarial examples for the task of image classification over *three* benchmark datasets: MNIST, CIFAR-10, and ImageNet. In the following, we introduce our experimental settings including setups of the target model, attack methods, and threat models. Under each threat model, we evaluate and compare the adversarial detection performance. In particular, we examine the robustness of our detector against adaptive adversaries and transferred attacks – either across different attack forms or across different target models.

4.1 Experimental Settings

Target models. For MNIST and CIFAR-10, we used the designated training set for training and the designated test set

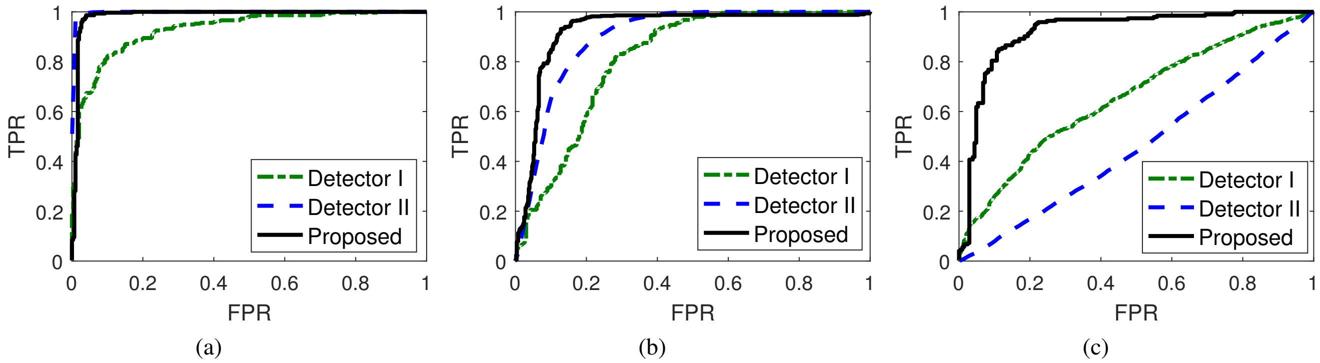


Figure 2: ROC curves against the BIM (l_∞) attack over (a) MNIST, (b) CIFAR-10, (c) ImageNet. The proposed method is compared with two represenative schemes of Detector I: *SafetyNet* that is model dependent, and Detector II: *Feature Squeezing* that is model agnostic. The model classifier is trained *without* noise augmentation.

for testing. For ImageNet, we used a pretrained DNN classifier and the first $10,000$ samples of validation set as our test examples for evaluation. Table 1 summarizes the standard classification accuracy measures of the targets, which are comparable with state-of-the-art results considering size of the model.

Attack methods. For each target model, we generate adversarial examples from test samples and use only those that can attack successfully before deploying any countermeasure to the target model in all of our experiments. We conduct *untargeted* attacks to each target model with *five* representative attack algorithms, namely FGSM, BIM, JSMA, DeepFool, and CW attacks, as introduced in Section 2. In particular, the iterative optimization-based approaches of DeepFool and CW attacks are considered stronger with higher success rates under the same norm objective in the white-box setting [Athalye *et al.*, 2018]. On the other hand, the FGSM methods transfer better [Su *et al.*, 2018]. BIM can be viewed as a PGD inside an l_∞ ball, which is among the strongest attack in the first-order family [Cisse *et al.*, 2017]. Our implementations are based on the Cleverhans2.0 library [2].

Threat models. Denote the target classifier by F and the detector by D. We consider the following threat models by knowledge and capability of the adversary:

- *Oblivious Adversary* follows the generic white-box setting that assumes a full access and knowledge to F but is not aware of D in place.
- *Adaptive Adversary* knows the model details of both F and D but *cannot* decide the test-time randomness. In our context, the test-time randomness is η sampled m times *i.i.d.* from $\mathcal{D}^{\text{noise}}$. The adversary may use his knowledge about F and D to construct more powerful *adaptive white-box* attacks.
- *Transferred Adversary* exploits the transferrability of adversarial examples. Here, we consider two scenarios: 1) The adversary knows F but cannot access D trained with an attack strategy A. Alternatively, he may generate adversarial examples using another strategy B to attack F. Ideally, D should still be able to detect those unseen examples generated by B. We refer to this scenario as the *generalizability analysis* of the detector.

  2) The adversary cannot access F. Alternatively, he may build adversarial examples from another model $\tilde{F}$ to attack F. Ideally, D should still able to detect those unseen examples generated targeting $\tilde{F}$. We refer to this scenario as the *black-box attack* across target models.

Detector		BIM	DeepFool	CW
		l_∞	l_2	l_2
MNIST	I)	0.931	0.908	0.890
	II)	**0.997**	**0.995**	0.995
	Ours	0.986	**0.995**	**0.998**
CIFAR-10	I)	0.814	0.814	0.820
	II)	0.897	0.898	0.916
	Ours	**0.928**	**0.984**	**0.957**
ImageNet	I)	0.656	0.423	0.685
	II)	0.461	0.898	0.827
	Ours	**0.919**	**0.910**	**0.869**

Table 2: AUC scores of ROC on adversarial detection comparing the proposed method with two other detectors: I) *SafetyNet* and II) *Feature Squeezing*. The best results are highlighted in **bold**.

4.2 Adversarial Detection Performance

We regard adversarial examples as the positive class and natural images as the negative class, and randomly select 80% of samples from each class to train the detector classifier, and use the remaining 20% for test. By changing the threshold value of the detector classifier, we provide a detection capability with varying trade-off between the *false positive rate* (FPR) and the *false negative rate* (FNR). The *true positive rate* (TPR) is computed as (1-FNR).

We first evaluate under the oblivious adversary model. For illustration, Fig. 2 plots the ROC curves against the BIM

[2] https://github.com/tensorflow/cleverhans

(l_∞) attack. We compare the performance of the proposed approach *without* noise augmentation with two other detectors: I) SafetyNet [Lu *et al.*, 2017] and II) Feature Squeezing [Xu *et al.*, 2017]. For the latter, we use the best attack-specific single squeezer for each dataset and the recommended detection threshold value as reported in the paper. Both comparing methods are representative in adversarial detection. In particular, SafetyNet represents the few approaches that are resilient against adaptive adversaries. As reviewed in Section 2, many other detection methods work well under oblivious white-box attacks, i.e., assuming an access to the target classifier but not aware of the detector in place. However, they are often defeated by adaptive attacks that use full knowledge of both models [Carlini and Wagner, 2017a]. Feature Squeezing is a well-known *model-agnostic* approach that treats the underlying classifier as a black box, which is inherently more robust to transferred examples. The detector is built based on the input-and-output analysis in the same paradigm as ours.

In Fig. 2, the two model-agnostic approaches work very well on the MNIST dataset, while the proposed approach outperforms the comparing detectors over CIFAR-10 and ImageNet. This can be seen more clearly in Table 2 which summarizes the AUC scores of ROC for adversarial detection performance against more attacks over the three benchmark datasets. The best results are highlighted in **bold**. It is worth noting that the results in Fig. 2 and Table 2 are performed *without* noise augmentation. In terms of both measures, the proposed approach either outperforms or is on a par with the other two detectors.

In adversarial detection, the cost of admitting an adversarial example from the positive class is far more severe than that of rejecting a legitimate one from the negative class. For convenience, it is sometimes desirable to report the adversarial detection accuracy (TPR) at a tolerable FPR. As introduced in Section 3.1, our detection performance can be significantly improved by *noise augmentation* for small FPR as shown in Fig. 2. We evaluated the adversarial detection performance using our approach with the complementary step. The results of TPR at 5% FPR are show in Table 3.

Although noise augmentation is included, the difference of relevance score by additive random perturbation still plays the most critical part in the proposed detection framework. To show this, we have further conducted an ablation test based on the experiment settings in Table 3. Without noise augmentation, for example, the detection accuracy drops to about 0.771 against the CW attack over CIFAR-10. We believe that the performance reduction is related to the inherent classification accuracy of the CIFAR-10 model (see Table 1). As discussed in Section 3.1, in such cases, the complementary step of noise augmentation can help to stabilize the model prediction on legitimate inputs with benign noise.

We also tested the proposed approach by training the detector classifier on different types of attacks. For example, the detection accuracy (TPR@5%FPR) of CW examples over the CIFAR-10 dataset drops only slightly from 0.987 to 0.95 when the detector is trained on BIM instead of CW. This is in fact a type of *Transferred Adversary* as described in Section 4.1. We will report more about the robustness of our detector under this threat in Section 4.4.

Attacks		MNIST		CIFAR-10	
		$\|\|\boldsymbol{\delta}\|\|$	TPR	$\|\|\boldsymbol{\delta}\|\|$	TPR
FGSM	l_∞	0.1	0.968	0.001	0.985
	l_2	2.0	0.872	0.05	0.772
BIM	l_∞	0.1	0.996	0.001	0.912
	EOT	0.1	0.940	0.001	0.967
	l_2	2.0	0.998	0.05	0.826
JSMA	l_∞	0.5	1.0	0.30	0.984
DeepFool	l_2	1.97	1.0	0.31	0.998
	EOT	1.86	1.0	0.32	0.944
CW	l_2	1.96	1.0	0.29	0.987

Table 3: Detection accuracy (TPR@5%FPR) of our approach against various attack forms. The model classifier is trained with the complementary step of noise augmentation.

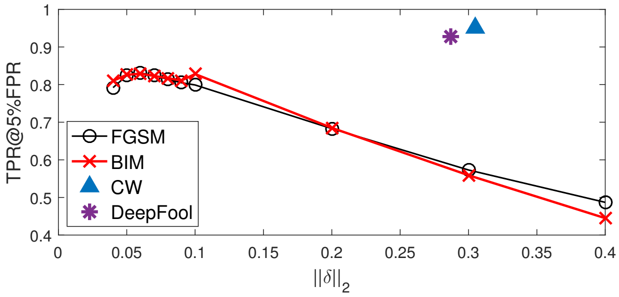


Figure 3: Detection robustness to increasing $||\boldsymbol{\delta}||_2$ over CIFAR-10.

4.3 Adaptive Adversary

In this section, we consider under the threat model of *adaptive adversary* four possible methods to modify the evasion attack over the proposed detector.

Increasing Adversarial Distortion

In this scenario, the adversary may increase the distortion level to attack a pretrained detector at test time. Thus, we conduct experiments to see if our detector is robust to the change. According to Remarks in Section 3.2, our approach can detect adversarial perturbation $\delta \leq \kappa \cdot \sigma$ in l_2 norm given detector parameters σ and m that generate the random perturbations. Fig. 3 plots the detection accuracy w.r.t. an increasing $||\boldsymbol{\delta}||_2$. Here, $\kappa = 2.0537$ for $m = 50$ and $\sigma = 0.05$ for CIFAR-10.

The experimental result shows that our detector is robust for $\delta \leq 0.1$, which coincides with the theoretical result. The detection accuracy then scales down linearly with $||\boldsymbol{\delta}||_2$. To deal with a larger distortion, we may increase the σ value. However, the additive random noise after increasing to some level may decrease the standard classification accuracy albeit enhancing the model with noise augmentation. There seems to be a fundamental trade-off between adversarial robustness and standard accuracy requirements of a model [Su *et al.*, 2018]. In our experiments, installing the robustness to random noise with $\sigma = 0.05$ is at a cost of about 1-3% on normal classification accuracy over the datasets.

Increasing Prediction Confidence

Many previous detection methods failed on high-confident adversarial examples generated by CW attacks [Carlini and Wagner, 2017a]. In Table 3, for example, the CW examples have an average score of 0.73 for their prediction confidence by the CIFAR-10 model. To test the robustness of our detector, we tuned the CW function parameters to generate 2000 new adversarial examples such that the mean of their prediction score is set at 0.97. We then used these CW examples of high confidence to test our detector previously trained on the BIM examples. The detection accuracy drops by about 12% to 0.837, which is quite resilient comparing with other detection methods such as those reported in [Carlini and Wagner, 2017a].

Attacking Randomized Detector

It was suggested that defenses employing randomized input tranformations may be defeated by applying the technique of Expectation of Transformation (EOT) to overcome the problem of stochastic gradients [Athalye *et al.*, 2018]. Here, we demonstrate that the adversarial adaptation is not effective to our detector.

EOT works as follows [Athalye *et al.*, 2018]. Let D randomly transform the input example $\mathbf{x}$ according to some function $t(\cdot)$ sampled from a distribution of transformations T. To attack F with a loss function $J_F(\cdot)$, the adversary computes gradients over the expectation of sampled transformations $\mathrm{E}_{t\sim T} J_F(t(\mathbf{x}))$ instead of $J_F(\mathbf{x})$. In our case, $t(\cdot)$ is the additive random perturbation $\eta_i \sim \mathcal{D}^{\text{noise}}$, for $i = 1, \ldots, m$. Take BIM (l_∞) for example. In each step of iterations, we update the gradient estimation with EOT by

$$g(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m} \nabla_{\mathbf{x}} J_F(\mathbf{x} + \boldsymbol{\eta}_i) \qquad (1)$$

and compute the adversarial example as usual. The results are reported in Table 3. It can be seen that the detector performance is even higher than that without EOT. Similar result can be observed for DeepFool with l_2 norm. This indicates that the EOT estimated gradient is not effective and hence the attack fails in the adaptive setting.

Incorporating Detector Loss

Following the approach in [Carlini and Wagner, 2017a], we modify the CW attack by introducing to its adversarial objective an additional loss term $J_D(\mathbf{x} + \boldsymbol{\delta})$ for penalizing being detected:

$$\min\{||\boldsymbol{\delta}||_p + \alpha \cdot J_F(\mathbf{x} + \boldsymbol{\delta}) + \beta \cdot J_D(\mathbf{x} + \boldsymbol{\delta})\} \qquad (2)$$

where J_F is the loss of F and

$$J_D(\mathbf{x} + \boldsymbol{\delta}) = \max\{0, 1 + D(\mathbf{x} + \boldsymbol{\delta})\}$$

such that $\mathbf{x}+\delta$ aims to fool F as an adversarial example and D as a normal example simultaneously. In practice, we mount this attack in two phases. First, we solve the original CW formulation to obtain $\mathbf{x}^{\text{adv}}$ which typically will be detected by D. Then, we use $\mathbf{x}^{\text{adv}}$ to initialize $\mathbf{x}$ in solving (2).

The above optimization problem is typically solved by gradient descents if all parts are differentiable. However, the detector gradients w.r.t. the input, i.e., $\nabla_{\mathbf{x}} J_D$, cannot be easily estimated in our case. There are *two* obstacles. First, our detector relies on the statistical feature drawn from m random vectors created by model responses to random perturbations $\boldsymbol{\eta}_i \sim \mathcal{D}^{\text{noise}}$ on the input $\mathbf{x}$. Second, the statistical feature is based on quantile discretization. To estimate $\nabla_{\mathbf{x}} J_D$, the adversary has to overcome the stochastic and shattered gradients at the same time. Both components obfuscate the actual gradient information required in the gradient-based optimization method [Athalye *et al.*, 2018].

Alternatively, we try to solve (2) by *gradient checking* where the computation of $\nabla_{\mathbf{x}} J_D$ is approximated with numerical differentiations incident on the input $\mathbf{x}$ for each of its dimensions as

$$g_i(\mathbf{x}) = \frac{J(\mathbf{x} + \Delta \cdot \mathbf{e}_i) - J(\mathbf{x} - \Delta \cdot \mathbf{e}_i)}{2\Delta} \approx \frac{\partial J_D(\mathbf{x})}{\partial \mathbf{x}_i} \qquad (3)$$

where $\mathbf{e}_i$ denotes the i-th elementary basis.

In practice, the coordinate-wise gradient estimation is very computationally expensive as each step of the gradient checking algorithm involves an update of (3) over each pixel of $\mathbf{x}$. Here, we only constructed the adapted attacks on MNIST for test. We found that $J_D(\mathbf{x})$ has a non-negligible expectation of variation (0.018) even when $\Delta = 0$. We consider it due to the stochastic nature of classifier scores in response to the random perturbations in our approach. Therefore, any small Δ (e.g., $< 10^{-4}$) tends to amplify this inherent stochastic variation and cause randomized gradient estimates. Indeed, it turns out that our detector cannot be easily defeated by incorporating the detector loss in this way. The detection accuracy remains as high as 95.6% against the adaptive attack with perfect knowledge of D.

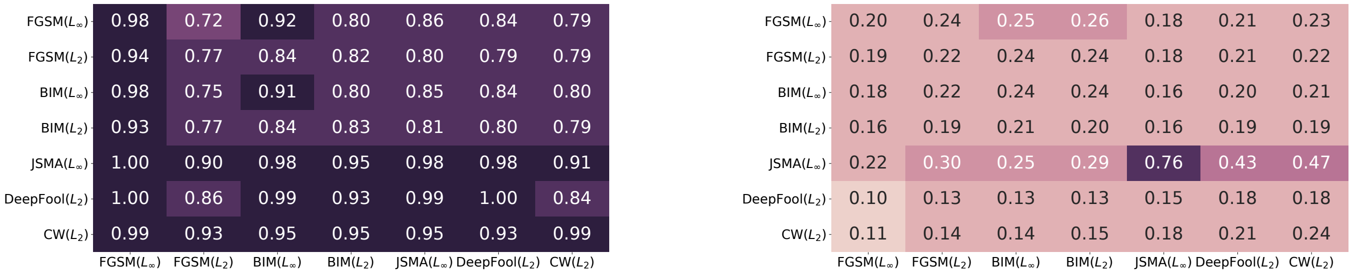
4.4 Transferred Adversary

Generalizability Analysis

Figure 4 shows the generalizability heatmaps by our approach comparing with Detector I. The detectors are trained with one of the attack forms listed in the columns and tested against one another listed in the rows with different norm objectives (49 pairs in total). The detection rate is measured by TPR@5%FPR under the same setting as for running Table 3. Unlike Detector I that performs relatively better on the first-order family, the proposed approach is more robust to transferred examples deliberately constructed by the more complex iterative methods of JSMA, DeepFool and CW attacks. It is also interesting to see that our detector trained with one-step FGSM (l_∞) generalizes particular well against various attack forms generated with different algorithms and norm objectives. This may be related to the previous observation that the plain FGSM has relatively better transferrability [Su *et al.*, 2018].

Black-Box Attacks

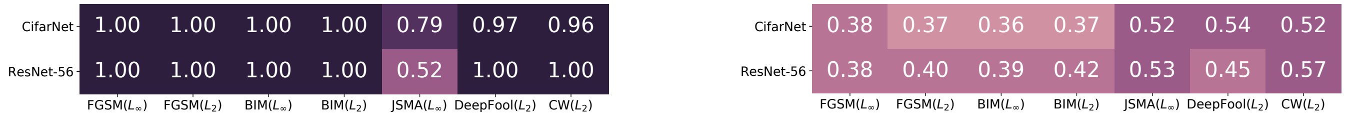
It was observed that adversarial examples tend to transfer better within the same model family, and that models with lower capacity and higher test accuracy are endowed with stronger capability for transfer-based attacks [Su *et al.*, 2018].



(a)

(b)

Figure 4: Generalizability analysis of detection across different attacks: (a) Ours and (b) SafetyNet. Both detectors are trained with one attack (column) and tested by another (row) over CIFAR-10.



(a)

(b)

Figure 5: Black-box detection against transferred attacks across different classifier models: (a) Ours and (b) Feature Squeezing. Our detector was trained with FGSM (l_∞) on ResNet-20 and tested by another attack (column) built on two substitute models, namely CifarNet (upper row) and ResNet-56 (lower row) over CIFAR-10.

Therefore, we test the detection robustness by transferring adversarial examples generated from two substitute models: 1) ResNet-56 with the same architecture but deeper, and 2) CifarNet with a different architecture and lower capacity. For performance evaluation, we used only those transferred examples that can attack successfully before deploying any detection measure to the target model, i.e., ResNet-20 for CIFAR-10. Figure 5 illustrates the detection accuracy (TPR@5%FPR) in heatmaps where the columns are the black-box attacks. Our detector was trained with FGSM (l_∞) and DeepFool (l_2) examples, respectively. We also evaluated Detector II using the threshold values recommended in [Xu *et al.*, 2017]. We note that the transferred attacks often have lower success rates and reduced strength on other target models. Our detector is able to catch the property and achieve high accuracy in detecting such transferred examples in the black-box setting.

5 Discussion

In Section 3.2 and Fig. 3, we apply the Gaussian perturbation to make it easier for theoretical analysis and empirical verifications. Nevertheless, the proposed approach can accommodate other forms of random perturbations. We tested by drawing random perturbations from $\texttt{Uniform}(-0.2, 0.2)$ and $\texttt{Laplace}(0, 0.1)$ respectively without using the complementary step of noise augmentation. Our experimental results on MNIST show that the detection accuracy (TPR@5%FPR) by these *non-Gaussian* perturbations is very close to that obtained by Gaussian perturbations under the attacks shown in Table 3. We conjecture that our analytical bounds still hold for all sub-Gaussian distributions. To accommodate non-Gaussian perturbations in the theoretical analysis, we only require to obtain the tail bound of L_2 norm of the resulting random perturbation vector.

6 Conclusion

We proposed a simple yet highly robust adversarial detection method based on statistical analysis of model responses to the input with additive random perturbations. Our method targets at the inherent constraint on all adversarial examples whose magnitude of the adversarial perturbation is bounded. Accordingly, we provided a theoretical analysis to understand its effectiveness, which matches the empirical results well. Our method has been demonstrated to be resilient to the powerful CW attacks under four possible variations by an adaptive adversary. We have performed extensive experimental evaluations to show that our method is more robust in different settings including transferred adversaries across different target models and generalizes well to unseen attacks even without noise augmentation in the training process. The proposed approach does not rely on specific model architecture nor data distribution, which is a salient property for being model agnostic. Thus, it can be mounted to any target model and possibly work in conjunction with a robust classifier or model-based methods to provide comprehensive protections.

Acknowledgments

The work of Y. Wang was supported in part by the National Natural Science Foundation of China under Grant 61876038 and in part by Dongguan University of Technology under Project KCYKYQD2017003. The work of W. Wang was supported in part by the Australian Research Council under Grant DP170103710 and Grant DP180103411, and in part by D2DCRC under Grant DC25002 and Grant DC25003.

References

[Akhtar *et al.*, 2018] Naveed Akhtar, Jian Liu, and Ajmal Mian. Defense Against Universal Adversarial Perturbations. In *IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 3389–3398, June 2018.

[Athalye *et al.*, 2018] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples. In *Int. Conf. Mach. Learn.*, pages 274–283, Stockholm, Sweden, July 2018.

[Biggio *et al.*, 2013] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion Attacks against Machine Learning at Test Time. In Hendrik Blockeel, Kristian Kersting, Siegfried Nijssen, and Filip Železný, editors, *Mach. Learn. Knowl. Discov. Databases*, pages 387–402, Berlin, Heidelberg, 2013.

[Carlini and Wagner, 2017a] Nicholas Carlini and David Wagner. Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods. In *ACM Work. Artif. Intell. Secur.*, pages 3–14, Dallas, Texas, USA, October 2017.

[Carlini and Wagner, 2017b] Nicholas Carlini and David Wagner. Towards Evaluating the Robustness of Neural Networks. In *IEEE Symp. Secur. Priv.*, pages 39–57, San Jose, CA, USA, May 2017.

[Cisse *et al.*, 2017] Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier. Parseval Networks: Improving Robustness to Adversarial Examples. In *Int. Conf. Mach. Learn.*, pages 854–863, Sydney, Australia, August 2017.

[Fawzi *et al.*, 2016] Alhussein Fawzi, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Robustness of classifiers: from adversarial to random noise. In *Neural Inf. Process. Syst.*, pages 1632–1640, Barcelona, Spain, December 2016.

[Goodfellow *et al.*, 2015] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples. In *Int. Conf. Learn. Represent.*, pages 1–11, San Diego, CA, USA, May 2015.

[Grosse *et al.*, 2017] Kathrin Grosse, Praveen Manoharan, Nicolas Papernot, Michael Backes, and Patrick McDaniel. On the (Statistical) Detection of Adversarial Examples. In *IEEE Int. Conf. Comput. Vis.*, pages 446–454, Venice, Italy, October 2017.

[Guo *et al.*, 2017] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens van der Maaten. Countering Adversarial Images using Input Transformations. In *Int. Conf. Learn. Represent.*, pages 1–12, Vancouver, Canada, October 2017.

[Kurakin *et al.*, 2017] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial Machine Learning at Scale. In *Int. Conf. Learn. Represent.*, pages 1–17, Toulon, France, October 2017.

[Lecuyer *et al.*, 2018] Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified Robustness to Adversarial Examples with Differential Privacy. In *Int. Conf. Learn. Represent.*, arXiv preprint, pages 1–18, Vancouver, BC, Canada, February 2018.

[Lu *et al.*, 2017] Jiajun Lu, Theerasit Issaranon, and David Forsyth. SafetyNet: Detecting and Rejecting Adversarial Examples Robustly. In *IEEE Int. Conf. Comput. Vis.*, pages 446–454, October 2017.

[Ma *et al.*, 2018] Xingjun Ma, Bo Li, Yisen Wang, Sarah M. Erfani, Sudanthi Wijewickrema, Grant Schoenebeck, Dawn Song, Michael E. Houle, and James Bailey. Characterizing Adversarial Subspaces Using Local Intrinsic Dimensionality. In *Int. Conf. Learn. Represent.*, pages 1–15, Vancouver, Canada, February 2018.

[Madry *et al.*, 2018] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. In *Int. Conf. Learn. Represent.*, arXiv preprint, pages 1–27, Vancouver, Canada, February 2018.

[Moosavi-Dezfooli *et al.*, 2016] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. DeepFool: a simple and accurate method to fool deep neural networks. In *IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 2574–2582, June 2016.

[Papernot *et al.*, 2016] Nicolas Papernot, Patrick Mcdaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *IEEE Eur. Symp. Secur. Priv.*, pages 372–387, April 2016.

[Papernot *et al.*, 2018] Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael Wellman. Towards the Science of Security and Privacy in Machine Learning. In *IEEE Eur. Symp. Secur. Priv.*, pages 1–19, London, UK, September 2018.

[Su *et al.*, 2018] Dong Su, Huan Zhang, Hongge Chen, Jinfeng Yi, Pin-Yu Chen, and Yupeng Gao. Is Robustness the Cost of Accuracy? A Comprehensive Study on the Robustness of 18 Deep Image Classification Models. In *Eur. Conf. Comput. Vis.*, pages 644–661, Munich, Germany, September 2018. Springer, Cham.

[Szegedy *et al.*, 2014] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *Int. Conf. Learn. Represent.*, pages 1–10, Banff, Canada, April 2014.

[Tramèr *et al.*, 2018] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Dan Boneh, and Patrick McDaniel. Ensemble Adversarial Training: Attacks and Defenses. In *Int. Conf. Learn. Represent.*, pages 1–20, Vancouver, Canada, May 2018.

[Xu *et al.*, 2017] Weilin Xu, David Evans, and Yanjun Qi. Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks. In *Netw. Distrib. Syst. Secur. Symp.*, pages 1–15, San Diego, CA, USA, February 2017.