Benchmarking Adversarial Robustness on Image Classification

Yinpeng Dong[1], Qi-An Fu[1], Xiao Yang[1], Tianyu Pang[1], Hang Su[1*], Zihao Xiao[2], Jun Zhu[1*]
[1] Dept. of Comp. Sci. and Tech., BNRist Center, Institute for AI, THBI Lab
[1] Tsinghua University, Beijing, 100084, China [2] RealAI
{dyp17, fqa19, yangxiao19, pty17}@mails.tsinghua.edu.cn, {suhangss, dcszj}@tsinghua.edu.cn, zihao.xiao@realai.ai

Abstract

Deep neural networks are vulnerable to adversarial examples, which becomes one of the most important research problems in the development of deep learning. While a lot of efforts have been made in recent years, it is of great significance to perform correct and complete evaluations of the adversarial attack and defense algorithms. In this paper, we establish a comprehensive, rigorous, and coherent benchmark to evaluate adversarial robustness on image classification tasks. After briefly reviewing plenty of representative attack and defense methods, we perform large-scale experiments with two robustness curves as the fair-minded evaluation criteria to fully understand the performance of these methods. Based on the evaluation results, we draw several important findings that can provide insights for future research, including: 1) The relative robustness between models can change across different attack configurations, thus it is encouraged to adopt the robustness curves to evaluate adversarial robustness; 2) As one of the most effective defense techniques, adversarial training can generalize across different threat models; 3) Randomization-based defenses are more robust to query-based black-box attacks.

1. Introduction

Deep learning (DL) models are vulnerable to adversarial examples [53, 19], which are maliciously generated to induce erroneous predictions. As DL models have been integrated into various security-sensitive applications (*e.g.*, autonomous driving, healthcare, and finance), the study of the adversarial robustness issue has attracted increasing attention with an enormous number of adversarial attack and defense methods proposed. Therefore, it is crucial to conduct correct and rigorous evaluations of these methods for understanding their pros and cons, comparing their performance, and providing insights for building new methods [5].

The research on adversarial robustness is faced with an "*arms race*" between attacks and defenses, *i.e.*, a defense method proposed to prevent the existing attacks was soon evaded by new attacks, and vice versa [6, 7, 22, 1, 55, 65]. For instance, defensive distillation [41] was proposed to improve adversarial robustness, but was later shown to be ineffective against a strong attack [7]. Many methods were introduced to build robust models by causing obfuscated gradients, which can be defeated by the adaptive ones [1, 55]. As a result, it is particularly challenging to understand their effects, identify the real progress, and advance the field.

Moreover, the current attacks and defenses are often evaluated incompletely. First, most defenses are only tested against a small set of attacks under limited threat models, and many attacks are evaluated on a few models or defenses. Second, the robustness evaluation metrics are too simple to show the performance of these methods. The accuracy of a defense against an attack for a given perturbation budget [29] and the minimum distance of the adversarial perturbation [4] are used as the primary evaluation metrics, which are often insufficient to totally characterize the behaviour of the attacks and defenses. Consequently, the incomplete evaluation cannot provide a comprehensive understanding of the strengths and limitations of these methods.

In this paper, we establish a comprehensive, rigorous, and coherent benchmark to evaluate adversarial robustness, which can provide a detailed understanding of the effects of the existing methods under different scenarios, with a hope to facilitate future research. In particular, we focus on the robustness of image classifiers under the ℓ_p norm threat models where a large body of works have been devoted. We incorporate a lot of typical and state-of-the-art attack and defense methods for robustness evaluation, including 15 attack methods and 16 defense models—8 on CIFAR-10 [27] and 8 on ImageNet [46]. To fully demonstrate the performance of these methods, we adopt two complementary robustness curves as the major evaluation metrics to present the results. Then, we carry out large-scale experiments on the cross evaluation of the attack and defense methods under complete threat models, including 1) untargeted and targeted attacks; 2) ℓ_∞ and ℓ_2 attacks; 3) white-box, transfer-based, score-based, and decision-based attacks.

*Hang Su and Jun Zhu are corresponding authors.

2575-7075/20/$31.00 ©2020 IEEE
DOI 10.1109/CVPR42600.2020.00040

By analyzing the quantitative results, we have some important findings. First, the relative robustness between defenses against an attack can be different under varying perturbation budgets or attack iterations. Thus it is hard to conclude that a defense is more robust than another against an attack by using a specific configuration. However, it is common in previous works. Second, although various defense techniques have been proposed, the most robust defenses are still the adversarially trained models. Their robustness can also generalize to other threat models, under which they are not trained to be robust. Third, defenses based on randomization are generally more robust to query-based black-box attacks. More discussions can be found in Sec. 5.3.

We develop a new adversarial robustness platform called **RealSafe**[1] to conduct all evaluation experiments, since the existing platforms (*e.g.*, CleverHans [39] and Foolbox [44]) cannot fully support our evaluations (see Appendix A). We hope that our platform could continuously incorporate and evaluate more methods, and be helpful for future works.

2. Threat Models

Precisely defining threat models is fundamental to perform adversarial robustness evaluations. According to [5], a threat model specifies the adversary's goals, capabilities, and knowledge under which an attack is performed and a defense is built to be robust. We first define the notations and then illustrate the three aspects of a threat model.

A classifier can be denoted as $C(\boldsymbol{x}) : \mathcal{X} \rightarrow \mathcal{Y}$, where $\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^d$ is the input, and $\mathcal{Y} = \{1, 2, ..., L\}$ with L being the number of classes. Let y denote the ground-truth label of $\boldsymbol{x}$, and $\boldsymbol{x}^{adv}$ denote an adversarial example for $\boldsymbol{x}$.

2.1. Adversary's Goals

An adversary can have different goals of generating adversarial examples. We study the *untargeted* and *targeted* adversarial examples in this paper. An untargeted adversarial example aims to cause misclassification of the classifier, as $C(\boldsymbol{x}^{adv}) \neq y$. A targeted one is crafted to be misclassified as the adversary-desired target class by the classifier, as $C(\boldsymbol{x}^{adv}) = y^*$, where y^* is the target class.

2.2. Adversary's Capabilities

As adversarial examples are usually assumed to be indistinguishable from the corresponding original ones to human eyes [53, 19], the adversary can only make small changes to the inputs. In this paper, we study the well-defined and widely used ℓ_p norm threat models, although there also exist other threat models [58, 51, 18]. Under the ℓ_p norm threat models, the adversary is allowed to add a small perturbation measured by the ℓ_p norm to the original input. Specifically, we consider the ℓ_∞ and ℓ_2 norms.

[1]Code released at: https://github.com/thu-ml/realsafe.

To achieve the adversary's goal, two strategies could be adopted to craft adversarial examples with small perturbations. The first seeks to craft an adversarial example $\boldsymbol{x}^{adv}$ that satisfies $\|\boldsymbol{x}^{adv} - \boldsymbol{x}\|_p \leq \epsilon$, where ϵ is the perturbation budget, while misleads the model. This can be achieved by solving a constrained optimization problem. For instance, the adversary can get an untargeted adversarial example by maximizing a loss function $\mathcal{J}$ (*e.g.*, the cross-entropy loss) in the restricted region as

$$\boldsymbol{x}^{adv} = \underset{\boldsymbol{x}' : \|\boldsymbol{x}' - \boldsymbol{x}\|_p \leq \epsilon}{\arg\max} \mathcal{J}(\boldsymbol{x}', y). \tag{1}$$

We call it the adversarial example with a *constrained* perturbation. The second strategy is generating an adversarial example by finding the minimum perturbation as

$$\boldsymbol{x}^{adv} = \underset{\boldsymbol{x}' : \boldsymbol{x}' \text{is adversarial}}{\arg\min} \|\boldsymbol{x}' - \boldsymbol{x}\|_p. \tag{2}$$

We call it the adversarial example with an *optimized* perturbation. However, it is usually intractable to solve Eq. (1) or Eq. (2) exactly, and thus various attack methods have been proposed to get an approximate solution.

2.3. Adversary's Knowledge

An adversary can have different levels of knowledge of the target model, from white-box access to the model architectures and parameters, to black-box access to the training data or model predictions. Based on the different knowledge of the model, we consider four attack scenarios, including *white-box attacks*, *transfer-based*, *score-based*, and *decision-based black-box attacks*.

White-box attacks rely on detailed information of the target model, including architecture, parameters, and gradient of the loss w.r.t. the input. For defenses, the adversary can design adaptive attacks by considering the specific defense mechanisms. Transfer-based black-box attacks are based on the adversarial transferability [40], which assume the availability of training data. It is used to train a substitute model from which the adversarial examples are generated. Score-based black-box attacks can only acquire the output probabilities by querying the target model. And decision-based black-box attacks solely rely on the predicted classes of the queries. Score-based and decision-based attacks are also restricted by a limited number of queries to the target model.

3. Attacks and Defenses

In this section, we summarize the typical adversarial attack and defense methods.

3.1. Attack Methods

White-box Attacks: Most white-box attacks craft adversarial examples based on the input gradient. For solving Eq. (1), the fast gradient sign method (**FGSM**) [19] linearizes the loss function in the input space and generates

an adversarial example by an one-step update. The basic iterative method (**BIM**) [28] extends FGSM by iteratively taking multiple small gradient steps. Similar to BIM, the projected gradient descent method (**PGD**) [34] acts as a universal first-order adversary with random starts. For solving Eq. (2), **DeepFool** [35] has been proposed to generate an adversarial example with the minimum perturbation. The Carlini & Wagner's method (**C&W**) [7] takes a Lagrangian form and adopts Adam [26] for optimization. However, some defenses can be robust against these gradient-based attacks by causing obfuscated gradients [1]. To circumvent them, the adversary can use **BPDA** [1] to provide an approximate gradient when the true gradient is unavailable or useless, or **EOT** [2] when the gradient is random.

Transfer-based Black-box Attacks: Transfer-based attacks craft adversarial examples against a substitute model, which are probable to fool black-box models based on the transferability. Several methods have been proposed to improve the transferability. The momentum iterative method (**MIM**) [14] integrates a momentum term into BIM to stabilize the update direction during the attack iterations. The diverse inputs method (**DIM**) [62] applies the gradient of the randomly resized and padded input for adversarial example generation. The translation-invariant method (**TI**) [15] further improves the transferability for defense models.

Score-based Black-box Attacks: Under this setting, although the white-box access to the model gradient is unavailable, it can be estimated by the gradient-free methods through queries. **ZOO** [8] estimates the gradient at each coordinate by finite differences and adopts C&W for attacks based on the estimated gradient. **NES** [24] and **SPSA** [55] can give the full gradient estimation based on drawing random samples and acquiring the corresponding loss values. Prior-guided random gradient free method (**P-RGF**) [10] estimates the gradient more accurately with a transfer-based prior. $\mathcal{N}$**ATTACK** [30] does not estimate the gradient but learns a Gaussian distribution centered around the input such that a sample drawn from it is likely adversarial.

Decision-based Black-box Attacks: This setting is more challenging since the model only provides discrete hard-label predictions. The **Boundary** attack [3] is the first method in this setting based on random walk on the decision boundary. An **optimization-based** method [9] formulates this problem as a continuous optimization problem and estimates the gradient to solve it. The **evolutionary** attack method [16] is further proposed to improve the query efficiency based on the evolution strategy.

3.2. Defenses

Due to the threat of adversarial examples, extensive research has been conducted on building robust models to defend against adversarial attacks. In this paper, we roughly classify the defense techniques into five categories, including *robust training*, *input transformation*, *randomization*, *model ensemble*, and *certified defenses*. Note that these defense categories are not exclusive, *i.e*., a defense can belong to many categories. Below we introduce each category.

Robust Training: The basic principle of robust training is to make the classifier robust against small noises internally. One line of work is adversarial training [19, 54, 34, 25, 66], which augments the training data by adversarial examples. Another line of work trains robust models by other losses or regularizations, including variants on the network Lipschitz constant [11], input gradients [23, 45], perturbation norm [64], or the Max-Mahalanobis center loss [36].

Input Transformation: Several defenses transform the inputs before feeding them to the classifier, including JPEG compression [17], bit-depth reduction [63], total variance minimization [20], autoencoder-based denoising [31], and projecting adversarial examples onto the data distribution through generative models [47, 50]. However, these defenses can cause shattered gradients or vanishing/exploding gradients [1], which can be evaded by adaptive attacks.

Randomization: The classifiers can be made random to mitigate adversarial effects. The randomness can be added to either the input [60, 38] or the model [13, 32]. The randomness can also be modeled by Bayesian neural networks [33]. These methods partially rely on random gradients to prevent adversarial attacks, and can be defeated by attacks that take the expectation over the random gradients [22, 1].

Model Ensemble: An effective defense strategy in practice is to construct an ensemble of individual models [29]. Besides aggregating the output of each model in the ensemble, some different ensemble strategies have been proposed. Random self-ensemble [32] averages the predictions over random noises injected to the model, which is equivalent to ensemble an infinite number of noisy models. Pang *et al*. [37] propose to promote the diversity among the predictions of different models, and introduce an adaptive diversity promoting regularizer to achieve this.

Certified Defenses: There are a lot of works [42, 49, 56, 57, 43, 59] on training certified defenses, which are provably guaranteed to be robust against adversarial perturbations under some threat models. Recently, certified defenses [67, 12] can apply to ImageNet [46], showing the scalability of this type of defenses.

4. Evaluation Methodology

With the growing number of adversarial attacks and defenses being proposed, the correct and rigorous evaluation of these methods becomes increasingly important to help us better understand the strengths and limitations of these methods. However, there still lacks a comprehensive understanding of the effects of these methods due to the incorrect or incomplete evaluations. To address this issue and further

CIFAR-10 [27]				ImageNet [46]			
Defense Model	Category	Intended Threat	Acc.	Defense Model	Category	Intended Threat	Acc.
Res-56 [21]	natural training	-	92.6	Inc-v3 [52]	natural training	-	78.0
PGD-AT [34]	robust training	ℓ_∞ ($\epsilon = 8/255$)	87.3	Ens-AT [54]	robust training	ℓ_∞ ($\epsilon = 16/255$)	73.5
DeepDefense [64]	robust training	ℓ_2	79.7	ALP [25]	robust training	ℓ_∞ ($\epsilon = 16/255$)	49.0
TRADES [66]	robust training	ℓ_∞ ($\epsilon = 0.031$)	84.9	FD [61]	robust training	ℓ_∞ ($\epsilon = 16/255$)	64.3
Convex [57]	(certified) robust training	ℓ_∞ ($\epsilon = 2/255$)	66.3	JPEG [17]	input transformation	General	77.3
JPEG [17]	input transformation	General	80.9	Bit-Red [63]	input transformation	General	61.8
RSE [32]	rand. & ensemble	ℓ_2	86.1	R&P [60]	(input) rand.	General	77.0
ADP [37]	ensemble	General	94.1	RandMix [67]	(certified input) rand.	General	52.4

Table 1: We show the defense models that are incorporated into our benchmark for adversarial robustness evaluation. We also show the defense type, original intended threat model (*i.e.*, the threat model under which the defense is trained to be robust or evaluated in the original paper; 'General' means the defense can be used for any threat model), and accuracy (%) on clean data of each method. The accuracy is re-calculated by ourselves. More details about their model architectures are shown in Appendix B.

advance the field, we establish a comprehensive, rigorous, and coherent benchmark to evaluate adversarial robustness empirically. We incorporate 15 attack methods and 16 defense models on two image datasets in our benchmark for robustness evaluation. We also adopt two complementary robustness curves as the fair-minded evaluation metrics.

4.1. Evaluation Metrics

Given an attack method $\mathcal{A}_{\epsilon,p}$ that generates an adversarial example $\boldsymbol{x}^{adv} = \mathcal{A}_{\epsilon,p}(\boldsymbol{x})$ for an input $\boldsymbol{x}$ with perturbation budget ϵ under the ℓ_p norm[2], and a (defense) classifier C defined in Sec. 2, the accuracy of the classifier against the attack is defined as

$$\mathrm{Acc}(C, \mathcal{A}_{\epsilon,p}) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\big(C(\mathcal{A}_{\epsilon,p}(\boldsymbol{x}_i)) = y_i\big),$$

where $\{\boldsymbol{x}_i, y_i\}_{i=1}^N$ is the test set, $\mathbf{1}(\cdot)$ is the indicator function. The attack success rate of an untargeted attack on the classifier is defined as

$$\mathrm{Asr}(\mathcal{A}_{\epsilon,p}, C) = \frac{1}{M}\sum_{i=1}^{N} \mathbf{1}\big(C(\boldsymbol{x}_i) = y_i \wedge C(\mathcal{A}_{\epsilon,p}(\boldsymbol{x}_i)) \neq y_i\big),$$

where $M = \sum_{i=1}^{N} \mathbf{1}\big(C(\boldsymbol{x}_i) = y_i\big)$, while the attack success rate of a targeted attack is defined as

$$\mathrm{Asr}(\mathcal{A}_{\epsilon,p}, C) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\big(C(\mathcal{A}_{\epsilon,p}(\boldsymbol{x}_i)) = y_i^*\big).$$

where y_i^* is the target class corresponding to $\boldsymbol{x}_i$.

The previous methods usually report the point-wise accuracy or attack success rate for some chosen perturbation budgets ϵ, which may not reflect their behaviour totally. In this paper, we adopt two complementary robustness curves to clearly and thoroughly show the robustness and resistance of the classifier against the attack, as well as the effectiveness and efficiency of the attack on the classifier.

The first one is the *accuracy (attack success rate) vs. perturbation budget* curve, which can give a global understanding of the robustness of the classifier and the effectiveness of the attack. To generate such a curve, we need to calculate the accuracy or attack success rate for all values of ϵ. This can be efficiently done for attacks that find the minimum perturbations, by counting the number of the adversarial examples, the ℓ_p norm of whose perturbations is smaller than each ϵ. For attacks that craft adversarial examples with constrained perturbations, we perform a binary search on ϵ to find its minimum value that enables the generated adversarial example to fulfill the adversary's goal.

[2]For attacks that find minimum perturbations, *e.g.*, DeepFool, C&W, we let $\mathcal{A}_{\epsilon,p}(\boldsymbol{x}) = \boldsymbol{x}$ if the ℓ_p norm of the perturbation is larger than ϵ.

Attack Method	Knowledge	Goals	Capability	Distance
FGSM [19]	white & transfer	un. & tar.	constrained	ℓ_∞, ℓ_2
BIM [28]	white & transfer	un. & tar.	constrained	ℓ_∞, ℓ_2
MIM [14]	white & transfer	un. & tar.	constrained	ℓ_∞, ℓ_2
DeepFool [35]	white	un.	optimized	ℓ_∞, ℓ_2
C&W [7]	white	un. & tar.	optimized	ℓ_2
DIM [62]	transfer	un. & tar.	constrained	ℓ_∞, ℓ_2
ZOO [8]	score	un. & tar.	optimized	ℓ_2
NES [24]	score	un. & tar.	constrained	ℓ_∞, ℓ_2
SPSA [55]	score	un. & tar.	constrained	ℓ_∞, ℓ_2
$\mathcal{N}$ATTACK [30]	score	un. & tar.	constrained	ℓ_∞, ℓ_2
Boundary [3]	decision	un. & tar.	optimized	ℓ_2
Evolutionary [16]	decision	un. & tar.	optimized	ℓ_2

Table 2: We show the attack methods that are implemented in our benchmark for adversarial robustness evaluation. We also show the adversary's knowledge (white-box, transfer-based, score-based, or decision-based), goals ('un.' stands for untargeted; 'tar.' stands for targeted), capability (constrained or optimized perturbations), and distance metrics of each attack method.

The second curve is the *accuracy (attack success rate) vs. attack strength* curve, where the attack strength is defined as the number of iterations or model queries based on different attack methods. This curve can show the efficiency of the attack, as well as the resistance of the classifier to the attack, *e.g.*, a defense whose accuracy drops to zero against an attack with 100 iterations is considered to be more resistant to this attack than another defense that is totally broken by the same attack with 10 iterations, although the worst-case accuracy of both models is zero.

4.2. Evaluated Datasets and Algorithms

Datasets: We use the CIFAR-10 [27] and ImageNet [46] datasets to perform adversarial robustness evaluation in this paper. We use the test set containing $10,000$ images of CIFAR-10, and randomly choose $1,000$ images from the ImageNet validation set for evaluation. For each image, we select a target class uniformly over all other classes except

Attack		Res-56	PGD-AT	DeepDefense	TRADES	Convex	JPEG	RSE	ADP
White	FGSM	0.005/21.6%	0.039/56.0%	0.001/9.2%	0.047/60.9%	0.017/36.6%	0.012/31.2%	0.020/29.0%	0.037/56.0%
	BIM	0.002/0.0%	0.030/48.3%	0.001/0.0%	0.037/56.8%	0.016/34.3%	0.008/3.2%	0.018/23.5%	0.008/12.2%
	MIM	0.003/0.0%	0.032/50.9%	0.001/0.0%	0.040/58.1%	0.016/34.9%	0.008/6.1%	0.019/25.1%	0.010/16.7%
	DeepFool	0.003/0.0%	0.040/56.5%	0.001/0.0%	0.047/60.6%	0.015/32.9%	0.007/3.1%	0.021/35.9%	0.016/28.7%
Transfer	FGSM	0.067/72.9%	0.067/71.3%	0.048/62.1%	0.087/73.6%	0.050/57.5%	0.051/62.8%	0.048/62.0%	0.066/73.4%
	BIM	0.049/70.3%	0.055/70.2%	0.041/58.8%	0.069/72.2%	0.044/56.7%	0.039/58.9%	0.041/60.0%	0.048/71.4%
	MIM	0.052/71.5%	0.056/70.4%	0.041/59.4%	0.067/72.2%	0.045/56.6%	0.041/59.9%	0.043/59.8%	0.050/70.4%
	DIM	0.052/73.3%	0.056/70.0%	0.043/58.8%	0.063/70.5%	0.044/55.3%	0.043/61.1%	0.043/60.2%	0.051/73.4%
Score	NES	0.004/0.0%	0.048/65.5%	0.002/0.0%	0.055/66.7%	0.025/44.0%	0.001/2.1%	0.293/79.7%	0.007/12.1%
	SPSA	0.003/0.0%	0.042/61.1%	0.002/0.0%	0.049/64.9%	0.021/39.7%	0.001/2.1%	0.208/78.7%	0.007/9.7%
	$\mathcal{N}$ATTACK	0.002/0.0%	0.030/48.6%	0.001/0.0%	0.037/55.8%	0.016/33.1%	0.000/0.0%	0.031/48.6%	0.005/2.4%

Table 3: The point-wise results of the 8 models on CIFAR-10 against untargeted attacks under the ℓ_∞ norm given by the previous evaluation criteria. Each entry shows the median ℓ_∞ distance of the minimum adversarial perturbations across all samples (left) as well as the model's accuracy for the fixed $\epsilon = 8/255$ (right).

its true class at random, which is used for targeted attacks.

Defense Models: For fair evaluation, we test 16 representative defense models whose original source codes and pre-trained models are publicly available. These models cover all defense categories and include the state-of-the-art models in each category. On CIFAR-10, we choose 8 models—naturally trained ResNet-56 (Res-56) [21], PGD-based adversarial training (PGD-AT) [34], DeepDefense [64], TRADES [66], convex outer polytope (Convex) [57], JPEG compression [17], random self-ensemble (RSE) [32], and adaptive diversity promoting (ADP) [37]. On ImageNet, we also choose 8 models—naturally trained Inception v3 (Inc-v3) [52], ensemble adversarial training (Ens-AT) [54], adversarial logit pairing (ALP) [25], feature denoising (FD) [61], JPEG compression [17], bit-depth reduction (Bit-Red) [63], random resizing and padding (R&P) [60], and RandMix [67]. We use the natural models as the backbone classifiers for defenses based on input transformation (*e.g.*, JPEG). Table 1 shows the defense details. The reason why we choose many *weak* defenses based on randomization or input transformation, which are already broken [1], is that we want to show their behaviour under various threat models comprehensively, and we indeed draw some findings for these defenses.

Attacks: We implement 15 typical and widely used attack methods in our benchmark, including 5 white-box attacks—FGSM, BIM, MIM, DeepFool, and C&W, 4 transfer-based attacks—FGSM, BIM, MIM, and DIM, 4 score-based attacks—ZOO, NES, SPSA, and $\mathcal{N}$ATTACK, and 2 decision-based attacks—Boundary and Evolutionary. More details of these attacks are outlined in Table 2. Note that 1) we do not evaluate PGD since PGD and BIM are very similar and often result in similar performance; 2) for transfer-based attacks, we craft adversarial examples by those white-box methods on a substitute model; 3) for defenses that rely on obfuscated gradients, we implement the white-box attacks adaptively by replacing the true gradient with an approximate one when it is unavailable or an expected one when it is random, such that the white-box attacks can identify the worst-case robustness of the models.

Platform: All attacks and defenses are implemented on a new adversarial robustness platform—**RealSafe**. We also conduct the experiments based on the platform. Our platform takes a modular implementation, which is easily extendable, as detailed in Appendix A. We acknowledge that many works are not included in our current benchmark. We hope that our platform could continuously incorporate and evaluate more methods, and be helpful for future works.

5. Evaluation Results

We present the evaluation results on CIFAR-10 in Sec. 5.1, and ImageNet in Sec. 5.2. Due to the space limitation, we mainly provide the accuracy vs. perturbation budget and attack strength curves of the defense models against untargeted attacks under the ℓ_∞ norm in this section, and leave the full experimental results (including targeted attacks under the ℓ_∞ norm, untargeted and targeted attacks under the ℓ_2 norm, and attack success rate curves) in Appendix C. We also report some key findings in Sec. 5.3.

5.1. Evaluation Results on CIFAR-10

In this section, we show the accuracy of the 8 models on CIFAR-10 against white-box, transfer-based, score-based, and decision-based attacks . To get the *accuracy vs. perturbation budget* curves, we fix the attack strength (*i.e.*, attack iterations or queries) for different budgets. To generate the *accuracy vs. attack strength* curves, we use a fixed perturbation budget as $\epsilon = 8/255$ for ℓ_∞ attacks and $\epsilon = 1.0$ for ℓ_2 attacks, with images in $[0, 1]$. The detailed parameters of each attack are provided in Appendix B. We let the attack parameters be the same for evaluating all defense models, and leave the study of attack parameters on robustness performance in future works. To better show the superiority of the robustness curves adopted in this paper compared with the previous evaluation criteria (*i.e.*, the median distance of the minimum adversarial perturbations [4] and the accuracy of a model against an attack for a given perturbation budget [29]), we show the evaluation results based on the previous evaluation criteria in Table 3.

White-box Attacks: We show the *accuracy vs. perturbation budget* curves of the 8 models against untargeted FGSM, BIM, MIM, and DeepFool attacks under the ℓ_∞

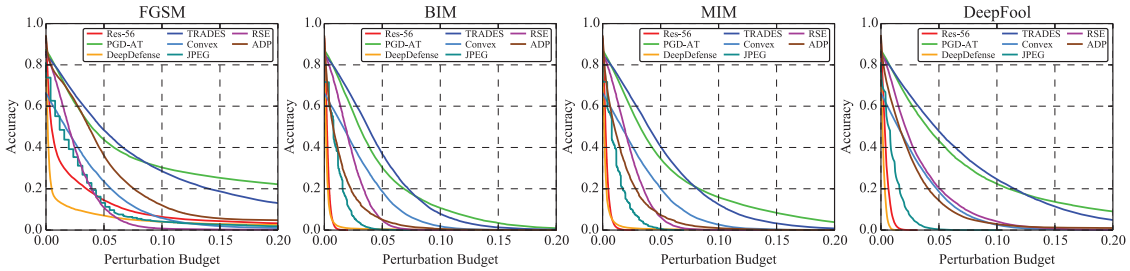

Figure 1: The *accuracy vs. perturbation budget* curves of the 8 models on CIFAR-10 against untargeted white-box attacks under the ℓ_∞ norm.

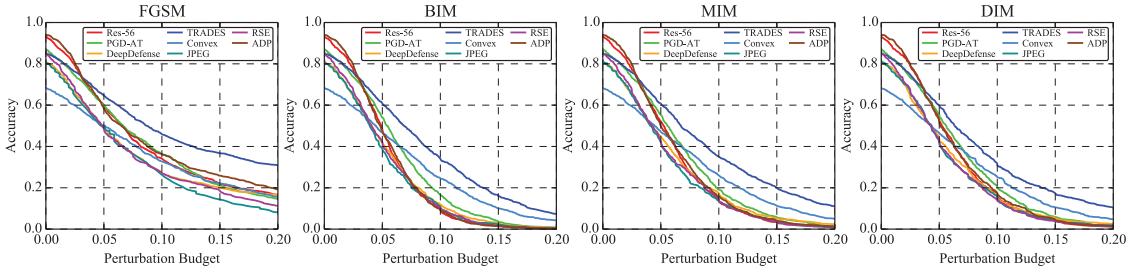

Figure 2: The *accuracy vs. perturbation budget* curves of the 8 models on CIFAR-10 against untargeted transfer-based attacks under the ℓ_∞ norm.

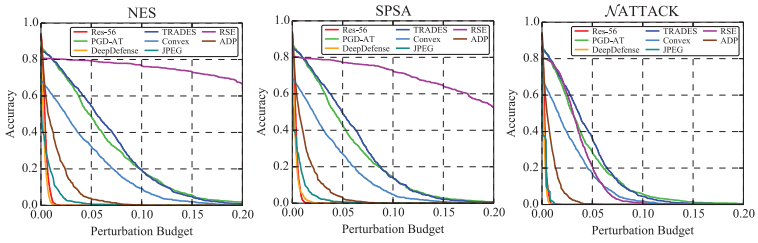

Figure 3: The *accuracy vs. perturbation budget* curves of the 8 models on CIFAR-10 against untargeted score-based attacks under the ℓ_∞ norm.

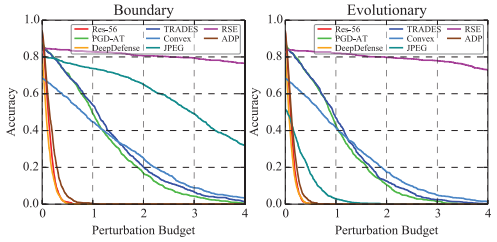

Figure 4: The *accuracy vs. perturbation budget* curves of the 8 models on CIFAR-10 against untargeted decision-based attacks under the ℓ_2 norm.

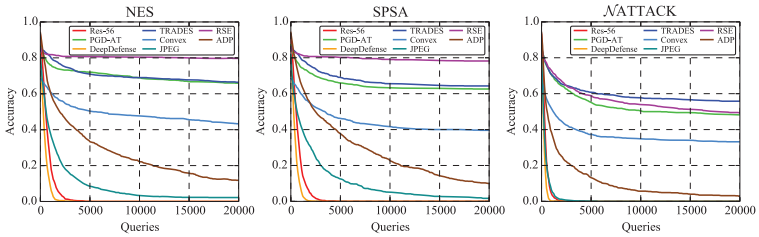

Figure 5: The *accuracy vs. attack strength* curves of the 8 models on CIFAR-10 against untargeted score-based attacks under the ℓ_∞ norm.

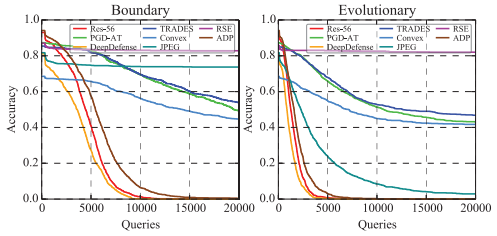

Figure 6: The *accuracy vs. attack strength* curves of the 8 models on CIFAR-10 against untargeted decision-based attacks under the ℓ_2 norm.

norm in Fig. 1 and leave the *accuracy vs. attack strength* curves in Appendix C. The accuracy of the models drops to zero against iterative attacks with the increasing perturbation budget. Based on the results, we observe that under white-box attacks, the adversarially trained models (*i.e.*, PGD-AT, TRADES) are more robust than other models, because they are trained on the worst-case adversarial examples. We also observe that the relative robustness between two models against an attack could be different under different perturbation budgets or attack iterations (shown in Appendix C). For instance, the accuracy of TRADES is higher than that of PGD-AT against white-box attacks when the perturbation budget is small (*e.g.*, $\epsilon = 0.05$), but is lower when it is large (*e.g.*, $\epsilon = 0.15$). This finding implies that the comparison between the defense models at a chosen perturbation budget or attack iteration, which is common in previous works, cannot fully demonstrate the performance of a model. But the robustness curves adopted in this paper can better show the global behaviour of these methods, compared with the point-wise evaluation results in Table 3.

Transfer-based Black-box Attacks: We show the *accuracy vs. perturbation budget* curves of the 8 models against untargeted transfer-based FGSM, BIM, MIM, and DIM attacks under the ℓ_∞ norm in Fig. 2, and leave the *accuracy vs. attack strength* curves in Appendix C. In this experiment, we choose TRADES as the substitute model to attack the others, and use PGD-AT to attack TRADES, since these two models demonstrate superior white-box robustness compared with the other models, and thus the adversarial examples generated on the other models can rarely transfer to TRADES and PGD-AT. From the results, the accuracy of the defenses also drops with the increasing per-

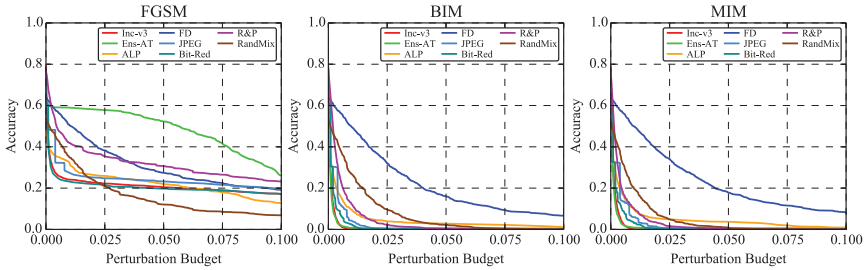

Figure 7: The *accuracy vs. perturbation budget* curves of the 8 models on ImageNet against untargeted white-box attacks under the ℓ_∞ norm.

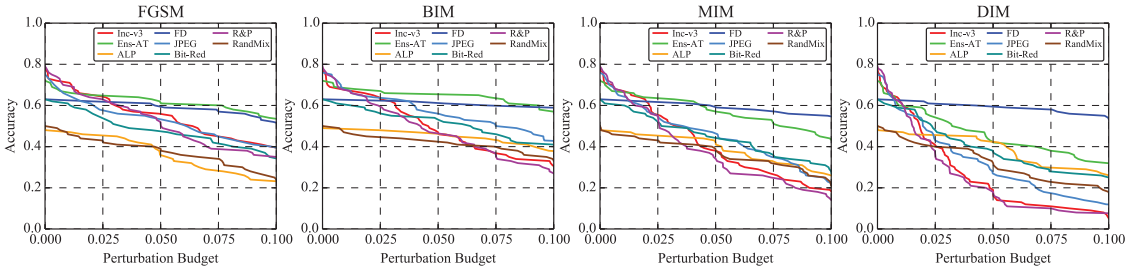

Figure 8: The *accuracy vs. perturbation budget* curves of the 8 models on ImageNet against untargeted transfer-based attacks under the ℓ_∞ norm.

turbation budget. We also observe that the recent attacks (*e.g.*, MIM, DIM) for improving the transferability do not actually perform better than the baseline BIM method.

Score-based Black-box Attacks: We show the curves of the *accuracy vs. perturbation budget* and *accuracy vs. attack strength (queries)* of the 8 models against untargeted score-based NES, SPSA, and $\mathcal{N}$ATTACK under the ℓ_∞ norm in Fig. 3 and Fig. 5. We set the maximum number of queries as $20,000$ in these attack methods. The accuracy of the defenses also decreases along with the increasing perturbation budget or the number of queries. $\mathcal{N}$ATTACK is more effective as can be seen from the figures. From the results, we notice that RSE is quite resistant to score-based attacks, especially NES and SPSA. We think that the randomness of the predictions given by RSE makes the estimated gradients of NES and SPSA useless for attacks.

Decision-based Black-box Attacks: Since the decision-based Boundary and Evolutionary attack methods can be only used for ℓ_2 attacks, we present the accuracy curves of the 8 models against untargeted Boundary and Evolutionary attacks under the ℓ_2 norm in Fig. 4 and Fig. 6. The behaviour of the defenses is similar to that of the score-based attacks. It can be observed that RSE is also resistant to decision-based attacks compared with the other defenses due to the randomness of the predictions.

5.2. Evaluation Results on ImageNet

We present the experimental results on ImageNet in this section. We use the same settings with those on CIFAR-10 to get the evaluation curves. Since the input image size is different for the ImageNet defenses, we adopt the normalized ℓ_2 distance defined as $\bar{\ell}_2(\boldsymbol{a}) = \|\boldsymbol{a}\|_2/\sqrt{d}$ as the measurement for ℓ_2 attacks, where d is the dimension of a vector $\boldsymbol{a}$. To get the *accuracy (attack success rate) vs. attack strength* curves, we fix the perturbation budget as $\epsilon = 16/255$ for ℓ_∞ attacks and $\epsilon = \sqrt{0.001}$ for ℓ_2 attacks.

White-box Attacks: We show the *accuracy vs. perturbation budget* curves of the 8 models on ImageNet against untargeted FGSM, BIM, and MIM under the ℓ_∞ norm in Fig. 7. We also leave the *accuracy vs. attack strength* curves in Appendix C. We find that FD exhibits superior performance over all other models. FD is also trained by the adversarial training method in [34], demonstrating the effectiveness of PGD-based adversarial training on ImageNet.

Transfer-based Black-box Attacks: We use a ResNet-152 model [21] as the substitute model. The *accuracy vs. perturbation budget* curves of the defenses against untargeted transfer-based FGSM, BIM, MIM, and DIM under the ℓ_∞ norm are shown in Fig. 8. Different from the results on CIFAR-10, MIM and DIM improve the transferability of adversarial examples over FGSM and BIM, resulting in lower accuracy of the black-box models. A potential reason is that the image size of ImageNet is much larger, and the adversarial examples crafted by BIM can "overfit" the substitute model [14], making them hard to transfer to others.

Score-based and Decision-based Attacks: Fig. 9 and Fig. 11 show the *accuracy vs. perturbation budget* and *accuracy vs. attack strength (queries)* curves of the defense models on ImageNet against untargeted score-based attacks under the ℓ_∞ norm, while Fig. 10 and Fig. 12 show the two sets of curves for untargeted decision-based attacks under the ℓ_2 norm. Similar to the results on CIFAR-10, we find that the two defenses based on randomization, *i.e.*, R&P and RandMix, have higher accuracy than the other methods in most cases. JPEG and Bit-Red that are based on input transformations also improve the robustness over the baseline model (*i.e.*, Inc-v3).

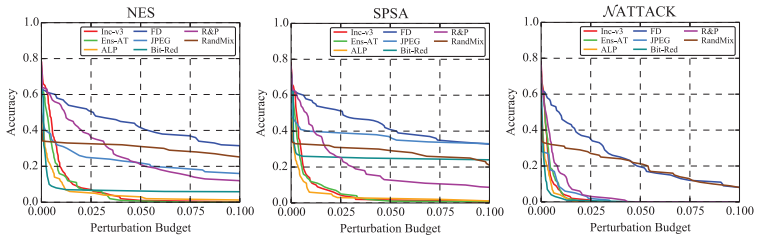

Figure 9: The *accuracy vs. perturbation budget* curves of the 8 models on ImageNet against untargeted score-based attacks under the ℓ_∞ norm.

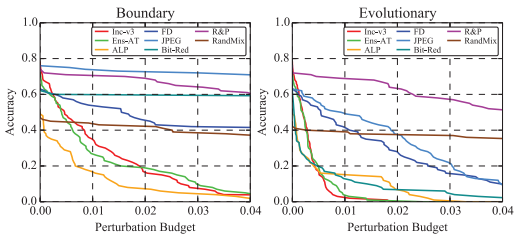

Figure 10: The *accuracy vs. perturbation budget* curves of the 8 models on ImageNet against untargeted decision-based attacks under the ℓ_2 norm.

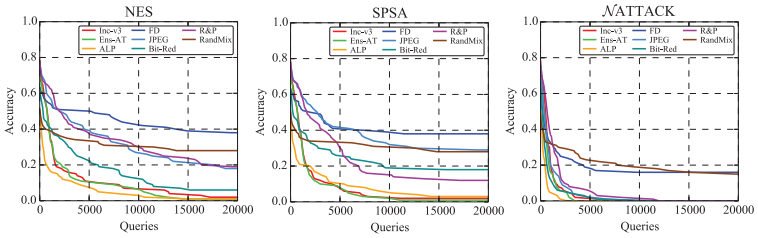

Figure 11: The *accuracy vs. attack strength* curves of the 8 models on ImageNet against untargeted score-based attacks under the ℓ_∞ norm.

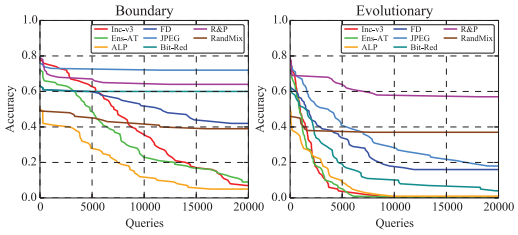

Figure 12: The *accuracy vs. attack strength* curves of the 8 models on ImageNet against untargeted decision-based attacks under the ℓ_2 norm.

5.3. Discussions

Based on the above results and more results in Appendix C, we highlight some key findings.

First, the relative robustness between defenses against the same attack can be different under varying attack parameters, such as the perturbation budget or the number of attack iterations. Not only the results of PGD-AT and TRADES in Fig. 1 can prove it, but also the results in many different scenarios show the similar phenomenon. Given this observation, the comparison between defenses at a specific attack configuration cannot fully demonstrate the superiority of a method upon another. We therefore strongly *advise the researchers to adopt the robustness curves as the major evaluation metrics to present the robustness results.*

Second, among the defenses studied in this paper, we find that the most robust models are obtained by PGD-based adversarial training. Their robustness not only is good for the threat model under which they are trained (*i.e.*, the ℓ_∞ threat model), but can also generalize to other threat models (*e.g.*, the ℓ_2 threat model). However, adversarial training usually leads to a reduction of natural accuracy and high training cost. A research direction is to develop new methods that maintain the natural accuracy or reduce the training cost. And we have seen several works [48] in this direction.

Third, we observe that the defenses based on randomization are quite resistant to score-based and decision-based attacks, which rely on the query feedback of the black-box models. We argue that the robustness of the randomization-based defenses against these attacks is due to the random predictions given by the models, making the estimated gradients or search directions unreliable for attacks. A potential research direction is to develop more powerful score-based and decision-based attacks that can efficiently evade the randomization-based defenses.

Fourth, the defenses based on input transformations (*e.g.*, JPEG, Bit-Red) sightly improve the robustness over undefended ones, and sometimes get much higher accuracy against score-based and decision-based attacks. Since these methods are quite simple, they may be combined with other types of defenses to build more powerful defenses.

Fifth, we find that different transfer-based attack methods exhibit similar performance on CIFAR-10, while the recent methods (*e.g.*, MIM, DIM) can improve the transferability of adversarial examples over BIM on ImageNet. One potential reason is that the input dimension of the models on ImageNet is much higher than that on CIFAR-10, and thus the adversarial examples generated by BIM can easily "overfit" the substitute model [14], resulting in poor transferability. The recent methods proposed to solve this issue can generate more transferable adversarial examples.

6. Conclusion

In this paper, we established a comprehensive, rigorous, and coherent benchmark to evaluate adversarial robustness of image classifiers. We performed large-scale experiments with two robustness curves as the fair-minded evaluation criteria to facilitate a better understanding of the representative and state-of-the-art adversarial attack and defense methods. We drew some key findings based on the evaluation results, which may be helpful for future research.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (No. 2017YFA0700904), NSFC Projects (Nos. 61620106010, U19B2034, U1811461), Beijing NSF Project (No. L172037), Beijing Academy of Artificial Intelligence (BAAI), Tsinghua-Huawei Joint Research Program, a grant from Tsinghua Institute for Guo Qiang, Tiangong Institute for Intelligent Computing, the JP Morgan Faculty Research Program and the NVIDIA NVAIL Program with GPU/DGX Acceleration. Yinpeng Dong is supported by MSRA, Baidu Fellowships.

References

[1] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning (ICML)*, 2018. 1, 3, 5

[2] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *International Conference on Machine Learning (ICML)*, 2018. 3

[3] Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In *International Conference on Learning Representations (ICLR)*, 2018. 3, 4

[4] Wieland Brendel, Jonas Rauber, Alexey Kurakin, Nicolas Papernot, Behar Veliqi, Sharada P. Mohanty, Florian Laurent, Marcel Salathé, Matthias Bethge, Yaodong Yu, Hongyang Zhang, Susu Xu, Hongbao Zhang, Pengtao Xie, Eric P. Xing, Thomas Brunner, Frederik Diehl, Jérôme Rony, Luiz Gustavo Hafemann, Shuyu Cheng, Yinpeng Dong, Xuefei Ning, Wenshuo Li, and Yu Wang. Adversarial vision challenge. In *The NeurIPS '18 Competition*, pages 129–153, 2020. 1, 5

[5] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, and Aleksander Madry. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019. 1, 2

[6] Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *ACM Workshop on Artificial Intelligence and Security*, 2017. 1

[7] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*, 2017. 1, 3, 4

[8] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 15–26. ACM, 2017. 3, 4

[9] Minhao Cheng, Thong Le, Pin-Yu Chen, Jinfeng Yi, Huan Zhang, and Cho-Jui Hsieh. Query-efficient hard-label black-box attack: An optimization-based approach. In *International Conference on Learning Representations (ICLR)*, 2019. 3

[10] Shuyu Cheng, Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Improving black-box adversarial attacks with a transfer-based prior. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 3

[11] Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier. Parseval networks: Improving robustness to adversarial examples. In *International Conference on Machine Learning (ICML)*, 2017. 3

[12] Jeremy M Cohen, Elan Rosenfeld, and J Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning (ICML)*, 2019. 3

[13] Guneet S Dhillon, Kamyar Azizzadenesheli, Zachary C Lipton, Jeremy Bernstein, Jean Kossaifi, Aran Khanna, and Anima Anandkumar. Stochastic activation pruning for robust adversarial defense. In *International Conference on Learning Representations (ICLR)*, 2018. 3

[14] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3, 4, 7, 8

[15] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[16] Yinpeng Dong, Hang Su, Baoyuan Wu, Zhifeng Li, Wei Liu, Tong Zhang, and Jun Zhu. Efficient decision-based black-box adversarial attacks on face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3, 4

[17] Gintare Karolina Dziugaite, Zoubin Ghahramani, and Daniel M Roy. A study of the effect of jpg compression on adversarial images. *arXiv preprint arXiv:1608.00853*, 2016. 3, 4, 5

[18] Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. Exploring the landscape of spatial robustness. In *International Conference on Machine Learning (ICML)*, 2019. 2

[19] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015. 1, 2, 3, 4

[20] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens van der Maaten. Countering adversarial images using input transformations. In *International Conference on Learning Representations (ICLR)*, 2018. 3

[21] 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 (CVPR)*, 2016. 4, 5, 7

[22] Warren He, Bo Li, and Dawn Song. Decision boundary analysis of adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2018. 1, 3

[23] Matthias Hein and Maksym Andriushchenko. Formal guarantees on the robustness of a classifier against adversarial manipulation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 3

[24] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *International Conference on Machine Learning (ICML)*, 2018. 3, 4

[25] Harini Kannan, Alexey Kurakin, and Ian Goodfellow. Adversarial logit pairing. *arXiv preprint arXiv:1803.06373*, 2018. 3, 4, 5

[26] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 3

[27] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. 1, 4

[28] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *International Conference on Learning Representations (ICLR) Workshops*, 2017. 3, 4

[29] Alexey Kurakin, Ian Goodfellow, Samy Bengio, Yinpeng Dong, Fangzhou Liao, Ming Liang, Tianyu Pang, Jun Zhu, Xiaolin Hu, Cihang Xie, et al. Adversarial attacks and defences competition. *arXiv preprint arXiv:1804.00097*, 2018. 1, 3, 5

[30] Yandong Li, Lijun Li, Liqiang Wang, Tong Zhang, and Boqing Gong. Nattack: Learning the distributions of adversarial examples for an improved black-box attack on deep neural networks. In *International Conference on Machine Learning (ICML)*, 2019. 3, 4

[31] Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, and Jun Zhu. Defense against adversarial attacks using high-level representation guided denoiser. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[32] Xuanqing Liu, Minhao Cheng, Huan Zhang, and Cho-Jui Hsieh. Towards robust neural networks via random self-ensemble. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 3, 4, 5

[33] Xuanqing Liu, Yao Li, Chongruo Wu, and Cho-Jui Hsieh. Adv-bnn: Improved adversarial defense through robust bayesian neural network. In *International Conference on Learning Representations (ICLR)*, 2019. 3

[34] 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 (ICLR)*, 2018. 3, 4, 5, 7

[35] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3, 4

[36] Tianyu Pang, Kun Xu, Yinpeng Dong, Chao Du, Ning Chen, and Jun Zhu. Rethinking softmax cross-entropy loss for adversarial robustness. *arXiv preprint arXiv:1905.10626*, 2019. 3

[37] Tianyu Pang, Kun Xu, Chao Du, Ning Chen, and Jun Zhu. Improving adversarial robustness via promoting ensemble diversity. In *International Conference on Machine Learning (ICML)*, 2019. 3, 4, 5

[38] Tianyu Pang, Kun Xu, and Jun Zhu. Mixup inference: Better exploiting mixup to defend adversarial attacks. *arXiv preprint arXiv:1909.11515*, 2019. 3

[39] Nicolas Papernot, Fartash Faghri, Nicholas Carlini, Ian Goodfellow, Reuben Feinman, Alexey Kurakin, Cihang Xie, Yash Sharma, Tom Brown, Aurko Roy, et al. Technical report on the cleverhans v2. 1.0 adversarial examples library. *arXiv preprint arXiv:1610.00768*, 2016. 2

[40] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against deep learning systems using adversarial examples. *arXiv preprint arXiv:1602.02697*, 2016. 2

[41] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *IEEE Symposium on Security and Privacy*, 2016. 1

[42] Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2018. 3

[43] Aditi Raghunathan, Jacob Steinhardt, and Percy S Liang. Semidefinite relaxations for certifying robustness to adversarial examples. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 3

[44] Jonas Rauber, Wieland Brendel, and Matthias Bethge. Foolbox v0. 8.0: A python toolbox to benchmark the robustness of machine learning models. *arXiv preprint arXiv:1707.04131*, 2017. 2

[45] Andrew Slavin Ross and Finale Doshi-Velez. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In *Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*, 2018. 3

[46] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 1, 3, 4

[47] Pouya Samangouei, Maya Kabkab, and Rama Chellappa. Defense-gan: Protecting classifiers against adversarial attacks using generative models. In *International Conference on Learning Representations (ICLR)*, 2018. 3

[48] Ali Shafahi, Mahyar Najibi, Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 8

[49] Aman Sinha, Hongseok Namkoong, and John Duchi. Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations (ICLR)*, 2018. 3

[50] Yang Song, Taesup Kim, Sebastian Nowozin, Stefano Ermon, and Nate Kushman. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2018. 3

[51] Yang Song, Rui Shu, Nate Kushman, and Stefano Ermon. Constructing unrestricted adversarial examples with generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 2

[52] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 4, 5

[53] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014. 1, 2

[54] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Dan Boneh, and Patrick McDaniel. Ensemble adversarial train-

ing: Attacks and defenses. In *International Conference on Learning Representations (ICLR)*, 2018. 3, 4, 5

[55] Jonathan Uesato, Brendan O'Donoghue, Aaron van den Oord, and Pushmeet Kohli. Adversarial risk and the dangers of evaluating against weak attacks. In *International Conference on Machine Learning (ICML)*, 2018. 1, 3, 4

[56] Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning (ICML)*, 2018. 3

[57] Eric Wong, Frank Schmidt, Jan Hendrik Metzen, and J Zico Kolter. Scaling provable adversarial defenses. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 3, 4, 5

[58] Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song. Spatially transformed adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2018. 2

[59] Kai Y Xiao, Vincent Tjeng, Nur Muhammad Shafiullah, and Aleksander Madry. Training for faster adversarial robustness verification via inducing relu stability. In *International Conference on Learning Representations (ICLR)*, 2019. 3

[60] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. In *International Conference on Learning Representations (ICLR)*, 2018. 3, 4, 5

[61] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 4, 5

[62] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3, 4

[63] Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. In *Proceedings of the Network and Distributed System Security Symposium (NDSS)*, 2018. 3, 4, 5

[64] Ziang Yan, Yiwen Guo, and Changshui Zhang. Deep defense: Training dnns with improved adversarial robustness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 3, 4, 5

[65] Huan Zhang, Hongge Chen, Zhao Song, Duane Boning, Inderjit S Dhillon, and Cho-Jui Hsieh. The limitations of adversarial training and the blind-spot attack. In *International Conference on Learning Representations (ICLR)*, 2019. 1

[66] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P Xing, Laurent El Ghaoui, and Michael I Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning (ICML)*, 2019. 3, 4, 5

[67] Y. Zhang and P. Liang. Defending against whitebox adversarial attacks via randomized discretization. In *Artificial Intelligence and Statistics (AISTATS)*, 2019. 3, 4, 5