Fortify the Guardian, Not the Treasure: Resilient Adversarial Detectors

Anonymous authors
Paper under double-blind review

Abstract

This paper presents RADAR—**R**obust **A**dversarial **D**etection via **A**dversarial **R**etraining—an approach designed to enhance the robustness of adversarial detectors against *adaptive attacks*, while maintaining classifier performance. An adaptive attack is one where the attacker is aware of the defenses and adapts their strategy accordingly. Our proposed method leverages adversarial training to reinforce the ability to detect attacks, without compromising clean accuracy. During the training phase, we integrate into the dataset adversarial examples, which were optimized to fool *both* the classifier *and* the adversarial detector, enabling the adversarial detector to learn and adapt to potential attack scenarios. Experimental evaluations on the CIFAR-10 and SVHN datasets demonstrate that our proposed algorithm significantly improves a detector's ability to accurately identify adaptive adversarial attacks—without sacrificing clean accuracy.

1 Introduction

Rapid advances in deep learning systems and their applications in various fields such as finance, healthcare, and transportation, have greatly enhanced human capabilities (Huang et al., 2020; Miotto et al., 2018; Chen et al., 2024a; Shamshirband et al., 2021; Lv et al., 2020). However, these systems continue to be vulnerable to adversarial attacks, where unintentional or intentionally designed inputs—known as *adversarial examples*—can mislead the decision-making process (Goodfellow et al., 2014; Szegedy et al., 2013; Tamam et al., 2023; Chen et al., 2024b; Lapid et al., 2023; Carlini & Wagner, 2017b; Lapid & Sipper, 2023b; Andriushchenko et al., 2020; Lapid et al., 2022; Lapid & Sipper, 2023a). Such adversarial attacks pose severe threats both to the usage and trustworthiness of deep learning technologies in critical applications (Finlayson et al., 2019; Liu et al., 2023).

Extensive research efforts have been dedicated to developing robust machine learning classifiers that are resilient to adversarial attacks. One popular and effective way relies on *adversarial training*, which involves integrating adversarial examples within the training process, to improve the classifier's resistance to attack (Madry et al., 2017; Zhao et al., 2022). However, the emergence of adversarial detectors designed to identify malicious inputs (Lu et al., 2017; Grosse et al., 2017a; Gong & Wang, 2023; Lust & Condurache, 2020; Metzen et al., 2016) has opened new avenues for enhancing or thwarting the security of AI systems.

This paper presents a novel approach to adversarially train detectors, combining the strengths both of classifier robustness and detector sensitivity to adversarial examples.

Adversarial detectors (Figure 1), which aim to distinguish adversarial examples from benign ones, have gained momentum recently, but their robustness is unclear. The purpose of adversarial detection is to enhance the robustness of machine learning systems by identifying and mitigating adversarial attacks—thereby preserving the reliability and integrity of the classifiers in critical applications.

Several studies have found that these detectors—even when combined with robust classifiers that have undergone adversarial training—can be fooled by designing tailored attacks; this engenders a cat-and-mouse cycle of attacking and defending (Carlini & Wagner, 2017a; Grosse et al., 2017b). It is therefore critical to

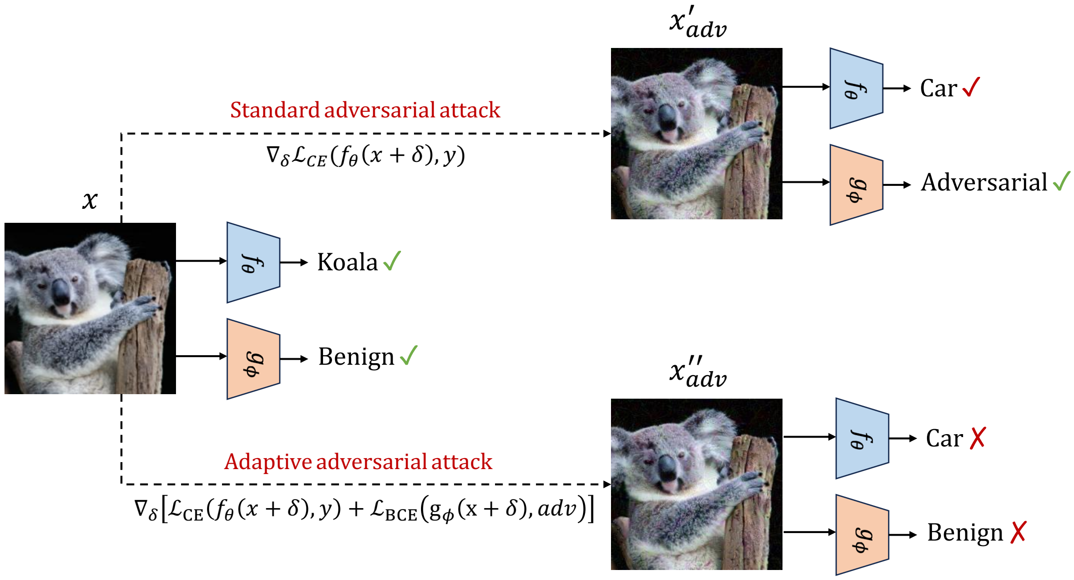

Figure 1: General scheme of adversarial attacks. x: original image. $x'_{\texttt{adv}}$: standard adversarial attack. $x''_{\texttt{adv}}$: adaptive adversarial attack, targeting both f_θ (classifier) and g_ϕ (detector). The attacker's goal is to fool the classifier into misclassifying the image *and* at the same time fool the detector into reporting the attack as benign (i.e., fail to recognize an attack).

thoroughly assess the robustness of adversarial detectors and ensure that they are able to withstand adaptive adversarial attacks.

The attacker's goal is to fool the classifier into misclassifying the image *and* at the same time fool the the detector into reporting the attack as benign (i.e., fail to recognize an attack). The attacker, in essence, carries out a two-pronged attack, targeting both the "treasure" and the treasure's "guardian".

An *adaptive* adversarial attack is *aware of the defenses*—and attempts to bypass them. To our knowledge, research addressing such attacks is lacking.

We thus face a crucial problem: How can adversarial detectors be made robust? This question underscores the need for a comprehensive understanding of the interplay between adversarial-training strategies and the unique characteristics of adversarial detectors.

We hypothesize that adversarially training adversarial detectors will lead to their improvement both in terms of robustness and accuracy, thus creating a more-resilient defense mechanism against sophisticated adversarial attacks. This supposition forms the basis of our investigation and serves as a guiding principle for the experiments and analyses presented in this paper.

There are two major motivations to our work herein:

1. Contemporary adversarial detectors are predominantly evaluated within constrained threat models, assuming attackers lack information about the adversarial detector itself. However, this approach likely does not align with real-world scenarios, where sophisticated adversaries possess partial knowledge of the defense mechanisms.

We believe it is necessary to introduce a more-realistic evaluation paradigm that considers the potential information available to an attacker. Such a paradigm will improve—perhaps significantly—the reliability of adversarial-detection mechanisms.

2. Our fundamental assumption is that adversarial training of the *adversarial detector*—in isolation—has the potential to render the resultant system significantly more challenging for adversaries to thwart. Unlike conventional approaches, where increasing the robustness of the *classifier* often leads to decreased accuracy, our approach focuses solely on enhancing the capabilities of the detector, thereby avoiding such trade-offs. By bolstering the detector through adversarial training, we introduce a distinct tier of defense that adversaries must contend with, amplifying the overall adversarial resilience of the system.

This paper focuses on the adversarial training of a defensive adversarial detector to improve its resilience against robust and sophisticated adversarial attacks. **We suggest a paradigm shift: instead of defending the classifier—strengthen the adversarial detector**. Thus, we ask:

(Q) *How can adversarial detectors be made robust?*

To our knowledge, problem **(Q)** remains open in the literature. Our contributions are:

- Paradigm shift: Instead of robustifying the classifier, we robustify the adversarial detector.
- We propose `RADAR`, a novel adversarial training approach to improve the resilience of adversarial detectors.
- Our study conducts extensive evaluations on two datasets and six detection architectures, providing insights into the generalizability and efficacy of adversarial training across varied data distributions and model architectures.

The next section describes previous work on adversarial training. Section 3 delineates our methodology, followed by the experimental framework in Section 4. We describe our results in Section 5, ending with conclusions in Section 6.

2 Previous Work

The foundation of adversarial training was laid by Madry et al. (2017), who demonstrated its effectiveness on the MNIST (LeCun et al., 1998) and CIFAR-10 (Krizhevsky, 2009) datasets. Their approach used Projected Gradient Descent (PGD) attacks to generate adversarial examples and incorporate them into the training process alongside clean (unattacked) data.

Subsequent work by Kurakin et al. (2016) explored diverse adversarial training methods, highlighting the trade-off between robustness and accuracy. These early pioneering works laid the groundwork for the field of adversarial training and showcased its potential for improving classifier robustness.

Adversarial training has evolved significantly since its inception. Zhang et al. (2017) introduced MixUp training, probabilistically blending clean and adversarial samples during training, achieving robustness without significant accuracy loss.

Tramer & Boneh (2019) investigated multi-attack adversarial training, aiming for broader robustness against diverse attack types.

A number of works explored formal-verification frameworks to mathematically certify classifier robustness against specific attacks, demonstrating provably robust defense mechanisms (Cohen et al., 2019; Lecuyer et al., 2019; Li et al., 2023; Singh et al., 2018).

Zhang et al. (2019) provided theoretical insights into the robustness of adversarially trained models, offering a deeper understanding of the underlying mechanisms and limitations. Their work elucidated the trade-offs between robustness, expressiveness, and generalization in adversarial training paradigms, contributing to a more comprehensive theoretical framework for analyzing and designing robust classifiers.

In the pursuit of improving robustness against adversarial examples, Xie et al. (2019) proposed a novel approach using feature denoising. Their work identified noise within the extracted features as a vulnerability during adversarial training. To address this, they incorporated denoising blocks within the network architecture of a convolutional neural network (CNN). These blocks leverage techniques such as non-local means filters to remove the adversarial noise, leading to cleaner and more-robust features. Their approach achieved significant improvements in adversarial robustness on benchmark datasets, when compared to baseline adversarially trained models. This work highlights feature denoising as a promising defense mechanism that complements existing adversarial training techniques.

Building on the limitations of standard adversarial training, requiring a large amount of labeled data, Carmon et al. (2019) explored leveraging unlabeled data for improved robustness. Their work demonstrates that incorporating unlabeled data through semi-supervised learning techniques, such as self-training, can significantly enhance adversarial robustness. This approach bridges the gap in sample complexity, achieving high robust accuracy with the same amount of labeled data needed for standard accuracy. Their findings were validated empirically on datasets such as CIFAR-10, achieving state-of-the-art robustness against various adversarial attacks. This research highlights the potential of semi-supervised learning as a cost-effective strategy to improve adversarial training and achieve robust models.

Despite its successes, adversarial training faces several challenges. Transferability of adversarial examples remains a concern (Cheng et al., 2019; Dong et al., 2021), because attacks often lack effectiveness across different classifiers and architectures. The computational cost of generating and training on adversarial examples can be significant, especially for large models and datasets. Moreover, adversarial training can lead to a reduction in clean-data accuracy, requiring effective optimization strategies to mitigate this trade-off (Yang et al., 2020). Additionally, carefully crafted evasive attacks can sometimes bypass adversarially trained classifiers, highlighting the need for further research and diversification of defense mechanisms.

There are but a few examples of works that compare adversarial attacks that take into account full knowledge of the defense, i.e., adaptive attacks. He et al. (2022) showed promising results, at the cost of multiple model inferences coupled with multiple augmented neighborhood images. Yang et al. (2022) presented a novel encoder-generator architecture that compared the generated image label with the original image label. We believe this was a step in the right direction, yet with a computationally expensive generator that had difficulties differentiating semantically close classes. Klingner et al. (2022) demonstrated a detection method based on edge extraction of features such as depth estimation and semantic segmentation, and comparison to natural image edges based on the SSIM metric. We believe the underlying assumption of adversarial examples containing abnormal edges might not hold in unnatural settings.

3 Methodology

Before delineating the methodology we note that the overall framework comprises three distinct components:

1. **Classifier**: Which is not modified (i.e., no learning).
2. **Detector**: With a "standard" loss function (to be discussed).
3. **Attacker**: With an adaptive loss function (to be discussed).

3.1 Threat Model

We assume a strong adversary, with full access both to the classifier f and the adversarial detector g. The attacker possesses comprehensive knowledge of the internal workings, parameters, and architecture of both models.

The attacker has two goals: **1)** To manipulate the classifier's (f) class prediction y such that it outputs a class different from the correct class, i.e., $f(x) \neq y$. Additionally, the adversarial noise δ should be sufficiently small, constrained through an upper bound ϵ on the l_p norm. This problem can be formulated as:

$$\underset{k=0,1,\ldots K-1}{\arg\max}\ f_k(x+\delta) \neq y,\ (x,y) \in \mathcal{D},\ \text{ s.t. } \|\delta\|_p \leq \epsilon. \tag{1}$$

We can solve this problem using the following optimization objective:

$$\underset{\delta:\|\delta\|_p \leq \epsilon}{\arg\max}\, \mathcal{L}_{\text{CE}}(f_\theta(x+\delta), y), \tag{2}$$

where $\mathcal{L}_{\text{CE}}$ is the cross-entropy loss.

The attacker's second goal is: **2)** To cause the detector g to predict that the image is benign, i.e., $g(x) \neq \texttt{adv}$, while also maintaining the l_p norm constraint:

$$\underset{k=\texttt{ben},\texttt{adv}}{\arg\max}\, g_k(x+\delta) \neq \texttt{adv},\ x \in \mathcal{D},\ \text{s.t. } \|\delta\|_p \leq \epsilon. \tag{3}$$

We can solve this problem using the following optimization objective:

$$\underset{\delta:\|\delta\|_p \leq \epsilon}{\arg\max}\, \mathcal{L}_{\text{BCE}}(g_\phi(x+\delta), \texttt{adv}), \tag{4}$$

where $\mathcal{L}_{\text{BCE}}$ is the binary cross-entropy loss and ϵ is the allowed perturbation norm.

Combining the two goals, the attacker's overall goal is defined as:

$$\underset{\delta:\|\delta\|_p \leq \epsilon}{\arg\max}\, \mathcal{L}_{\text{CE}}(f_\theta(x+\delta), y) + \mathcal{L}_{\text{BCE}}(g_\phi(x+\delta), \texttt{adv}). \tag{5}$$

3.2 Problem Definition

We now formalize the problem of enhancing the robustness of an adversarial detector (g) through adversarial training. The setup involves a classifier (f) and an adversarial detector (g), both initially trained on a clean dataset $\mathcal{D}$ containing pairs (x, y), where x is an input data point and y is the ground-truth label. We denote by θ and ϕ the weight parameters of the classifier and the detector, respectively. Note that $\mathcal{D}$ is the clean dataset—adversarial examples are represented by $x + \delta$.

Adversarial training is formulated as a minimax game:

$$\min_\theta \left\{ \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\delta:\|\delta\|_p \leq \epsilon} \mathcal{L}_{\text{CE}}(f_\theta(x+\delta), y) \right] \right\}. \tag{6}$$

This minimax game pits the classifier against an adversary: the adversary maximizes the classifier's loss on perturbed inputs, while the classifier minimizes that loss, across all such perturbations, ultimately becoming robust to attacks.

Our main objective is to improve the robustness of detector g using adversarial training, by iteratively updating ϕ based on adversarial examples. Formally:

$$\min_\phi \left\{ \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\delta:\|\delta\|_p \leq \epsilon} \mathcal{L}_{\text{BCE}}(g_\phi(x+\delta), \texttt{adv}) \right] \right\}. \tag{7}$$

As with the classifier, we have a minimax game for the detector.

However, the above equation does not take into account the classifier, f. So we added the classifier outputs to our (almost) final objective:

$$\min_\phi \left\{ \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\delta:\|\delta\|_p \leq \epsilon} (\mathcal{L}_{\text{CE}}(f_\theta(x+\delta), y) + \mathcal{L}_{\text{BCE}}(g_\phi(x+\delta), \texttt{adv})) \right] \right\}. \tag{8}$$

But a potential conflict manifests with this objective. While gradient descent allows for simultaneous maximization of the cross-entropy loss ($\mathcal{L}_{\text{CE}}$) and the binary cross-entropy loss ($\mathcal{L}_{\text{BCE}}$), they might contradict each other, adopting different optimization strategies. To address this challenge we employ the selective and orthogonal approaches proposed Bryniarski et al. (2021).

Selective Projected Gradient Descent (SPGD) (Bryniarski et al., 2021) optimizes only with respect to a constraint that has not been satisfied yet—while not optimizing against a constraint that has been optimized. The loss function used by SPGD is:

$$\mathcal{L}_{\text{CE}}(f_\theta(x+\delta), y) \cdot \mathbb{1}[f_\theta(x) = y] + \mathcal{L}_{\text{BCE}}(g_\phi(x+\delta), \texttt{adv}) \cdot \mathbb{1}[g_\phi(x+\delta) = \texttt{adv}]. \tag{9}$$

Rather than minimize a convex combination of the two loss functions, the idea is to optimize either the left-hand side of the equation, if the classifier's prediction is still y, meaning the optimization hasn't completed; or optimize the right-hand side of the equation, if the detector's prediction is still `adv`.

This approach ensures that updates consistently enhance the loss on either the classifier or the adversarial detector, with the attacker attempting to "push" the classifier away from the correct class y, and also "push" the detector away from raising the adversarial "flag".

Orthogonal Projected Gradient Descent (OPGD) (Bryniarski et al., 2021) focuses on modifying the update direction to ensure it remains orthogonal to the constraints imposed by previously satisfied objectives. This orthogonal projection ensures that the update steers the input towards the adversarial objective without violating the constraints imposed by the classifier's decision boundary, allowing for the creation of adversarial examples that bypass the detector. Thus, the update rule is slightly different:

$$\mathcal{L}_{\text{update}}(x, y) = \begin{cases} \nabla\mathcal{L}_{\text{CE}}(f_\theta(x+\delta), y) - \text{proj}_{\nabla\mathcal{L}_{\text{CE}}(f_\theta(x+\delta),y)} \nabla\mathcal{L}_{\text{BCE}}(g_\phi(x+\delta), \texttt{adv}) & \text{if} \quad f_\theta(x) = y \\ \nabla\mathcal{L}_{\text{BCE}}(g_\phi(x+\delta), \texttt{adv}) - \text{proj}_{\nabla\mathcal{L}_{\text{BCE}}(g_\phi(x+\delta),\texttt{adv})} \nabla\mathcal{L}_{\text{CE}}(f_\theta(x+\delta), y) & \text{if} \quad g_\phi(x) = \texttt{adv}. \end{cases} \tag{10}$$

In essence we have set up a minimax game: the attacker wants to maximize loss with respect to the image, while the defender wants to minimize loss with respect to the detector's parameters.

Since optimization is customarily viewed as minimizing a loss function we will consider that the attacker uses minimization instead of maximization. A failed attack would thus be observed though a large loss value (which we will indeed observe in Section 5).

All aforementioned objectives are intractable and thus approximated using iterative gradient-based attacks, such as PGD. Algorithm 1 shows the pseudocode of our method for adversarially training an adversarial detector.

4 Experimental Framework

Datasets and models. We used the VGG-{11,13,16} and ResNet-{18, 34, 50} architectures both for classification and for adversarial detection. We modified the classification head of the detector, $C \in \mathbb{R}^{K \times e}$, where K is the number of classes and e is the embedding-space size, to $C \in \mathbb{R}^{2 \times e}$, for binary classification (benign/adversarial), and added a sigmoid transformation at the end.

We used the definition of attack success rate presented by Bryniarski et al. (2021) as part of their evaluation methodology. Attack efficacy is measured through Attack Success Rate at N, SR@N: proportion of deliberate attacks achieving their objectives subject to the condition that the defense's false-positive rate is configured to N%. The underlying motivation is that in real-life scenarios we must strike a delicate balance between security and precision. A 5% false positive is acceptable, while extreme cases might even use 50%. Our results proved excellent so we set N to a low 5%, i.e., SR@5.

Throughout the experiments, the allowed perturbation was constrained by an L-infinity norm, denoted as $\|\cdot\|_\infty$, with a maximum magnitude set to $\epsilon = 16/255$. We employed an adversarial attack strategy, specifically utilizing 100 iterations of PGD with a step size parameter α set to 0.03. This approach was employed to generate adversarial instances and assess the resilience of the classifier under scrutiny. We then evaluated the classifiers on the attacked test set—the results are delineated in Table 1.

Algorithm 1: `RADAR`

Input: dataset $\mathcal{D}$, classifier f_θ, detector, g_ϕ, detector's loss function $\mathcal{L}_{\text{BCE}}$, classifier's loss function $\mathcal{L}_{\text{CE}}$, batch size B, step size α, epsilon ϵ, number of steps I, number of epochs E, learning rate η

Output: adversarially trained g_ϕ

Set f to eval mode

for $e = 1, ..., E$ **do**

  for $X_{ben} \subseteq \mathcal{D}$, *s.t.* $|X_{ben}| = B$ **do**

    Set g to eval mode

    Compute adversarial examples $X_{\texttt{adv}}$:

$$X_{\texttt{adv}} \leftarrow \text{OPGD/SPGD}(f_\theta, g_\phi, X_{\texttt{ben}}, \epsilon, \alpha, I, \mathcal{L}_{\text{BCE}}, \mathcal{L}_{\text{CE}})$$

    Form a new batch:

$$X_{\text{mixed}} \leftarrow X_{\texttt{ben}} \oplus X_{\text{adv}}$$
$$Y_{\text{mixed}} \leftarrow \{\texttt{ben}\}_{i=1}^{|B|} \oplus \{\texttt{adv}\}_{i=1}^{|B|}$$

    Set g to train mode

    Compute average loss gradient for X_{mixed}:

$$\nabla_{\text{mixed}} \leftarrow \frac{1}{2|B|} \sum_{j=1}^{2|B|} \nabla_\phi \mathcal{L}_{\text{BCE}}(x_{\text{mixed}}, y_{\text{mixed}}; \phi)$$

    Update parameters:

$$\phi \leftarrow \phi + \eta \nabla_{\text{mixed}}$$

  end

end

Table 1: Performance (accuracy percentage) of original classifiers both on clean and adversarial, PGD-perturbed test sets.

Dataset	Classifier	Accuracy (`ben`)	Accuracy (`adv`)
CIFAR-10	VGG-11	92.40	0.35
	VGG-13	94.21	0.19
	VGG-16	94.00	5.95
	ResNet-18	92.60	0.14
	ResNet-34	93.03	0.21
	ResNet-50	93.43	0.26
SVHN	VGG-11	93.96	0.00
	VGG-13	94.85	0.01
	VGG-16	94.95	0.48
	ResNet-18	94.88	0.02
	ResNet-34	94.84	0.11
	ResNet-50	94.33	0.10

Afterwards, we split the training data into training (70%) and validation sets (30%). We trained 3 VGG-based and 3 ResNet-based adversarial detectors for 20 epochs, using the Adam optimizer (with default β_1 and β_2 values), with a learning rate of $1e-4$, `CosineAnnealing` learning-rate scheduler with $T_{\max} = 10$, and a batch size of 256. We tested the adversarial detectors on the test set that was comprised of one half clean images and one half attacked images: They all performed almost perfectly.

After the initial clean training phase we performed adversarial finetuning on the adversarial detectors. This involved attacking them with an adaptive adversarial attack, e.g., OPGD. We ran the adversarial finetuning for 20 epochs, using the same optimizer, `ReduceLROnPlateau` learning rate scheduler with patience parameter set to 3, and batch size of 256. Then, we tested on a test set—with equal distribution of clean images and OPGD-attacked images.

5 Results

Initially, we conducted an assessment of classifier performance following the generation of adversarial perturbations employing the PGD method. The outcomes of this evaluation are delineated in Table 1, encompassing classifier performance metrics both on clean and perturbed test sets. It is evident that the classifiers exhibit considerable susceptibility to adversarial manipulations, with the crafted attacks achieving near-perfect efficacy in deceiving the classifiers.

After training involving the use of PGD-generated adversarial images, Table 2 shows the performance outcomes of the adversarial detectors, as assessed through the ROC-AUC metric, before the deployment of `RADAR`.

Table 2: **Without `RADAR`**: Performance of detectors on several classifiers and datasets. In this and subsequent tables, boldface marks best performance. Note: for ROC AUC, higher is better.

		CIFAR-10			SVHN		
Detector	**$\mathbf{AUC_{Avg.}}$**	AUC_{PGD}	AUC_{OPGD}	AUC_{SPGD}	AUC_{PGD}	AUC_{OPGD}	AUC_{SPGD}
VGG-11	0.33	1.00	0.00	0.00	1.00	0.00	0.00
VGG-13	0.33	1.00	0.00	0.00	1.00	0.00	0.00
VGG-16	**0.53**	1.00	0.61	0.59	1.00	0.00	0.00
ResNet-18	0.33	1.00	0.00	0.00	1.00	0.00	0.00
ResNet-34	0.33	1.00	0.00	0.00	1.00	0.00	0.00
ResNet-50	0.39	1.00	0.15	0.14	1.00	0.03	0.05
Avg.	0.37	1.00	0.13	0.12	1.00	0.00	0.00

Table 3 presents results regarding the adversarial detectors' performance, based on the SR@5 metric, prior to using `RADAR`.

Table 3: **Without `RADAR`**: Performance of detectors on several classifiers and datasets. VGG-16 is the most robust detector in terms of SR@5. Note: for SR@5, lower is better.

		CIFAR-10		SVHN	
Detector	**$\mathbf{SR@5_{Avg.}}$**	$SR@5_{OPGD}$	$SR@5_{SPGD}$	$SR@5_{OPGD}$	$SR@5_{SPGD}$
VGG-11	0.97	0.97	0.97	0.98	0.99
VGG-13	0.98	0.97	0.99	0.99	0.99
VGG-16	**0.68**	0.37	0.40	0.99	0.99
ResNet-18	0.97	0.96	0.97	0.99	0.99
ResNet-34	0.97	0.96	0.97	0.99	0.99
ResNet-50	0.88	0.81	0.83	0.96	0.94
Avg.	0.90	0.84	0.85	0.98	0.98

We then deployed `RADAR`. The resultant outcomes pertaining to the efficacy of the adversarial detectors, as measured by the ROC-AUC metric, after the integration of the `RADAR` framework, are delineated in Table 4.

Table 4: **With `RADAR`**: Performance of employing `RADAR` with OPGD, on several classifiers and datasets.

		CIFAR-10			SVHN		
Detector	**$\mathbf{AUC_{Avg.}}$**	AUC_{PGD}	AUC_{OPGD}	AUC_{SPGD}	AUC_{PGD}	AUC_{OPGD}	AUC_{SPGD}
VGG-11	0.98	0.99	0.95	0.94	1.00	1.00	1.00
VGG-13	**0.99**	0.99	0.99	0.99	1.00	1.00	1.00
VGG-16	**0.99**	1.00	0.99	0.99	1.00	1.00	1.00
ResNet-18	0.98	0.99	0.96	0.96	1.00	0.99	1.00
ResNet-34	**0.99**	0.99	0.99	1.00	1.00	1.00	1.00
ResNet-50	**0.99**	0.99	0.99	0.99	1.00	1.00	0.99
Avg.	0.99	0.99	0.98	0.98	1.00	0.99	0.99

Table 5 presents results regarding the adversarial detectors' performance, based on the SR@5 metric, upon deploying `RADAR`.

Table 5: **With** `RADAR`: Performance of detectors on several classifiers and datasets.

		CIFAR-10		SVHN	
Detector	**SR@5$_{\text{Avg.}}$**	SR@5$_{\text{OPGD}}$	SR@5$_{\text{SPGD}}$	SR@5$_{\text{OPGD}}$	SR@5$_{\text{SPGD}}$
VGG-11	0.02	0.04	0.06	0.00	0.00
VGG-13	**0.00**	0.00	0.00	0.00	0.00
VGG-16	**0.00**	0.00	0.00	0.00	0.00
ResNet-18	0.02	0.05	0.05	0.00	0.00
ResNet-34	**0.00**	0.00	0.00	0.00	0.00
ResNet-50	**0.00**	0.00	0.00	0.00	0.00
Avg.	0.00	0.01	0.02	0.00	0.00

Table 6 and Table 7 show the generalization performance of `RADAR`-trained detectors on classifiers they were not trained on.

Table 6: Generalization performance of adversarially trained detectors, trained on CIFAR-10, on different classifiers. A value represents the ROC AUC of the respective detector/classifier pair, for OPGD and SPGD, respectively. For example, the two top-right values represent the ROC AUC for OPGD and SPGD, of a ResNet50 detector model, trained on a ResNet50 classifier model, with its performance then evaluated on attacks involving a VGG-11 classifier model. **Clf**: Classifier model. **Det**: Adversarial detector model.

Clf \ **Det**	VGG-11	VGG-13	VGG-16	ResNet-18	ResNet-34	ResNet-50
VGG-11	0.95, 0.94	0.99, 0.99	0.99, 0.99	0.96, 0.99	0.99, 0.99	0.99, 0.99
VGG-13	0.95, 1.00	0.99, 0.99	0.99, 0.99	0.96, 0.99	0.99, 0.99	0.99, 0.99
VGG-16	0.95, 0.95	0.99, 0.99	0.99, 0.99	0.96, 0.99	0.99, 0.99	0.99, 0.99
ResNet-18	0.95, 0.94	0.99, 0.99	0.99, 0.99	0.96, 0.96	0.99, 0.99	0.99, 0.99
ResNet-34	0.95, 0.94	0.99, 0.99	0.99, 0.99	0.96, 0.99	0.99, 1.00	0.99, 0.99
ResNet-50	0.95, 0.94	0.99, 0.99	0.99, 0.99	0.96, 0.99	0.99, 0.99	1.00, 0.99

Table 7: Generalization performance of adversarially trained detectors, trained on SVHN, on different classifiers. A cell value represents the ROC AUC of the respective detector/classifier pair. **Clf**: Classifier model. **Det**: Adversarial detector model.

Clf \ **Det**	VGG-11	VGG-13	VGG-16	ResNet-18	ResNet-34	ResNet-50
VGG-11	1.00, 1.00	1.00, 1.00	1.00, 1.00	0.99, 0.99	1.00, 1.00	1.00, 0.99
VGG-13	1.00, 1.00	1.00, 1.00	1.00, 1.00	0.99, 0.99	1.00, 1.00	1.00, 0.99
VGG-16	1.00, 1.00	1.00, 1.00	1.00, 1.00	0.99, 0.99	1.00, 1.00	1.00, 0.99
ResNet-18	1.00, 1.00	1.00, 1.00	1.00, 1.00	0.99, 1.00	1.00, 1.00	1.00, 0.99
ResNet-34	1.00, 1.00	1.00, 1.00	1.00, 1.00	0.99, 0.99	1.00, 1.00	1.00, 0.99
ResNet-50	1.00, 1.00	1.00, 1.00	1.00, 1.00	0.99, 0.99	1.00, 1.00	1.00, 0.99

A notable enhancement is observed across all detectors with respect to ROC-AUC and SR@5. Moreover, our findings suggest that adversarial training did not optimize solely for adaptability to specific adversarial techniques, but also demonstrated efficacy against conventional PGD attacks.

Before the incorporation of adversarial training, the optimization process exhibited a trend of rapid convergence towards zero loss within a few iterations, as can be seen in the top rows of Figure 2, Figure 3, Figure 4, and Figure 5. Once we integrated adversarial training into the detector, we observed a distinct shift in the optimization behavior.

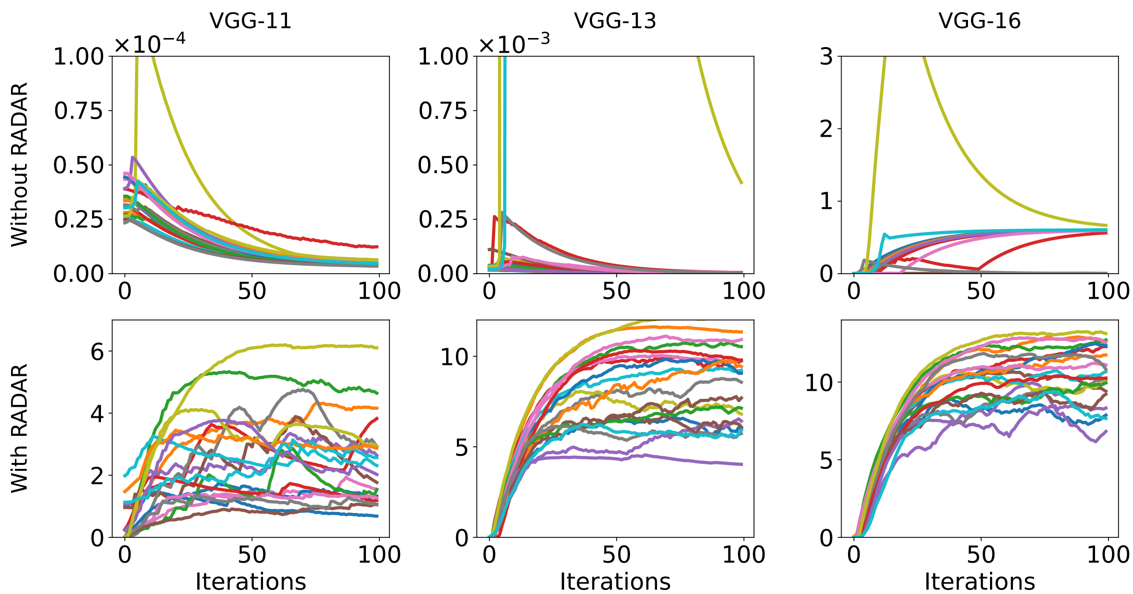

Figure 2: CIFAR-10. Binary cross-entropy loss metrics, from the point of view of an attacker, are herein presented in the context of crafting an adversarial instance from the test set, using VGG-11 (left), VGG-13 (middle), and VGG-16 (right). These plots illustrate the progression of loss over 20 different images of orthogonal projected gradient descent (OPGD), with the main goal being to minimize the loss. Top: Prior to adversarial training, the loss converges to zero after a small number of iterations. Bottom: After adversarial training, the incurred losses are significantly higher by orders of magnitude (note the difference in scales), compared to those observed in their standard counterparts. This shows that the detector is now resilient, i.e., far harder to fool.

Upon deploying `RADAR` the optimization process displayed a tendency to plateau after a small number of iterations, with loss values typically higher by orders of magnitude, as compared to those observed prior to deployment. This plateau phase persisted across diverse experimental settings and datasets, indicating a fundamental change in the optimization landscape, induced by `RADAR`.

This behavior can be seen as a robustness-enhancing effect, because the detector appears to resist rapid convergence towards trivial solutions, thereby enhancing its generalization capabilities. This showcases the efficacy of our method in bolstering the defenses of detection systems against adversarial threats—without sacrificing clean accuracy.

6 Conclusions

We presented a paradigmatic shift in the approach to adversarial training of deep neural networks, transitioning from fortifying classifiers to fortifying networks dedicated to adversarial detection. Our findings illuminate the prospective capacity to endow deep neural networks with resilience against adversarial attacks. Rigorous empirical inquiries substantiate the efficacy of our developed adversarial training methodology.

There is no trade-off between clean and adversarial classification because the classifier is not modified.

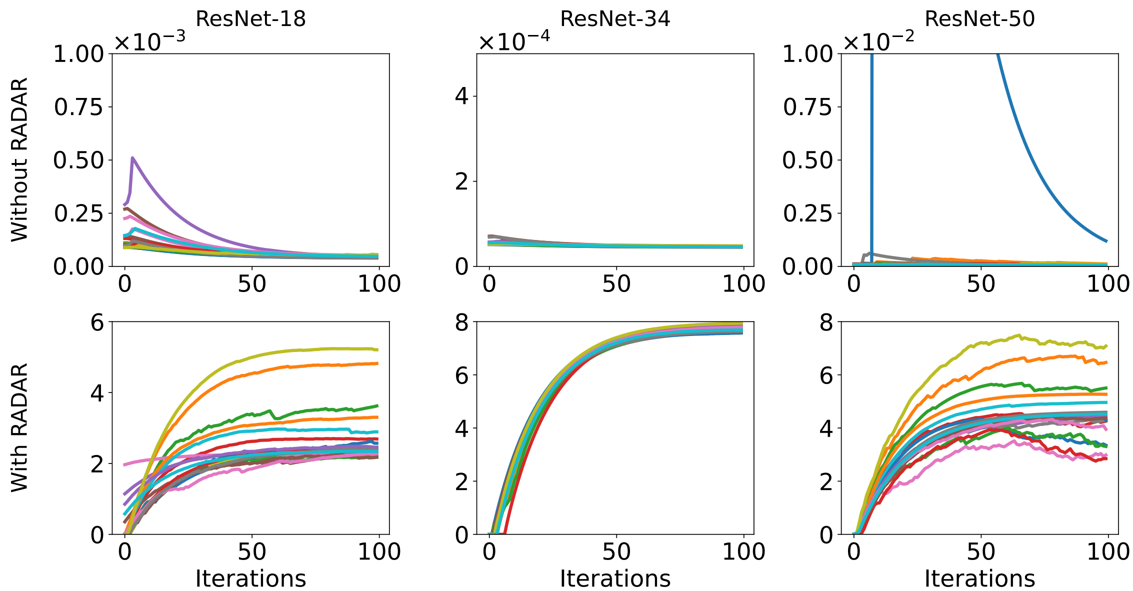

Figure 3: CIFAR10. Binary cross-entropy loss metrics, from the point of view of an attacker, are herein presented in the context of crafting an adversarial instance from the test set using ResNet-18 (left), ResNet-34 (middle), and ResNet-50 (right).

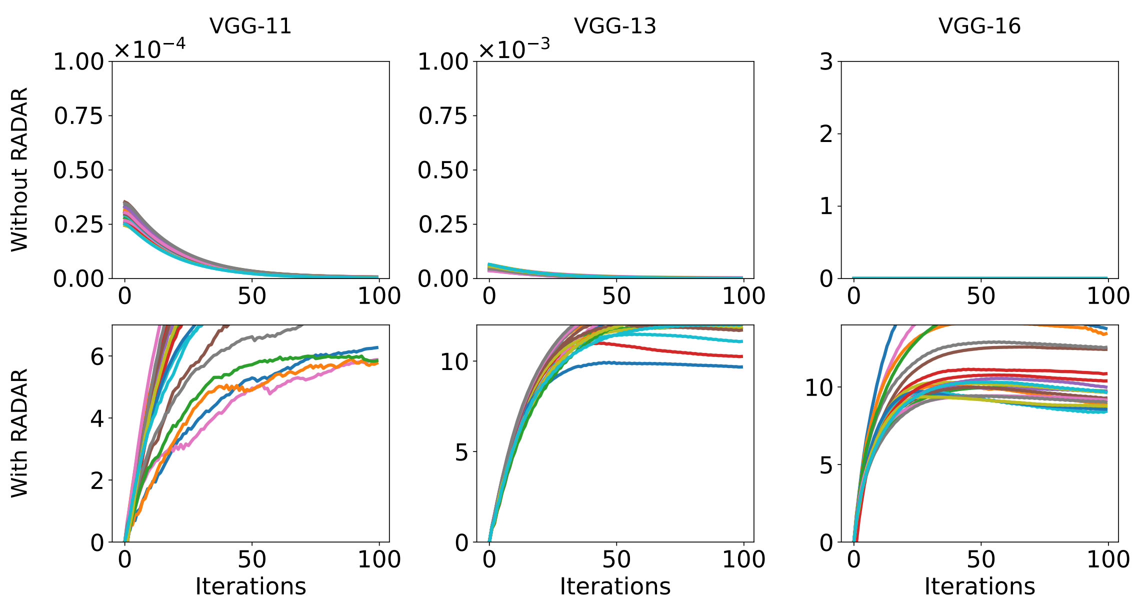

Figure 4: SVHN. Binary cross-entropy loss metrics, from the point of view of an attacker, are herein presented in the context of crafting an adversarial instance from the test set using VGG-11 (left), VGG-13 (middle), and VGG-16 (right).

Our results bring forth, perhaps, a sense of optimism regarding the attainability of adversarially robust deep learning detectors. Of note is the significant robustness exhibited by our networks across the datasets examined, manifested in increased accuracy against a diverse array of potent ∞-bound adversaries.

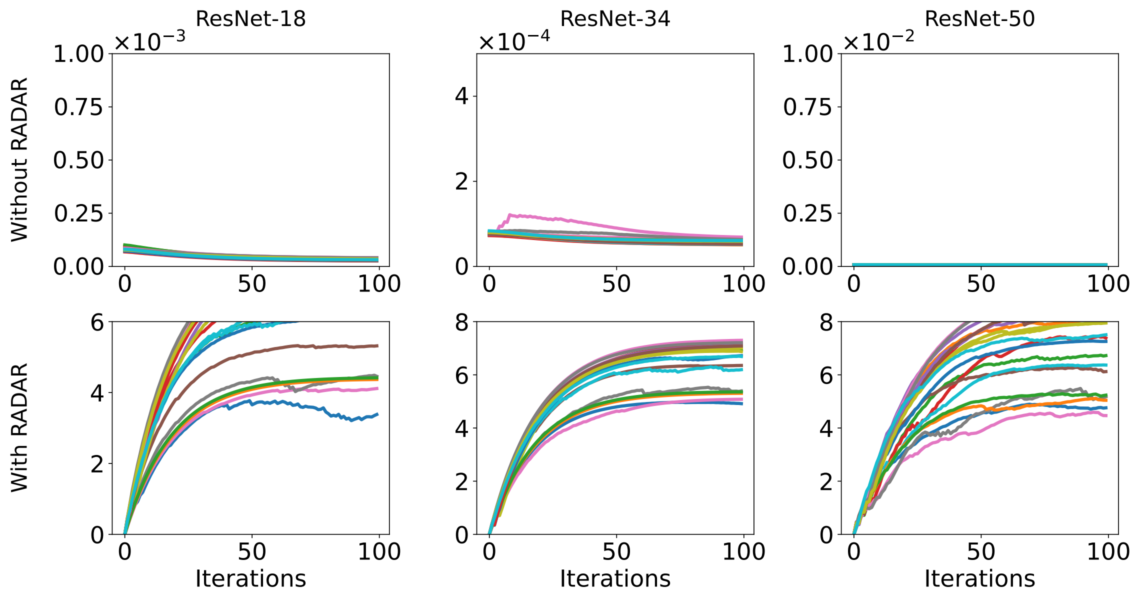

Figure 5: SVHN. Binary cross-entropy loss metrics, from the point of view of an attacker, are herein presented in the context of crafting an adversarial instance from the test set using ResNet-18 (left), ResNet-34 (middle), and ResNet-50 (right).

Our study not only contributes valuable insights into enhancing the resilience of deep neural networks against adversarial attacks but also underscores the importance of continued exploration in this domain. Therefore, we encourage researchers to persist in their endeavors to advance the frontier of adversarially robust deep learning detectors.

Acknowledgments

Anonymized.

References

Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search. In *European Conference on Computer Vision*, pp. 484–501. Springer, 2020.

Oliver Bryniarski, Nabeel Hingun, Pedro Pachuca, Vincent Wang, and Nicholas Carlini. Evading adversarial example detection defenses with orthogonal projected gradient descent. In *International Conference on Learning Representations*, 2021.

Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pp. 3–14, 2017a.

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

Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. Unlabeled data improves adversarial robustness. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Luyang Chen, Markus Pelger, and Jason Zhu. Deep learning in asset pricing. *Management Science*, 70(2): 714–750, 2024a.

Zhaoyu Chen, Bo Li, Shuang Wu, Kaixun Jiang, Shouhong Ding, and Wenqiang Zhang. Content-based unrestricted adversarial attack. *Advances in Neural Information Processing Systems*, 36, 2024b.

Shuyu Cheng, Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Improving black-box adversarial attacks with a transfer-based prior. *Advances in Neural Information Processing Systems*, 32, 2019.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pp. 1310–1320. PMLR, 2019.

Yinpeng Dong, Shuyu Cheng, Tianyu Pang, Hang Su, and Jun Zhu. Query-efficient black-box adversarial attacks guided by a transfer-based prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9536–9548, 2021.

Samuel G Finlayson, John D Bowers, Joichi Ito, Jonathan L Zittrain, Andrew L Beam, and Isaac S Kohane. Adversarial attacks on medical machine learning. *Science*, 363(6433):1287–1289, 2019.

Zhitao Gong and Wenlu Wang. Adversarial and clean data are not twins. In *Proceedings of the Sixth International Workshop on Exploiting Artificial Intelligence Techniques for Data Management*, pp. 1–5, 2023.

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

Kathrin Grosse, Praveen Manoharan, Nicolas Papernot, Michael Backes, and Patrick McDaniel. On the (statistical) detection of adversarial examples. *arXiv preprint arXiv:1702.06280*, 2017a.

Kathrin Grosse, Nicolas Papernot, Praveen Manoharan, Michael Backes, and Patrick McDaniel. Adversarial examples for malware detection. In *Computer Security–ESORICS 2017: 22nd European Symposium on Research in Computer Security, Oslo, Norway, September 11-15, 2017, Proceedings, Part II 22*, pp. 62–79. Springer, 2017b.

Zhiyuan He, Yijun Yang, Pin-Yu Chen, Qiang Xu, and Tsung-Yi Ho. Be your own neighborhood: Detecting adversarial example by the neighborhood relations built on self-supervised learning, 2022.

Jian Huang, Junyi Chai, and Stella Cho. Deep learning in finance and banking: A literature review and classification. *Frontiers of Business Research in China*, 14(1):1–24, 2020.

Marvin Klingner, Varun Ravi Kumar, Senthil Yogamani, Andreas Bär, and Tim Fingscheidt. Detecting adversarial perturbations in multi-task perception. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, October 2022. doi: 10.1109/iros47612.2022.9981559. URL http://dx.doi.org/10.1109/IROS47612.2022.9981559.

Alex Krizhevsky. Learning multiple layers of features from tiny images, 2009. URL https://api.semanticscholar.org/CorpusID:18268744.

Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial machine learning at scale. In *International Conference on Learning Representations*, 2016.

Raz Lapid and Moshe Sipper. Patch of invisibility: Naturalistic black-box adversarial attacks on object detectors. *arXiv preprint arXiv:2303.04238*, 2023a.

Raz Lapid and Moshe Sipper. I see dead people: Gray-box adversarial attack on image-to-text models. In *5th Workshop on Machine Learning for Cybersecurity, part of ECMLPKDD 2023*, 2023b.

Raz Lapid, Zvika Haramaty, and Moshe Sipper. An evolutionary, gradient-free, query-efficient, black-box algorithm for generating adversarial instances in deep convolutional neural networks. *Algorithms*, 15(11): 407, 2022.

Raz Lapid, Ron Langberg, and Moshe Sipper. Open sesame! universal black box jailbreaking of large language models. *arXiv preprint arXiv:2309.01446*, 2023.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*, pp. 656–672. IEEE, 2019.

Linyi Li, Tao Xie, and Bo Li. Sok: Certified robustness for deep neural networks. In *2023 IEEE Symposium on Security and Privacy (SP)*, pp. 1289–1310. IEEE, 2023.

Mingqian Liu, Zhenju Zhang, Yunfei Chen, Jianhua Ge, and Nan Zhao. Adversarial attack and defense on deep learning for air transportation communication jamming. *IEEE Transactions on Intelligent Transportation Systems*, 2023.

Jiajun Lu, Theerasit Issaranon, and David Forsyth. Safetynet: Detecting and rejecting adversarial examples robustly. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 446–454, 2017.

Julia Lust and Alexandru Paul Condurache. Gran: An efficient gradient-norm based detector for adversarial and misclassified examples. *arXiv preprint arXiv:2004.09179*, 2020.

Zhihan Lv, Shaobiao Zhang, and Wenqun Xiu. Solving the security problem of intelligent transportation system with deep learning. *IEEE Transactions on Intelligent Transportation Systems*, 22(7):4281–4290, 2020.

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.

Jan Hendrik Metzen, Tim Genewein, Volker Fischer, and Bastian Bischoff. On detecting adversarial perturbations. In *International Conference on Learning Representations*, 2016.

Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19(6):1236–1246, 2018.

Shahab Shamshirband, Mahdis Fathi, Abdollah Dehzangi, Anthony Theodore Chronopoulos, and Hamid Alinejad-Rokny. A review on deep learning approaches in healthcare systems: Taxonomies, challenges, and open issues. *Journal of Biomedical Informatics*, 113:103627, 2021.

Gagandeep Singh, Timon Gehr, Matthew Mirman, Markus Püschel, and Martin Vechev. Fast and effective robustness certification. *Advances in Neural Information Processing Systems*, 31, 2018.

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.

Snir Vitrack Tamam, Raz Lapid, and Moshe Sipper. Foiling explanations in deep neural networks. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=wvLQMHtyLk.

Florian Tramer and Dan Boneh. Adversarial training and robustness for multiple perturbations. *Advances in Neural Information Processing Systems*, 32, 2019.

Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan L Yuille, and Kaiming He. Feature denoising for improving adversarial robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 501–509, 2019.

Yao-Yuan Yang, Cyrus Rashtchian, Hongyang Zhang, Russ R Salakhutdinov, and Kamalika Chaudhuri. A closer look at accuracy vs. robustness. *Advances in Neural Information Processing Systems*, 33:8588–8601, 2020.

Yijun Yang, Ruiyuan Gao, Yu Li, Qiuxia Lai, and Qiang Xu. What you see is not what the network infers: Detecting adversarial examples based on semantic contradiction. In *Proceedings 2022 Network and Distributed System Security Symposium*, NDSS 2022. Internet Society, 2022. doi: 10.14722/ndss.2022.24001. URL http://dx.doi.org/10.14722/ndss.2022.24001.

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*, pp. 7472–7482. PMLR, 2019.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

Weimin Zhao, Sanaa Alwidian, and Qusay H Mahmoud. Adversarial training methods for deep learning: A systematic review. *Algorithms*, 15(8):283, 2022.