Benchmarking Adversarial Robustness on Image Classification

Yinpeng Dong\textsuperscript{1}, Qi-An Fu\textsuperscript{1}, Xiao Yang\textsuperscript{1}, Tianyu Pang\textsuperscript{1}, Hang Su\textsuperscript{1}\textsuperscript{*}, Zihao Xiao\textsuperscript{2}, Jun Zhu\textsuperscript{1}\textsuperscript{*}

\textsuperscript{1}Dept. of Comp. Sci. and Tech., BNrist Center, Institute for AI, THBI Lab
\textsuperscript{1}Tsinghua University, Beijing, 100084, China \textsuperscript{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 \cite{Goodfellow2014, Carlini2017}, which are maliciously generated to induce erroneous predictions. As DL models have been integrated into various security-sensitive applications (\textit{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 \cite{Athalye2018}.

The research on adversarial robustness is faced with an “arms race” between attacks and defenses, \textit{i.e.}, a defense method proposed to prevent the existing attacks was soon evaded by new attacks, and vice versa \cite{Carlini2017, Kurakin2017a, Szegedy2013, Madry2017, Papernot2016, Tramer2017}. For instance, defensive distillation \cite{Madry2017} was proposed to improve adversarial robustness, but was later shown to be ineffective against a strong attack \cite{Carlini2017}. Many methods were introduced to build robust models by causing obfuscated gradients, which can be defeated by the adaptive ones \cite{Arjovsky2017, Papernot2016}. 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 \cite{Liu2017} and the minimum distance of the adversarial perturbation \cite{Athalye2018} 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 $\ell_p$ 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 \cite{Krizhevsky2009} and 8 on ImageNet \cite{Deng2009}. 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) $\ell_\infty$ and $\ell_2$ attacks; 3) white-box, transfer-based, score-based, and decision-based attacks.

*Hang Su and Jun Zhu are corresponding authors.
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 RealSafe1 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 $C(x) : \mathcal{X} \rightarrow \mathcal{Y}$, where $x \in \mathcal{X} \subset \mathbb{R}^d$ is the input, and $\mathcal{Y} = \{1, 2, ..., L\}$ with $L$ being the number of classes. Let $y$ denote the ground-truth label of $x$, and $x^{adv}$ denote an adversarial example for $x$.

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 $C(x^{adv}) \neq y$. A targeted one is crafted to be misclassified as the adversary-desired target class by the classifier, as $C(x^{adv}) = y^*$, 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 $\ell_p$ norm threat models, although there also exist other threat models [58, 51, 18]. Under the $\ell_p$ norm threat models, the adversary is allowed to add a small perturbation measured by the $\ell_p$ norm to the original input. Specifically, we consider the $\ell_\infty$ and $\ell_2$ norms.

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 $x^{adv}$ that satisfies $\|x^{adv} - x\|_p \leq \epsilon$, where $\epsilon$ 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

$$x^{adv} = \arg \max_{x : \|x' - x\|_p \leq \epsilon} \mathcal{J}(x', y).$$

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

$$x^{adv} = \arg \min_{x : x' \text{ is adversarial}} \|x' - x\|_p.$$ 

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

1Code released at: https://github.com/thu-ml/realsafe.
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. \( N^{\text{ATTACK}} [30] \) 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
advancing 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 \( A_{\epsilon,p} \) that generates an adversarial example \( x^{adv} = A_{\epsilon,p}(x) \) for an input \( x \) with perturbation budget \( \epsilon \) under the \( \ell_p \) norm\(^2\), and a (defense) classifier \( C \) defined in Sec. 2, the accuracy of the classifier against the attack is defined as

\[
\text{Acc}(C, A_{\epsilon,p}) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(C(A_{\epsilon,p}(x_i)) = y_i),
\]

where \( \{x_i, y_i\}_{i=1}^{N} \) is the test set, \( \mathbb{I}() \) is the indicator function. The attack success rate of an untargeted attack on the classifier is defined as

\[
\text{Asr}(A_{\epsilon,p}, C) = \frac{1}{M} \sum_{i=1}^{M} \mathbb{I}(C(x_i) = y_i \land C(A_{\epsilon,p}(x_i)) \neq y_i),
\]

where \( M = \sum_{i=1}^{N} \mathbb{I}(C(x_i) = y_i) \), while the attack success rate of a targeted attack is defined as

\[
\text{Asr}(A_{\epsilon,p}, C) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(C(A_{\epsilon,p}(x_i)) = y_i^*),
\]

where \( y_i^* \) is the target class corresponding to \( x_i \).

The previous methods usually report the point-wise accuracy or attack success rate for some chosen perturbation budgets \( \epsilon \), 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 \( \epsilon \). This can be efficiently done for attacks that find the minimum perturbations, by counting the number of the adversarial examples, the \( \ell_p \) norm of whose perturbations is smaller than each \( \epsilon \). For attacks that craft adversarial examples with constrained perturbations, we perform a binary search on \( \epsilon \) to find its minimum value that enables the generated adversarial example to fulfill the adversary’s goal.

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

---

\( ^2 \)For attacks that find minimum perturbations, e.g., DeepFool, C&W, we let \( A_{\epsilon,p}(x) = x \) if the \( \ell_p \) norm of the perturbation is larger than \( \epsilon \).
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) [57]. 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 N\(\text{\textsuperscript{\textdagger}}\)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 \(\ell_\infty\) norm in this section, and leave the full experimental results (including targeted attacks under the \(\ell_\infty\) norm, untargeted and targeted attacks under the \(\ell_2\) 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 \(\ell_\infty\) attacks and \(\epsilon = 1.0\) for \(\ell_2\) 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 \(\ell_\infty\) 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 $\ell_\infty$ 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 pertur-
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 \( \ell_\infty \) 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 \( \ell_2 \) attacks, we present the accuracy curves of the 8 models against untargeted Boundary and Evolutionary attacks under the \( \ell_2 \) 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 \( \ell_2 \) distance defined as \( \tilde{\ell}_2(\mathbf{a}) = \|\mathbf{a}\|_2/\sqrt{d} \) as the measurement for \( \ell_2 \) attacks, where \( d \) is the dimension of a vector \( \mathbf{a} \). To get the accuracy (attack success rate) vs. attack strength curves, we fix the perturbation budget as \( \epsilon = 16/255 \) for \( \ell_\infty \) attacks and \( \epsilon = \sqrt{0.001} \) for \( \ell_2 \) 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 \( \ell_\infty \) 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, and MIM under the \( \ell_\infty \) 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 \( \ell_\infty \) norm, while Fig. 10 and Fig. 12 show the two sets of curves for untargeted decision-based attacks under the \( \ell_2 \) 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).
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 $\ell_\infty$ threat model), but can also generalize to other threat models (e.g., the $\ell_2$ 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) slightly 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 NIVAIL Program with GPU/DGX Acceleration. Yinpeng Dong is supported by MSRA, Baidu Fellowships.
References


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


