Random Noise Defense Against Query-Based Black-Box Attacks

Anonymous Author(s)
Affiliation
Address
email

Abstract

The query-based black-box attacks have raised serious threats to machine learning models in many real applications. In this work, we study a lightweight defense method, dubbed *Random Noise Defense* (RND) which adds proper Gaussian noise to each query. We conduct the theoretical analysis about the effectiveness of RND against query-based black-box attacks and the corresponding adaptive attacks. Our theoretical results reveal that the defense performance of RND is determined by the magnitude ratio between the noise induced by RND and the noise added by the attackers for gradient estimation or local search. The large magnitude ratio leads to the stronger defense performance of RND and it's also critical for mitigating the adaptive attacks. Based on our analysis, we further propose to combine RND with a plausible Gaussian augmentation Fine-tuning (RND-GF). It enables RND to add larger noise to each query while maintaining the clean accuracy, to obtain a better trade-off between clean accuracy and defense performance. Additionally, RND can be flexible combined with the existing defense methods to further boost the adversarial robustness, such as adversarial training (AT). Extensive experiments on CIFAR-10 and ImageNet verify our theoretical findings and the effectiveness of RND and RND-GF.

1 Introduction

Deep neural networks (DNNs) have been successfully applied in many safety-critical tasks, such as autonomous driving, face recognition and verification, *etc*. However, it has been shown that DNN models are vulnerable to adversarial examples [19, 45], which are indistinguishable from natural examples but make a model produce erroneous predictions. For real-world applications, the DNN model as well as the training dataset, are often hidden from users. Instead, only the model feedback for each query (*e.g*., labels or confidence scores) are accessible. In this case, the product providers mainly face the severe threats from query-based black-box attacks, which don't require any knowledge about the attacked models.

In this work, we focus on efficient defense techniques against query-based black-box attacks, of which the main challenges are **1)** the defender should not significantly influence the model's feedback to normal queries, but it is difficult to know whether a query is normal or malicious; **2)** the defender has no information about what kinds of black-box attack strategies adopted by the attacker. Considerable efforts have been devoted to improving the adversarial robustness of DNNs [13, 31, 46, 47]. Among them, adversarial training (AT) is considered as the most effective defense techniques [3, 46]. However, the improved robustness from AT is often accompanied by significant degradation of the clean accuracy. Besides, the training cost of AT is much higher than that of standard training, and the it also often suffers from poor generalization to new samples or adversarial attacks [18, 36, 42, 43]. Thus, we argue that AT based defense is not a very suitable choice for black-box defense. In contrast, we expect that a good defense technique should satisfy the following requirements: *well keeping clean accuracy, lightweight, and plug-and-play*.

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

To this end, we study a lightweight defense strategy, dubbed *Random Noise Defense* (RND) against query-based black-box attacks. For query-based attacks [1, 2, 7, 9, 10, 11, 21, 24, 25, 29, 32], the core is to find a attack direction by gradient estimation or random search based on the exact feedback of consecutive queries, which leads to a decrease of the designed objective. RND is realized by adding a random noise to each query at the inference time. Therefore, the returned feedback with randomness result in the poor gradient estimation or random search and slow down the attack process.

To better understand the effectiveness of RND, we provide the theoretical analysis about the defense performance of RND against query-based attacks and the adaptive attacks. Our theoretical results reveal that the defense performance of RND is determined by the magnitude ratio between the noise induced by RND and the noise added by the attackers for gradient estimation or random search. For large magnitude ratio, the attack efficiency is greatly affected and the attackers need much more queries to find the adversarial examples or fail to find a success attack under query limited settings. That is, the larger ratio leads the better defense performance of RND. Apart from standard attacks, the adaptive attack (EOT) [3] has been considered as an effective strategy to mitigate the random effect induced by the defenders. We also conduct theoretical analysis about defense effect of RND against EOT attacks and demonstrate that the magnitude ratio is also crucial to mitigate adaptive attacks and the adaptive attacks has limited effect to evade the RND. On the other hand, large random noises to each query may lead to degradation of clean accuracy. To achieve a better trade-off between the defense effect and the clean accuracy while maintaining training time efficiency, we further propose to combine RND with a lightweight Gaussian augmentation Fine-tuning (RND-GF). RND-GF enables us to adopt larger noise in inference time to better disturb the query process.

We conduct extensive experiments on CIFAR-10 and ImageNet and the experimental results verify our theoretical results and demonstrate the effectiveness of RND-GF. It is worth noting that RND is plug-and-play and easy to be combined with existing defense methods such as AT to further boost the defense performance. To verify these, we evaluate the performance of RND combined with AT [20] and find that the RND can improve the robust accuracy of AT by up to 23.1% against the SOTA black-box attack Square attack [2] with maintaining clean accuracy.

The main contributions of this work are four-fold.

- We study an lightweight random noise defense (RND) against black-box attacks theoretically and empirically. Our theoretical results reveal that the effectiveness of RND is determined by the the magnitude ratio between the noise induced by RND and the noise added by the attackers for gradient estimation and random search.
- We theoretically analyse the performance of RND against the adaptive attack (EOT) and demonstrate that EOT has limited effect to evade the RND.
- Leveraging our theoretical analysis, we further propose an efficient and stronger defense strategy RND-GF by combining the Gaussian augmentation Fine-tuning and RND, towards a better trade-off between clean and adversarial performance.
- Extensive experiments verify our theoretical analysis and show the effectiveness of our defense methods against several state-of-the-art query-based attacks.

2 Related Work

Query-based Methods Here we mainly review the query-based black-box attack methods, which can be categorized into two classes, including ***gradient estimation*** and ***search-based*** methods. Gradient estimation methods are based on zero-order (ZO) optimization algorithms. In search-based methods, the attacker utilizes a direct search strategy to find the search direction, instead of explicitly estimating gradient. Furthermore, for our theoretical analysis, we focus on the *score-based* queries, where the continuous score (*e.g.*, the posterior probability or the logit) for each query is returned, in contrast to the decision-based queries which return hard labels. Specifically, [24] proposed the first limited query-based attack method by utilizing the Natural Evolutionary Strategies (NES) to estimate the gradient. [29] proposed the ZOsignSGD algorithm which is similar to NES. Based on Bandit Optimization, [25] proposed to combine the time and data dependent gradient prior with gradient estimation, which dramatically reduced the number of queries. For the ℓ_2 norm constrain, SimBA [21] randomly sampled a perturbation from orthonormal basises. SignHunter [1] focuses on estimating the sign of gradient and flipped the sign of perturbation to improve query efficiency. Square attack [2] is the state-of-art query-based attack method which selects localized square shaped updates at random positions of images.

Black-Box Defense Compared with the defense for white-box attacks, the defense specially designed for black-box attacks has not been well studied. Two recent works [8, 28] proposed to detect malicious queries based on the comparison with the history queries, since the malicious queries are more similar with each other compared with normal queries. AdvMind [34] proposed to infer the intent of adversary, and it also needed to store the history queries. However, if the attacker adopted the strategy of long-interval malicious queries, the defender has to store a long history, with very high cost of storage and comparison. [39] proposed the first black-box certified defense method, dubbed denoised smoothing. [5] also showed that adding random noise can defend against query-based attacks through experimental evaluations, without theoretical analysis of the defense effect. There are also a few randomization-based defenses. R&P [48] proposed a random input transform-based method. RSE [30] added large Gaussian noise to both the input and activation of each layer. PNI [23] combined adversarial training with adding Gaussian noise to the input or weight of each layer. However, these methods significantly sacrificed the accuracy of benign examples and also have the huge training cost. PixelDP [27] and random smoothing [13, 38] proposed to train classifiers with large Gaussian noise to get certified robustness under the ℓ_2 norm. However, they required to obtain a majority prediction of this query. Obviously, that places a huge burden to model inference and even helps the attacker get the more accurate gradient estimation. The too large Gaussian noise used in those methods also sacrifice clean accuracy. In contrast, RND maintains the good clean accuracy and only perturbs each query once, without any extra burden. [37] showed that the model with Gaussian augmentation training achieves the state-of-art defense against common corruptions, while the defense to black-box attacks is not evaluated.

3 Preliminaries

3.1 Score-based Black-Box Attack

In this work, we mainly focus on the score-based attacks. The analysis and evaluation of decision-based attacks are shown in **supplementary materials**. We denote the attacked model as $\mathcal{M} : \mathcal{X} \to \mathcal{Y}$ with $\mathcal{X}$ being the input space and $\mathcal{Y}$ being the output space. Given a benign data $(\boldsymbol{x}, y) \in (\mathcal{X}, \mathcal{Y})$, the goal of adversarial attack is to generate an adversarial example $\boldsymbol{x}_{adv}$ that is similar with $\boldsymbol{x}$, but to enforce prediction of $\mathcal{M}$ to be different with the true label y (*i.e.*, untargeted attack) or to be a target label h (*i.e.*, targeted attack). The optimization of untargeted attack can be formulated as follows,

$$\min_{\boldsymbol{x}_{adv} \in \mathcal{N}_R(\boldsymbol{x})} f(\boldsymbol{x}_{adv}) = \min_{\boldsymbol{x}_{adv} \in \mathcal{N}_R(\boldsymbol{x})} (\mathcal{M}_y(\boldsymbol{x}_{adv}) - \max_{j \neq y} \mathcal{M}_j(\boldsymbol{x}_{adv})). \tag{1}$$

For targeted attack, the objective function is $\max_{j \neq h} \mathcal{M}_j(\boldsymbol{x}_{adv}) - \mathcal{M}_h(\boldsymbol{x}_{adv})$. $\mathcal{M}_j(\boldsymbol{x}_{adv})$ denotes the logit or softmax output *w.r.t.* class j. $\mathcal{N}_R(\boldsymbol{x}) = \{\boldsymbol{x}' | \, \|\boldsymbol{x}' - \boldsymbol{x}\|_p \leq R\}$ indicates a ℓ_p ball around $\boldsymbol{x}$ (p is often specified as 2 or ∞), with $R > 0$. In attack evaluation, as long as $\mathcal{L}(\boldsymbol{x}_{adv})$ is less than 0, attackers consider the attack to be successful.

The score-based attack algorithms commonly adopt the projected gradient descent algorithms,

$$\boldsymbol{x}_{t+1} = \mathrm{Proj}_{\mathcal{N}_R(\boldsymbol{x})}(\boldsymbol{x}_t - \eta_t g(\boldsymbol{x}_t)). \tag{2}$$

For white-box attacks, $g(\boldsymbol{x}_t)$ is the gradient of $f(\boldsymbol{x})$ *w.r.t.* $\boldsymbol{x}_t$. However, for black-box attacks, the gradient $g(\boldsymbol{x}_t)$ cannot be directly obtained. So the attackers utilize the gradient estimation or random search to conduct the update $g(\boldsymbol{x}_t)$.

3.2 Zero-Order Optimization for Black-Box Attack

Zero-order optimization (ZO) has become the mainstream of black box attacks, dubbed ZO attack. The general idea of ZO attack methods is to estimate the gradient according to the objective values returned by querying. The gradient estimator widely used in score-based and decision-based attacks [7, 10, 12, 12, 16, 24, 25, 29, 33] is

$$g_\mu(\boldsymbol{x}) = \frac{f(\boldsymbol{x} + \mu \boldsymbol{u}) - f(\boldsymbol{x})}{\mu} \boldsymbol{u}, \tag{3}$$

where $\boldsymbol{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and $\mu \geq 0$. Based on this gradient estimator, the update of attack becomes

$$\boldsymbol{x}_{t+1} = \mathrm{Proj}_{\mathcal{N}_R(\boldsymbol{x})}(\boldsymbol{x}_t - \eta_t g_\mu(\boldsymbol{x}_t)). \tag{4}$$

3.3 Random Search for Black-Box Attack

The search-based attacks also depend on the value of $f(\boldsymbol{x} + \mu \boldsymbol{u}) - f(\boldsymbol{x})$ to find attack directions. Here, $\boldsymbol{u}$ is a searching direction sampled from some pre-defined distributions, such as gaussian noise in [12], orthogonal basis in [21] and squared perturbations in [2]. The μ is the size of perturbation in each search step such as the size of square in Square attack [2]. The work of [1, 32] adopt the fixed size schedule. If $f(\boldsymbol{x} + \mu \boldsymbol{u}) - f(\boldsymbol{x}) < 0$, the attackers consider $\boldsymbol{u}$ as a valuable attack direction. So,

the direction searching becomes

$$s(\boldsymbol{x}) = \mathbb{I}(h(\boldsymbol{x}) < 0) \cdot \mu \boldsymbol{u} \text{ where } h(\boldsymbol{x}) = f(\boldsymbol{x} + \mu \boldsymbol{u}) - f(\boldsymbol{x}), \tag{5}$$

where, $\mathbb{I}$ is the indicator function. And the updating of search-based attacks becomes

$$\boldsymbol{x}_{t+1} = \mathrm{Proj}_{\mathcal{N}_R(\boldsymbol{x})}(\boldsymbol{x}_t + s(\boldsymbol{x}_t)). \tag{6}$$

4 RND: Random Noise Defense Against Query-based Attack Methods

4.1 Random Noise Defense

In the gradient estimator (*e.g.*, Eq.(3)) and searching direction (*e.g.*, Eq.(5)), the attacker adds one small random perturbation $\mu \boldsymbol{u}$ to get the objective difference between two queries. Thus, if the defender can further disturb this random perturbation, the gradient estimation or searching direction are expected to be misled, so as to decrease the attack efficiency. However, it's also hard for defenders to identify whether a query is normal or malicious. Inspired by those observations, we study an lightweight defense method, dubbed Random Noise Defense (RND), that simply adds a random noise to each query. For RND, the feedback for one query $\boldsymbol{x}$ is $\mathcal{M}(\boldsymbol{x} + \nu \boldsymbol{v})$, with $\boldsymbol{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is standard Gaussian noise added by the defender, and the factor $\nu > 0$ controls the magnitude of random noise. Considering the task of defending query-based black-box attacks, RND should satisfies: **1)** the prediction of each individual query will not be changed significantly; **2)** the estimated gradient or direction searching should be perturbed as large as possible towards better defense performance. The gradient estimator and searching direction under RND become

$$g_{\mu,\nu}(\boldsymbol{x}) = \frac{f(\boldsymbol{x} + \mu \boldsymbol{u} + \nu \boldsymbol{v}_1) - f(\boldsymbol{x} + \nu \boldsymbol{v}_2)}{\mu} \boldsymbol{u}, \tag{7}$$

$$s_\nu(\boldsymbol{x}) = \mathbb{I}(h_\nu(\boldsymbol{x}) < 0) \cdot \mu \boldsymbol{u} \text{ where } h_\nu(\boldsymbol{x}) = f(\boldsymbol{x} + \mu \boldsymbol{u} + \nu \boldsymbol{v}_1) - f(\boldsymbol{x} + \nu \boldsymbol{v}_2), \tag{8}$$

where $\boldsymbol{v}_1, \boldsymbol{v}_2 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ are both standard Gaussian noise generated by the defender.

To satisfy the first requirement, one cannot add too large noise to each query. However, to satisfy the second requirement, ν should be large enough to change the gradient estimation or searching direction. We need to choose a proper ν to achieve a good trade-off between this two requirements. In the following, we provide the theoretical analysis of RND, which can shed light on the setting of ν.

4.2 Theoretical Analysis of Random Noise Defense Against ZO Attacks

In this section, we will present the theoretical analysis about the effect of RND against ZO attacks. Specifically, we study the convergence property of ZO attack in Eq.(4) with $g_{\mu,\nu}(\boldsymbol{x})$ in Eq.(7) being the gradient estimator. Throughout our analysis, the measure of adversarial perturbation is specified as ℓ_2 norm, corresponding to $\mathcal{N}_R(\boldsymbol{x}) = \{\boldsymbol{x}' | \|\boldsymbol{x}' - \boldsymbol{x}\|_2 \leq R\}$. To facilitate subsequent analyses, we first introduce some assumptions, notations, and definitions.

Assumption 1. *$f(\boldsymbol{x})$ is Lipschitz-continuous,* i.e., $|f(\boldsymbol{y}) - f(\boldsymbol{x})| \leq L_0(f)\|\boldsymbol{y} - \boldsymbol{x}\|$.

Assumption 2. *$f(\boldsymbol{x})$ is continuous and differentiable, and $\nabla f(\boldsymbol{x})$ is Lipschitz-continuous,* i.e., $\|\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})\| \leq L_1(f)\|\boldsymbol{y} - \boldsymbol{x}\|$.

Notations. We denote the sequence of standard Gaussian noises added by the attacker as $\boldsymbol{\mathcal{U}}_t = \{\boldsymbol{u}_0, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_t\}$, with t being the iteration index in the update Eq.(4). The sequence of standard Gaussian noises added by the defender is denoted as $\boldsymbol{\mathcal{V}}_t = \{\boldsymbol{v}_{01}, \boldsymbol{v}_{02}, \ldots, \boldsymbol{v}_{t1}, \boldsymbol{v}_{t2}\}$. The generated sequential solutions are denoted as $\{\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_Q\}$, and the benign example $\boldsymbol{x}$ is used as the initial solution, *i.e.*, $\boldsymbol{x}_0 = \boldsymbol{x}$. $d = |\mathcal{X}|$ denotes the input dimension.

Definition 1. *The Gaussian-Smoothing function corresponding to $f(\boldsymbol{x})$ with $\nu > 0, \boldsymbol{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is*

$$f_\nu(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}} \int f(\boldsymbol{x} + \nu \boldsymbol{v}) \cdot \mathrm{e}^{-\frac{1}{2}\|\boldsymbol{v}\|_2^2} \, \mathrm{d}\boldsymbol{v}. \tag{9}$$

Due to the noise inserted by the defender, f_ν becomes the objective function for the attacker. We also define the minimum of $f_\nu(\boldsymbol{x})$, $f_\nu^* = f_\nu(\boldsymbol{x}^*) = \min_{\boldsymbol{x} \in \mathcal{N}_R(\boldsymbol{x}_0)} f_\nu(\boldsymbol{x})$.

4.2.1 Theoretical Analysis under General Non-Convex Case

Here, we report theoretical analysis of RND under general non-convex case satisfying Assumption 1.

Theorem 1. *Under Assumption 1, for any $Q \geq 0$, consider a sequence $\{\boldsymbol{x}_t\}_{t=0}^Q$ generated according to the descent update Eq.(4) using the gradient estimator $g_{\mu,\nu}(\boldsymbol{x})$ in Eq.(7) , with the constant stepsize,*

$\eta = \left[\frac{R\epsilon}{\gamma(\alpha)d^3 L_0^3(f)(Q+1)}\right]^{1/2}$ *Then,*

$$\frac{1}{Q+1}\sum_{t=0}^{Q}\mathbb{E}_{\mathcal{U}_t,\mathcal{V}_t}(\|\nabla f_{\mu,\nu}(\boldsymbol{x}_t)\|^2) \leq \frac{2L_0(f)^{\frac{5}{2}}R^{\frac{1}{2}}d^{\frac{3}{2}}}{(Q+1)^{\frac{1}{2}}\epsilon^{\frac{1}{2}}}\sqrt{\gamma(\alpha)} \tag{10}$$

where $\frac{\nu}{\mu} = \alpha$ *and* $\gamma(\alpha) = (\frac{1}{2} + \sqrt{2}\alpha + \alpha^2)$*, which is always increasing function of* α*.*

Remark 1. *To bound the gap* ϵ *between* $f_{\mu,\nu}(\boldsymbol{x})$ *and* $f_\nu(\boldsymbol{x})$*,* i.e.*,* $|f_{\mu,\nu}(\boldsymbol{x}) - f_\nu(\boldsymbol{x})| \leq \epsilon$*,* $\forall\ \boldsymbol{x} \in \mathcal{N}_R(\boldsymbol{x})$*, we could choose* $\mu \leq \hat{\mu} = \frac{\epsilon}{d^{1/2}L_0(f)}$ *[33]. Due to the non-convexity assumption, we only guarantee the convergence to a stationary point of the function* $f_{\mu,\nu}(\boldsymbol{x})$*, which is a smoothing approximation of* f_ν*. The minimum of upper bound is denoted as* δ*, then the upper bound for the expected number of queries is* $O\left(\gamma(\alpha)\frac{d^3L_0^5(f)R}{\epsilon\delta^2}\right)$*.*

The convergence rate of ZO attacks is rather important, since the attackers need to effectively find adversarial examples within limited queries. Theorem 1 shows the convergence rate is positive related to the ratio $\frac{\nu}{\mu}$. The larger ratio $\frac{\nu}{\mu}$ will lead to the higher upper bound of convergence error and slower convergence rate. In consequence, the attackers need much more queries to find adversarial examples. Specifically, if $\frac{\nu}{\mu} \ll 1$, the query complexity is equivalent to the original constant term $O(\frac{d^3}{\epsilon\delta^2})$. When $\frac{\nu}{\mu} \geq 1$, the query complexity is really improved over $O(\frac{d^3}{\epsilon\delta^2})$. Therefore, under the queries limited setting, the attack efficiency will be significantly reduced, leading to failed attacks or a much larger number of queries for successful attacks. Therefore, **the large ratio $\frac{\nu}{\mu}$ leads the effectiveness of RND.** It is in accordance with our intuition that the defenders should insert larger noise (*i.e.*, ν) than that added by the attack (*i.e.*, μ) to achieve the satisfied defense effect.

Trade-off of Larger ν and Clean Accuracy: If $f(\boldsymbol{x})$ is Lipschitz-continuous, then $|f_\nu(\boldsymbol{x}) - f(\boldsymbol{x})| \leq \nu L_0(f)d^{1/2}$ [33]. The larger ν is, the larger the gap between $f_\nu(\boldsymbol{x})$ and $f(\boldsymbol{x})$. So the clean accuracy of model with adding larger noise will also decrease. This is also shown in Figure 1 (d), which forms a trade-off between defense performance of RND and clean accuracy.

Larger Noise Size μ Adopted by Attackers: The attacker may be aware of the defense mechanism, so they can also increase the adopted noise size μ. Our experiment results also verify this point. As shown in Figure 2 (a), for NES attack, the attack failure rate is almost 0, when $\nu = \mu = 0.01$. This shows adding small noise can't always guarantee the effective defense. Previous work [5, 15] don't consider about this and overestimate the effect of small noise. However, increasing the noise size μ will also lead less accurate gradient estimation and random search in Eq.(3) and Eq.(5), leading to a significant decrease in attack performance consequentially. Taking the NES [24] and Bandit attack [25] on ImageNet as examples, as shown in Figure 3, when μ increase from 0.01 to 0.05, the ℓ_∞ attack failure rates increase from 5.1%, 5.0% to 22.8%, 25.7% respectively. For Square attack [2], when μ increase from 0.3 to 0.5, the average number of queries has doubled. To guarantee the successful attack, attackers cannot always increase μ. Our experimental results show when the μ of ZO attacks and Square attack is larger than 0.01 and 0.3, the attack performance decreases significantly. Based on the above analysis, we propose the stronger defense method in Section 4.4. It enables us to add larger noise ($\nu > \mu$) and still maintains the good clean accuracy.

4.2.2 Theoretical Analysis of Random Noise Defense Against Adaptive Attacks

As suggested in recent studies of robust defense [3, 6, 46], the defender should take a robust evaluation against the corresponding adaptive attack, in which case the attacker is aware of the defense mechanism. Here we study the defense effect of RND against adaptive attacks. Since the idea of RND is to insert random noise to each query, an adaptive attacker could utilize Expectation Over Transformation (EOT) [4] to obtain an more accurate estimation, *i.e.*, querying one sample multiple times to obtain the average. Then, the gradient estimator used in ZO attacks Eq.(4) becomes

$$\tilde{g}_{\mu,\nu}(\boldsymbol{x}) = \frac{1}{M}\sum_{j=1}^{M}\frac{f(\boldsymbol{x}+\mu\boldsymbol{u}+\nu\boldsymbol{v}_{j1}) - f(\boldsymbol{x}+\nu\boldsymbol{v}_{j2})}{\mu}\boldsymbol{\mu}, \tag{11}$$

where $\boldsymbol{v}_{j1}, \boldsymbol{v}_{j2} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, with $j = 1, \ldots, M$. Note that here the definition of the sequential standard Gaussian noises added by the defender (see Section 4.2.1) should be updated to $\boldsymbol{\mathcal{V}}_t = \{\boldsymbol{v}_{01}, \ldots, \boldsymbol{v}_{0M}, \ldots, \boldsymbol{v}_{t1}, \ldots, \boldsymbol{v}_{tM}\}$. $\boldsymbol{v}_{ij} \in \boldsymbol{\mathcal{V}}_t$ contains $\boldsymbol{v}_{ij1}$ and $\boldsymbol{v}_{ij2}$. The convergence analysis of ZO attack with Eq.(11) against RND is presented in Theorem 2.

Theorem 2. *Under Assumption 1 and 2, for any $Q \geq 0$, consider a sequence $\{\boldsymbol{x}_t\}_{t=0}^{Q}$ generated according to the descent update Eq.(4) using the gradient estimator $\tilde{g}_{\mu,\nu}(\boldsymbol{x})$ Eq.(11), we have*

$$\begin{aligned}\frac{1}{S_Q}\sum_{t=0}^{Q}\eta_t\mathbb{E}_{\mathcal{U}_t,\mathcal{V}_t}(\|\nabla f_{\mu,\nu}(\boldsymbol{x}_t)\|^2) \leq & \frac{L_0(f)R}{S_Q} + \frac{1}{S_Q}\sum_{t=0}^{Q}\eta_t^2(L_0(f)^2L_1(f)d^2(\frac{1}{2}+\frac{2\nu^2}{\mu^2 M}) \\ & + \frac{\nu^2 L_0(f)L_1(f)^2}{\mu}d^{\frac{5}{2}} + \frac{\nu^4 L_1(f)^3(M+1)}{2\mu^2 M}d^3)\end{aligned} \tag{12}$$

The larger M for EOT: Theorem 2 shows that with larger M, the upper bound will be reduced. So, EOT can mitigate the defense effect caused by the randomness of RND. However, with M going to infinity, the upper bound of expected convergence error (*i.e.*, Eq. (12)) is still determined by the max term $\frac{\nu^4}{\mu^2}d^3$. **It implies that the attack improvement from EOT is limited, especially with the larger ratio $\frac{\nu}{\mu}$.** Experiments in Section 5.3 verify our analysis of EOT attack.

4.3 Theoretical Analysis of Random Noise Defense Against Search-based Attacks

In this section, we will show how RND affects the search-based attacks. Based on the Eq.(5) and Eq.(8), by adding noise $\nu\boldsymbol{v}$, the value of $h_\nu(\boldsymbol{x})$ will be different from that of $h(\boldsymbol{x})$, and there is certain probability that $\text{Sign}(h_\nu(\boldsymbol{x}))$ be different from $\text{Sign}(h(\boldsymbol{x}))$. When the random noise $\nu\boldsymbol{v}$ causes inconsistence between $\text{Sign}(h_\nu(\boldsymbol{x}))$ and $\text{Sign}(h(\boldsymbol{x}))$, RND will mislead the attackers to select the incorrect attack directions (*i.e.*, abandoning the descent direction *w.r.t.* f or selecting the ascent direction), so as to decrease the attack performance. We have the following theoretical analysis about this.

Theorem 3. *Under Assumption 1, considering the direction update Eq.(6) with Eq.(8) in search-based attacks, we have,*

$$P(\textit{Sign}(h(\boldsymbol{x})) \neq \textit{Sign}(h_\nu(\boldsymbol{x})) \leq \frac{2L_0(f)\nu\sqrt{d}}{|h(\boldsymbol{x})|} \tag{13}$$

The probability P is intuitively controlled by the relative values of the $h_\nu(\boldsymbol{x})$ and $h(\boldsymbol{x})$. The proof is shown in Section F of supplementary materials. Theorem 3 shows the probability of misleading attacker is positive correlated with $\frac{\nu}{|h(\boldsymbol{x})|}$. Here, $|h(\boldsymbol{x})| = |f(\boldsymbol{x}+\mu\boldsymbol{u}) - f(\boldsymbol{x})| \leq L_0(f)\mu\|\boldsymbol{u}\|$. Recall f is a smooth function and the perturbation is restricted in small neighborhoods around $\boldsymbol{x}$. The $|h(\boldsymbol{x})|$ is also positive correlated with the stepsize μ within the small neighborhoods. So the probability of changing the sign is positive correlated with $\frac{\nu}{\mu}$. **The larger ratio $\frac{\nu}{\mu}$ leads the larger the upper bound of the probability**. The attackers are more likely to be mislead and need much more queries to find adversarial examples. So, the defense performance of RND is better. As shown in Figure 1 (a-c), the evaluations on Square attack verify our analysis.

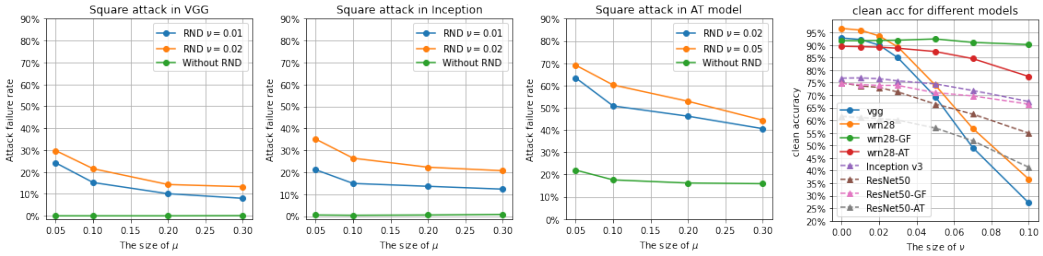
4.4 RND with Gaussian Augmentation Fine-tuning

Aforementioned theoretical analyses reveal that RND should choose a proper noise magnitude ν to achieve a good balance between clean accuracy and the defense effectiveness. To achieve a high-quality balance, we could reduce the sensitivity of the target model to random noises, such that the influence of the noise to the accuracy of each query will be reduced. To satisfy the lightweight requirement of black-box defense, we propose to utilize Gaussian Augmentation Fine-tuning (GF), which fine-tunes the deployed model with Gaussain augmentation. We add random Gaussian noise to each training sample, as a pre-processing step in the fine-tuning process. Consequently, the model fine-tuned with GF is expected to maintain good accuracy to each query, even though defenders add a relatively large random noise to each query. As shown in Figure 1 (d), **GF models still maintain the high clean accuracy under larger noise**.

5 Experiments

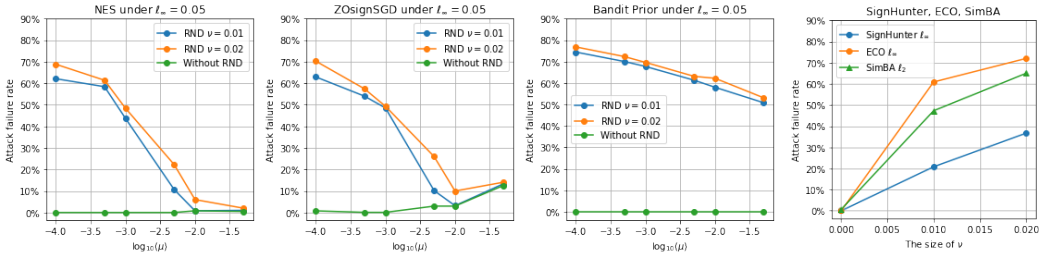
5.1 Experimental Settings

Datasets and Classification Models. We conduct experiments on two widely used benchmark datasets in adversarial machine learning: CIFAR-10 [26] and ImageNet [14]. For classification models, we use VGG-16 [40], WideResNet-16-10 and WideResNet-28-10 [51] on CIFAR-10. We conducted standard training and their clean accuracy on the test set is 92.76% 95.10, and 96.60%, respectively. For ImageNet, we adopt the pretrained Inception v3 model [44] and ResNet-50 model [22] provided by torchvision package and the clean accuracy are 76.80% and 74.90%, respectively.



(a) (b) (c) (d)

Figure 1: (a-c): Attack failure rate (%) of Square ℓ_∞ attacks on VGG-16(CIFAR-10), Inception v3(ImageNet) and AT model (ImageNet) under different values of μ and ν, where μ is the square size in Square attacks. (d): clean accuracy for different models on CIFAR-10 and ImageNet. The circle lines and triangle lines represent models on CIFAR-10 and ImageNet respectively.



(a) (b) (c) (d)

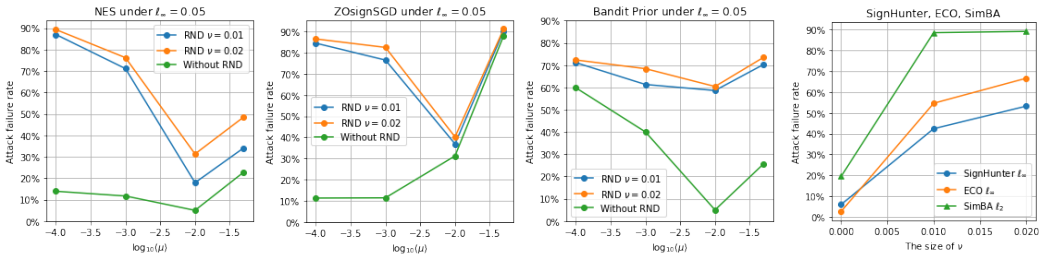
Figure 2: Attack failure rate (%) of query-based attacks on VGG-16 and CIFAR-10 under different values of μ and ν. We adopt logarithm scale in subplot (a-c) for better illustration. The complete evaluation under ℓ_2 attack is given in Section B.4 of supplementary materials.

Black-box Attack Methods and Compared Defense Methods. We consider several mainstreamed query-based black-box attack methods, including NES [24], ZO-signSGD (ZS) [29], Bandit [25], ECO [32], SimBA [21], SignHunter [1] and Square attack [2]. Note that the NES, ZS, Bandit are gradient estimation based attack methods and the other four methods are search-based methods. SimBA and ECO are only designed for ℓ_2 and ℓ_∞ attack, repectively. Following [24], we evaluate all the attack methods on whole test set of CIFAR-10 and 1,000 random sampled images from the validation set of ImageNet. We present the evaluation performance against the untargeted attack under both ℓ_∞ and ℓ_2 norm settings. The perturbation budget of ℓ_∞ is set to 0.05 for both datasets. For ℓ_2 attack, the perturbation budget is set to 1 and 5 on CIFAR-10 and ImageNet, respectively. The number of maximal queries is set to 10,000. **We adopt the attack failure rate and average number of query as evaluation metric.** The higher the attack failure rate and larger average number of query, the better the adversarial defense performance. We compare our methods with RSE [30] and PNI [23] on CIFAR-10. We adopt pre-trained AT model [20] which shows the better robustness than other AT models. For ImageNet, we compare pre-trained Feature Denoise (FD) model [49] and adopt pre-trained AT in Robustness Library [17]. The implementation details of those methods are given in the Section B.1 and B.2 of supplementary materials.

5.2 Evaluation of RND Against Query-based Black-Box Attacks

We first evaluate the defense performance of RND with various settings of μ and ν against query-based black-box attacks and verify the theoretical results in Section 4.2 and 4.3. Specifically, we set $\nu \in \{0.0, 0.01, 0.02\}$ and $\mu \in \{10^{-4}, 5{*}10^{-4}, 10^{-3}, 0.005, 0.01, 0.05\}$ for ZO attacks on CIFAR-10 and $\mu \in \{10^{-4}, 0.001, 0.01, 0.05\}$ on ImageNet. For Square attacks, we set $\mu \in \{0.05, 0.1, 0.2, 0.3\}$. Figure 1 (a-c) presents the defense performance of RND against Square attack on CIFAR-10 and ImageNet, repectively. Figure 2 (a-c) and 3 (a-c) present the defense performance of RND against NES, ZS and Bandit on CIFAR-10 and ImageNet, respectively.

The results in Figure 1 (a-c), 2 (a-c), and 3 (a-c) show that: the attack failure rate of all attack methods generally increases as the value of $\frac{\nu}{\mu}$ increases. Specifically, for a fixed value of ν (*e.g.*, 0.01), the attack failure rate increases as μ decreases. While for a certain value of μ (*e.g.*, 10^{-4}), the attack failure rate increases as ν increases. These collaborate our theoretical analysis that **the ratio of $\frac{\nu}{\mu}$ determines the probability of changing sign and the convergence rate of ZO attacks.** The larger $\frac{\nu}{\mu}$, the higher the probability of changing the sign and the convergence error of ZO attacks, which results in the poor attack performance under the query-limited settings.



(a) (b) (c) (d)

Figure 3: Attack failure rate (%) of query-based attacks on Inception v3 and ImageNet under different values of μ and ν. We adopt logarithm scale in subplot (a-c) for better illustration. The complete evaluation under ℓ_2 attack is given in Section B.4 of supplementary materials.

Table 1: The evaluation of EOT with ℓ_∞ attack on CIFAR-10 and ImageNet under the *adaptive and fixed query setting*. **The left part is the results on CIFAR-10 and the right part is on ImageNet.** The average number of query of successful attack as well as the attack failure rate are reported. The performance of EOT with ℓ_2 attack is reported in Section B.5 of supplementary materials.

settings	Methods	M=1	M=5	M= 10	Methods	M=1	M=5	M= 10
adaptive	NES	1448/0.484	4078/0.361	5763/0.342	NES	2532/0.762	5364/0.705	7582/0.691
	ZS	1489/0.493	3189/0.374	5912/0.349	ZS	2824/0.825	5735/0.761	7662/0.740
fixed	NES	1448/0.484	2528/0.452	3246/0.443	NES	2533/0.762	5240/0.775	5658/0.781
	ZS	1489/0.493	2765/0.448	3123/0.421	ZS	2824/0.825	4023/0.842	4652/0.861
	Bandit	436/0.696	276/0.582	314/0.543	Bandit	305/0.604	759/0.523	946/0.49
	Square	380/0.301	181/0.162	223/0.121	Square	93/0.353	145/0.20	328/0.171
	SignHunter	459/0.367	559/0.224	759/0.191	SignHunter	173/0.532	336/0.456	659/0.431
	ECO	904/0.720	1681/0.761	2560/0.793	ECO	1237/0.666	3065/0.678	3091/0.692
	SimBA	1353/0.650	3852/0.467	4103/0.396	SimBA	274/0.891	468/0.878	517/0.869

We also evaluate the defense performance of RND with various values of ν against other search-based black-box attacks and the results are shown in Figure 2 (d) for CIFAR-10 and Figure 3 (d) for ImageNet. As shown in the plots, the RND can significantly boost the defense performance and the attack failure rate increases as the value of ν increases. The whole evaluations, including the WideResNet-16 on CIFAR-10, the ℓ_2 attack as well as the mean and medium number of queries of success attacks are reported in Section B.4 of supplementary materials.

5.3 Evaluation of RND Against Adaptive Attacks

We then evaluate the defense effect of RND against the adaptive attack EOT and verify our theoretical analysis in Section 4.2.2. We evaluate EOT with $M \in \{1, 5, 10\}$ on CIFAR-10 and ImageNet, and ν is set to 0.02 and μ is set to 0.001 and 0.1 for ZO attacks and Square attack, respectively. The evaluation with even larger M and μ is presented in Section B.5 of supplementary materials.

We consider two settings of query budget: **1)** *adaptive query budget* that assigns query budget of 10,000 $\times M$ for different value of M; **2)** *fixed query budget* that adopt a fixed query budget of 10,000 for all M. We first evaluate the performance with adaptive query budget on NES and ZS and their numerical results are shown in Table 1. As shown in the Table 1, the attack failure rate decreases as M increases on both datasets. However, the average number of query of successful attack also greatly increases as M increases, which demonstrates that **the adaptive EOT attack can increase the attacking success rate with a sacrifice of query efficiency.** More importantly, we observe that **the relative performance improvements induced by EOT generally decrease as M increases.** These verify our theoretical analysis in Section 4.2.2.

The evaluation under the fixed query budget is also reported in Table 1. On small-scale dataset CIFAR-10, the attack failure rate of all attacks generally decreases as M increases. Yet we also observe the similar phenomenon that the relative performance improvements induced by EOT decreases as M increases. For the large-scale dataset ImageNet, EOT can increase the attack performance with a significantly sacrifice of query efficiency under most cases, except for NES and ZS. The attack performance of NES and ZS decreases as M increases. Specifically, for a fixed number of queries, the larger the M, the smaller the number of iterations for ZO attack under limited query setting. The poor performance of NES and ZS with larger M may be because the improvements induced by EOT are less than the potential degeneration caused by the decrease of iterations.

5.4 Evaluation of RND-GF and the Combination of RND with AT

In this section, we evaluate the performance of RND combined with Gaussian augmentation Fine-tuning, dubbed RNG-GF. In fine-tuning phase, we adopt the cyclic learning rate [41] to achieve superconvergence in 50 epochs. The training detail is shown in Section B.3 of supplementary

Table 2: The comparison of RND ($\nu = 0.02$), GF, RND-GF ($\nu = 0.05$), AT, RND-AT ($\nu = 0.05$), PNI, RSE, and FD on CIFAR-10 and Imagenet. The average number of queries of successful attack and the attack failure rates are reported. The best and second best attack failure rate under each attack are highlighted in bold and underlined, respectively. The evaluation under ℓ_2 attack is shown in Section B.6 of supplementary materials.

Datasets	Methods	Clean Acc	NES(ℓ_∞)	ZS(ℓ_∞)	Bandit(ℓ_∞)	Sign(ℓ_∞)	Square(ℓ_∞)	SimBA(ℓ_2)	ECO(ℓ_∞)
CIFAR-10 (WideNet-28)	Clean Model	**96.60**%	465.5/0.01	581.8/0.06	210.2/0.03	167.6/0.03	137.1/0.02	457.2/0.04	457.8/0.0
	GF	91.72%	999.0/0.407	759.9/0.544	744.5/0.116	348.3/0.027	581.0/0.061	1146.8/0.395	883.9/0.067
	RSE[30]	84.12%	1246.3/0.396	1327.8/0.422	281.7/0.372	243.7/0.221	413.3/0.243	498.3/0.337	578.3/0.534
	PNI[23]	87.20%	1071.4/0.725	1310.7/0.823	324.9/0.824	267.0/0.708	295.3/0.612	945.0/0.857	2342.2/0.623
	AT[20]	89.48%	821.6/0.807	614.9/0.862	1451.5/0.623	766.3/0.476	1135.4/0.499	1523.2/0.635	1180.4/0.484
	RND	93.60%	842.5/0.05	941.8/0.143	273.1/0.478	977.2/0.226	762.4/0.116	2112.6/0.549	912.8/0.688
	RND-GF	92.40%	2805.7/0.516	2966.3/0.730	1223.5/0.841	1017.1/0.407	1207.3/0.378	1220.2/0.863	687.2/0.872
	RND-AT	87.40%	2499.2/**0.842**	2625.7/**0.923**	891.5/**0.891**	767.9/**0.737**	1170.7/**0.730**	1787.4/**0.912**	687.4/ **0.911**
ImageNet (ResNet-50)	Clean Model	**74.90**%	1031.9/0.0	2013.0/0.235	329.2/0.02	264.1/0.03	76.5/0.0	1234.5/0.281	347.7/0.0
	GF[37]	74.70%	1685.5/0.03	1712.1/0.347	601.4/0.02	329.0/0.0	97.28/0.0	1417.4/0.112	362.4/0.0
	FD[49]	54.20%	1997.2/0.679	1555.5/0.775	1579.2/0.426	1633.1/0.332	1092.4/0.242	2607.9/0.613	1501.0/0.240
	AT[17]	61.60%	2113.4/0.724	1688.7/0.815	1091.5/0.416	1522.7/0.289	1109.0/0.159	2638.2/0.651	1440.6/0.200
	RND	73.00%	3041.5/0.245	2266.2/0.330	390.6/0.536	661.0/0.314	81.5/0.101	825.3/0.612	2435.5/0.540
	RND-GF	71.15%	2489.3/0.421	2053.5/0.563	495.9/0.603	514.0/0.348	1009.9/0.146	777.2/ 0.762	994.8/0.702
	RND-AT	58.15%	2556.6/**0.864**	2596.6/**0.870**	448.0/**0.810**	724.2/**0.632**	1306.3/**0.386**	1210.5/**0.953**	631.1/**0.865**

materials. We adopt the attack failure rate and average number of queries under the stronger adaptive attack EOT for evaluation. As that in the Section 5.2, we also tune the parameter μ for attackers and tune $M \in \{1, 5, 10\}$ for EOT to achieve the best attack performance for all attack methods.

The comparison with other defense methods is shown in Table 2, where GF refers to standard fine-tuning with Gaussian augmentation. RND denotes the clean model coupled with random noise defense and RND-GF denotes the GF model coupled with RND. As shown in the Table 2, RND can significant boost the defense performance of the Clean Model on both datasets. For example, RNG achieves $20\% \sim 40\%$ improvement against Bandit and SignHunter on on both datasets. The proposed GF can protect the clean accuracy from the larger noise induced by RND and we adopt a relative larger $\nu = 0.05$ for RND-GF towards better defense performance. Compared with RND, RNG-GF significantly improves the defense performance under all attack methods while maintaining the good clean accuracy. For example, RNG-GF achieves $40\% \sim 50\%$ and $20\% \sim 30\%$ improvements against ZO and search-based attacks on CIFAR-10. Compared with RSE, PNI, and FD, RND-GF achieves the higher failure rate against Bandit, SimBA, and ECO and leads to much more queries on other attacks. Besides, it also maintains the much better clean accuracy and lower training cost.

Without any extra modification, RND can be easily combined with many existing defense methods. We adopt the pre-trained AT WideNet-28 (ℓ_∞) model [20] and pre-trained AT ResNet-50 (ℓ_∞) model in Robustness Library [17] on CIFAR-10 and ImageNet, respectively. As shown in Figure 1 (d), AT models are less affected by adding noise like GF models. So we can adopt the larger noise ($\nu = 0.05$) in inference time. Combining AT with RND, RND-AT significantly improves the robustness against all attacks and achieves best performance among all methods. Based on AT models, RND-AT achieves 23.1% and 22.7% improvements against Square attack on CIFAR-10 and ImageNet.

5.5 Discussion and Limitations

As shown in Section 3.1, we focus on the score-based attacks. **The analysis and evaluation of RND against decision-based attacks are shown in Section C of supplementary materials**. Besides, as we mainly focus on practical query-based black-box attacks where the models as well as the training datasets are unknown to the attackers, we don't cover the attacks utilizing the transferability from surrogate to target models. These methods usually assume that the surrogate models are trained on the same training set with the target model. It is difficult to obtain the training set behind the target model in real scenarios. Meanwhile, there have been some defense strategies developed for transfer-based attacks [35, 47, 50]. It is interesting to explore the combination of RND and these works towards better defense performance against transfer-based attacks and we leave it to future works.

6 Conclusion

In this work, we study a lightweight black-box defense, dubbed Random Noise Defense (RND) for query-based black-box attacks, which is realized by adding proper Gaussian noise to each query. We give the first theoretical analysis of the effectiveness of RND against both standard and adaptive query-based black-box attacks. Based on our analysis, we further propose RND-GF which combines RND with Gaussian augmentation Fine-tuning to obtain the better trade-off between clean accuracy and robustness. Extensive experiments verify our theoretical analysis and demonstrate the effectiveness of the RND-GF against several state-of-the-art query-based attacks. Without any extra modification, RND can be easy to combined with the existing defense methods, such as AT. We further demonstrate that when combined with RND, RND-AT can significant boost the adversarial robustness.

References

[1] Abdullah Al-Dujaili and Una-May O'Reilly. Sign bits are all you need for black-box attacks. In *International Conference on Learning Representations*, 2020.

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

[3] 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*, pages 274–283, 2018.

[4] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *International conference on machine learning*, pages 284–293. PMLR, 2018.

[5] Junyoung Byun, Hyojun Go, and Changick Kim. Small input noise is enough to defend against query-based black-box attacks, 2021.

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

[7] Jianbo Chen, Michael I Jordan, and Martin J Wainwright. Hopskipjumpattack: A query-efficient decision-based attack. In *2020 ieee symposium on security and privacy (sp)*, pages 1277–1294. IEEE, 2020.

[8] Steven Chen, Nicholas Carlini, and David Wagner. Stateful detection of black-box adversarial attacks. In *Proceedings of the 1st ACM Workshop on Security and Privacy on Artificial Intelligence*, pages 30–39, 2020.

[9] Weilun Chen, Zhaoxiang Zhang, Xiaolin Hu, and Baoyuan Wu. Boosting decision-based black-box adversarial attacks with random sign flip. In *European Conference on Computer Vision*, pages 276–293. Springer, 2020.

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

[11] Minhao Cheng, Simranjit Singh, Patrick Chen, Pin-Yu Chen, Sijia Liu, and Cho-Jui Hsieh. Sign-opt: A query-efficient hard-label adversarial attack. *arXiv preprint arXiv:1909.10773*, 2019.

[12] Minhao Cheng, Simranjit Singh, Patrick H. Chen, Pin-Yu Chen, Sijia Liu, and Cho-Jui Hsieh. Sign-opt: A query-efficient hard-label adversarial attack. In *International Conference on Learning Representations*, 2020.

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

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

[15] Yinpeng Dong, Qi-An Fu, Xiao Yang, Tianyu Pang, Hang Su, Zihao Xiao, and Jun Zhu. Benchmarking adversarial robustness on image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 321–331, 2020.

[16] John C Duchi, Michael I Jordan, Martin J Wainwright, and Andre Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015.

[17] Logan Engstrom, Andrew Ilyas, Hadi Salman, Shibani Santurkar, and Dimitris Tsipras. Robustness (python library), 2019.

[18] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *ICLR*, 2019.

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

[20] Sven Gowal, Chongli Qin, Jonathan Uesato, Timothy Mann, and Pushmeet Kohli. Uncovering the limits of adversarial training against norm-bounded adversarial examples. *arXiv preprint arXiv:2010.03593*, 2020.

[21] Chuan Guo, Jacob Gardner, Yurong You, Andrew Gordon Wilson, and Kilian Weinberger. Simple black-box adversarial attacks. In *International Conference on Machine Learning*, pages 2484–2493, 2019.

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

[23] Zhezhi He, Adnan Siraj Rakin, and Deliang Fan. Parametric noise injection: Trainable randomness to improve deep neural network robustness against adversarial attack. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 588–597, 2019.

[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*, pages 2137–2146, 2018.

[25] Andrew Ilyas, Logan Engstrom, and Aleksander Madry. Prior convictions: Black-box adversarial attacks with bandits and priors. In *International Conference on Learning Representations*, 2018.

[26] Alex Krizhevsky et al. Learning multiple layers of features from tiny images. 2009.

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

[28] Huiying Li, Shawn Shan, Emily Wenger, Jiayun Zhang, Haitao Zheng, and Ben Y Zhao. Blacklight: Defending black-box adversarial attacks on deep neural networks. *arXiv preprint arXiv:2006.14042*, 2020.

[29] Sijia Liu, Pin-Yu Chen, Xiangyi Chen, and Mingyi Hong. signsgd via zeroth-order oracle. In *International Conference on Learning Representations*, 2018.

[30] 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)*, pages 369–385, 2018.

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

[32] Seungyong Moon, Gaon An, and Hyun Oh Song. Parsimonious black-box adversarial attacks via efficient combinatorial optimization. In *International Conference on Machine Learning*, pages 4636–4645. PMLR, 2019.

[33] Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

[34] Ren Pang, Xinyang Zhang, Shouling Ji, Xiapu Luo, and Ting Wang. Advmind: Inferring adversary intent of black-box attacks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1899–1907, 2020.

[35] Tianyu Pang, Kun Xu, Chao Du, Ning Chen, and Jun Zhu. Improving adversarial robustness via promoting ensemble diversity. In *International Conference on Machine Learning*, pages 4970–4979. PMLR, 2019.

[36] Leslie Rice, Eric Wong, and Zico Kolter. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*, pages 8093–8104. PMLR, 2020.

[37] Evgenia Rusak, Lukas Schott, Roland S Zimmermann, Julian Bitterwolf, Oliver Bringmann, Matthias Bethge, and Wieland Brendel. A simple way to make neural networks robust against diverse image corruptions. In *European Conference on Computer Vision*, pages 53–69. Springer, 2020.

[38] Hadi Salman, Jerry Li, Ilya Razenshteyn, Pengchuan Zhang, Huan Zhang, Sebastien Bubeck, and Greg Yang. Provably robust deep learning via adversarially trained smoothed classifiers. In *Advances in Neural Information Processing Systems*, pages 11292–11303, 2019.

[39] Hadi Salman, Mingjie Sun, Greg Yang, Ashish Kapoor, and J Zico Kolter. Denoised smoothing: A provable defense for pretrained classifiers. *Advances in Neural Information Processing Systems*, 33, 2020.

[40] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[41] Leslie N Smith. Cyclical learning rates for training neural networks. In *2017 IEEE winter conference on applications of computer vision (WACV)*, pages 464–472. IEEE, 2017.

[42] Jure Sokolic, Raja Giryes, Guillermo Sapiro, and Miguel R. D. Rodrigues. Generalization error of invariant classifiers. In Aarti Singh and Xiaojin (Jerry) Zhu, editors, *AISTATS*, pages 1094–1103, 2017.

[43] David Stutz, Matthias Hein, and Bernt Schiele. Confidence-calibrated adversarial training: Generalizing to unseen attacks. In *International Conference on Machine Learning*, pages 9155–9166. PMLR, 2020.

[44] C Szegedy, V Vanhoucke, S Ioffe, J Shlens, and Z Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016.

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

[46] Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1631–1643. Curran Associates, Inc., 2020.

[47] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017.

[48] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. In *International Conference on Learning Representations*, 2018.

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

[50] Huanrui Yang, Jingyang Zhang, Hongliang Dong, Nathan Inkawhich, Andrew Gardner, Andrew Touchet, Wesley Wilkes, Heath Berry, and Hai Li. Dverge: Diversifying vulnerabilities for enhanced robust generation of ensembles. *Advances in Neural Information Processing Systems*, 33, 2020.

[51] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See final part of Section 1.
   (b) Did you describe the limitations of your work? [Yes] See Section 5.5.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Supplementary Materials
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 4.2.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Supplementary Materials.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 5.1 and Supplementary Materials
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] We run all experiments 3 times and average all results over 3 random seeds. We find the results of each experiment were almost identical to each other.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Supplementary Materials
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Section 5.1
   (b) Did you mention the license of the assets? [Yes] See Supplementary Materials
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] Yes, we provide our code in the Supplemental Materials.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No] No, all the used data are well-known public benchmark datasets such as CIFAR10, ImageNet.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] No, all the used data are well-known public benchmark datasets such as CIFAR10, ImageNet.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] Not applicable.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] Not applicable.
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] Not applicable.