On the Effectiveness of Small Input Noise for Defending Against Query-based Black-Box Attacks

Junyoung Byun, Hyojun Go, Changick Kim
Korea Advanced Institute of Science and Technology (KAIST)
{bjyoung, gohyojun15, changick}@kaist.ac.kr

Abstract

While deep neural networks show unprecedented performance in various tasks, the vulnerability to adversarial examples hinders their deployment in safety-critical systems. Many studies have shown that attacks are also possible even in a black-box setting where an adversary cannot access the target model's internal information. Most black-box attacks are based on queries, each of which obtains the target model's output for an input, and many recent studies focus on reducing the number of required queries. In this paper, we pay attention to an implicit assumption of query-based black-box adversarial attacks that the target model's output exactly corresponds to the query input. If some randomness is introduced into the model, it can break the assumption, and thus, query-based attacks may have tremendous difficulty in both gradient estimation and local search, which are the core of their attack process. From this motivation, we observe even a small additive input noise can neutralize most query-based attacks and name this simple yet effective approach Small Noise Defense (SND). We analyze how SND can defend against query-based black-box attacks and demonstrate its effectiveness against eight state-of-the-art attacks with CIFAR-10 and ImageNet datasets. Even with strong defense ability, SND almost maintains the original classification accuracy and computational speed. SND is readily applicable to pre-trained models by adding only one line of code at the inference.

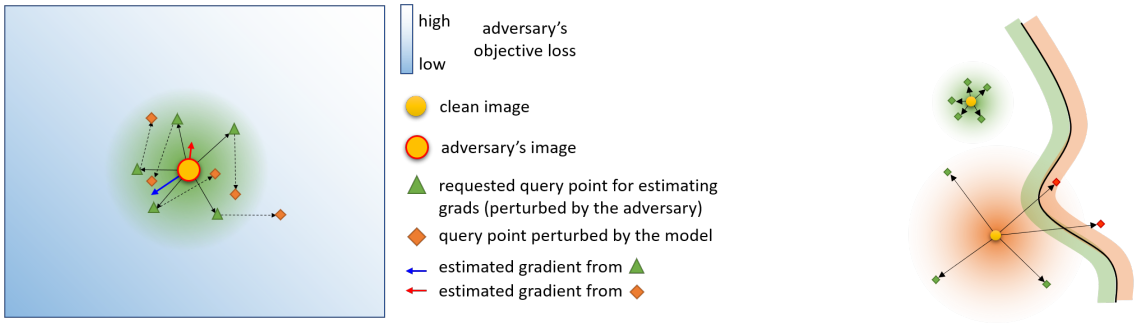
1. Introduction

Although deep neural networks perform well in various areas, it is now well-known that small and malicious input perturbation can cause them to malfunction [4, 35]. This vulnerability of AI models to adversarial examples hinders their deployment, especially in safety-critical areas. In the white-box setting, where the target model's parameters can be accessed, strong adversarial attacks such as Projected Gradient Descent (PGD) [27] can generate adversarial examples using the internal information. Recent studies have shown that adversarial examples can be generated even in a practical black-box setting where the model's interior is hidden to adversaries.

These black-box attacks can be largely divided into *transfer-based attacks* and *query-based attacks*. Transfer-based attacks train a substitute model that mimics the target model's behavior and take advantage of *transferability* that adversarial examples generated from a network can deceive other networks [30]. However, due to differences in training methods and model architectures, the transferability of adversarial examples can be significantly weakened, and thus, transfer-based attacks usually result in lower success rates [8]. For this reason, most black-box attacks are based on *queries*, each of which obtains the target model's output for an input. Query-based attacks create adversarial examples through an iterative process based on either local search with repetitive small input modifications or optimization with estimated gradients of an adversary's loss with respect to an input. Here, requesting many queries in their process takes a lot of time and financial loss. Moreover, many similar query images can be suspicious to system administrators. For this reason, researchers have focused on reducing the number of queries required to make a successful adversarial example [3].

Compared to the increasing number of studies on adversarial defenses in white-box settings, the number of defenses against query-based black-box attacks is still very small [3]. However, in a practical situation, black-box attacks are more realistic as attackers cannot know the target model's interiors. Since existing defenses developed for white-box attacks improve their robustness at the high cost of clean accuracy (accuracy on clean images) [37], it is necessary to develop a new defense strategy that targets query-based black-box attacks with minimal accuracy loss.

To defend against query-based black-box attacks, we pay attention to an implicit but important assumption of these attacks that *the target model's output exactly corresponds to the query input*. If some randomness is introduced into the model, it can break the assumption, and thus, query-based



(a) (b)

Figure 1: Illustrations of our intuitions. (a) Small noise can effectively disturb gradient estimation of query-based attacks which use finite difference. (b) Compared to large noise, small noise hardly affects predictions on clean images.

attacks may have tremendous difficulty in both gradient estimation and local search, which are the core of their attack process. This intuition is illustrated in Fig. 1(a).

Among previous studies, Dong *et al.* [14] empirically find that randomization-based defenses are more effective in defending against query-based black-box attacks than other types of defenses. However, existing randomization-based defenses introduce significant uncertainty into predictions, and thus, they also degrade clean accuracy.

In this paper, however, we highlight that simply adding small Gaussian noise into an input image is enough to defeat various query-based attacks by breaking the above core assumption while almost maintaining clean accuracy. One may think that additive Gaussian noise cannot defend against most adversarial attacks unless we introduce large randomness. This idea is valid for white-box attacks [15], but we will show that small noise is surprisingly effective against query-based black-box attacks.

Our second intuition is that sufficiently small Gaussian noise hardly affect predictions on clean images, as shown in Fig. 1(b). Dodge *et al.* [14] empirically find that classification accuracy decreases in proportion to the variance of Gaussian noise, but the accuracy drop is negligible for a sufficiently small variance.

We think an adversarial defense techniques should have the following goals: (1) preventing malfunction of a model against various attacks, (2) minimizing the computational overhead, (3) maintaining the accuracy on clean images, and (4) easily applicable to existing models. The proposed defense against query-based attacks meets all of the above objectives, and we name this simple yet effective defense technique *Small Noise Defense* (SND).

Our contributions can be listed as follows:

- We highlight the effectiveness of adding a small additive noise to input for defending against query-based black-box attacks. The proposed SND method can be readily applied to pre-trained models by adding only one line of code in the PyTorch framework [31] at the inference stage (`x = x + sigma * torch.randn_like(x)`) and almost maintains the performance of the model.
- We analyze how SND can efficiently interfere with gradient estimation and local search, which are the core of query-based attacks.
- We devise an adaptive attack method against SND and explain its limitations and the difficulty of evading SND.
- We have empirically shown that the proposed method can effectively defend against eight state-of-the-art query-based black-box attacks with the CIFAR-10 and ImageNet datasets. Specifically, four decision-based and four score-based attacks are used to show strong defense ability against various attacks, including local search-based and optimization-based methods.

2. Background

2.1. Adversarial Setting

Since we deal with adversarial attacks on the image classification task throughout the paper, we briefly explain adversarial attacks on the image classification task in this section.

Suppose that a neural network $f(\boldsymbol{x})$ classifies an image $\boldsymbol{x}$ among total N classes and returns a class-wise probability vector $\boldsymbol{y} = [y_1, ..., y_N]$ for $\boldsymbol{x}$. For notational convenience, we also denote the probability of i^{th} class (i.e., y_i) as $f(\boldsymbol{x})_i$ and the top-1 class index as $h(\boldsymbol{x}) = \arg\max_{i \in C} y_i$, where $C = \{1, ..., N\}$.

In a black-box threat model, an adversary has a clean image $\boldsymbol{x}_0$ whose class index is c_0 and wants to generate an adversarial example $\hat{\boldsymbol{x}} = \boldsymbol{x}_0 + \boldsymbol{\delta}$ to fool a target model f. In the following, we denote the adversarial example at t^{th} step in an iterative attack algorithm as $\hat{\boldsymbol{x}}_t = \boldsymbol{x}_0 + \boldsymbol{\delta}_t$. The adversary should generate an adversarial example within a

perturbation norm budget ϵ and query budget Q. If we let q be the number of queries used to make $\boldsymbol{\delta}_t$, then we can write the adversary's objective as follows:

$$\min_{\boldsymbol{\delta}_t} \ell(\boldsymbol{x}_0 + \boldsymbol{\delta}_t), \text{ subject to } ||\boldsymbol{\delta}_t||_p \leq \epsilon \text{ and } q \leq Q, \quad (1)$$

where $\ell(\boldsymbol{x}) = f(\boldsymbol{x})_{c_0} - \max_{c \neq c_0} f(\boldsymbol{x})_c$ for untargeted attacks and $\ell(\boldsymbol{x}) = \max_{c \neq \hat{c}} f(\boldsymbol{x})_c - f(\boldsymbol{x})_{\hat{c}}$ for targeted attacks with target class index $\hat{c}$. Since decision-based attacks cannot obtain $\ell(\boldsymbol{x})$, they have a different objective for untargeted attacks as follows:

$$\min_{\boldsymbol{\delta}_t} ||\boldsymbol{\delta}_t||_p, \text{ subject to } h(\boldsymbol{x}_0) \neq h(\boldsymbol{x}_0 + \boldsymbol{\delta}_t) \text{ and } q \leq Q. \quad (2)$$

Unless otherwise noted, in this paper, we use $p = 2$ and focus on untargeted attacks because it is more challenging for defenders. Besides, we assume that each pixel value is normalized into $[0, 1]$.

2.2. Taxonomy of query-based black-box attacks

Query-based attacks can be largely divided into score-based and decision-based attacks according to the available type of output of the target model (class-wise probabilities for score-based attacks and the top-1 class index for decision-based attacks). On the other hand, query-based attacks can be categorized into optimization-based attacks and local search-based attacks. Optimization-based methods optimize an adversary's objective loss with estimated gradients of the loss with respect to $\hat{\boldsymbol{x}}_t$. In contrast, local search-based attacks repeatedly update an image according to how the model's output changes after adding a small perturbation.

In the following, we briefly introduce various query-based attacks used in this paper.

Bandit optimization with priors (Bandit-TD). Ilyas *et al.* [21] observe that the image gradients in successive steps of an iterative attack have strong correlations. In addition, they find that the gradients of surrounding pixels also have correlations. Bandit-TD exploits this information as priors for efficient gradient estimation.

Simple Black-box Attack (SimBA & SimBA-DCT). For each iteration, SimBA [17] samples a vector $\boldsymbol{q}$ from a pre-defined set Q and modify the current image $\hat{\boldsymbol{x}}_t$ with $\hat{\boldsymbol{x}}_t - \boldsymbol{q}$ and $\hat{\boldsymbol{x}}_t + \boldsymbol{q}$ and updates the image in the direction of decreasing $\boldsymbol{y}_{c_0}$. Inspired by the observation that low-frequency components make a major contribution to misclassification [16], SimBA-DCT exploits DCT basis in low-frequency components for query-efficiency.

Boundary Attack (BA). BA [5] updates $\hat{\boldsymbol{x}}_t$ on the decision-boundary so that the perturbation norm gradually decreases via random walks while misclassification is maintained.

Sign-OPT. Cheng *et al.* [10] treat a decision-based attack as a continuous optimization problem of the nearest distance to the decision boundary. They use the randomized gradient-free method [29] for estimating the gradient of the distance. Cheng *et al.* [11] propose Sign-OPT, which uses the expectation of the sign of gradient with random directions to estimate the gradients efficiently without exhaustive binary searches.

Hop Skip Jump Attack (HSJA). Chen *et al.* [7] improve BA with gradient estimation. For each iteration of HSJA, it finds an image on the boundary with a binary search algorithm, and estimates the gradients, and calculates the step-size towards the decision boundary.

GeoDA. Rahmati *et al.* [34] propose a geometry-based attack that exploits a geometric prior that the decision boundary of the neural network has a small curvature on average near data samples. By linearizing the decision boundary in the vicinity of samples, it can efficiently estimate the normal vector of the boundary, which helps to reduce the number of required queries for generating adversarial examples.

2.3. Adversarial Defenses

As Dong *et al.* [14] observe that randomization is important for effective defense against query-based attacks, we focus on randomization-based defenses among various defense methods. In what follows, we briefly explain three randomization-based defenses along with PGD-adversarial training.

Random Self-Ensemble (RSE). RSE [26] adds Gaussian noise with $\sigma_{\text{inner}} = 0.1$ to the input of each convolutional layer, except for the first convolutional layer where $\sigma_{\text{init}} = 0.2$ is used. To stabilize the performance, they use an ensemble of multiple predictions for each image.

Parametric Noise Injection (PNI). He *et al.* [20] propose a method to increase the robustness of neural networks by adding trainable Gaussian noise to the activation or weight of each layer. They introduce learnable scale factors of noise and allow them to be learned with adversarial training.

Random Resizing and Padding (R&P). Xie *et al.* [39] propose a random input transform-based method. In front of network inference, it applies random resizing and random padding to its input sequentially, making adversaries obtain noisy gradients. It can be easily applied to a pre-trained model, but it increases total computational time due to the enlarged input image.

PGD-Adversarial Training (PGD-AT). Madry *et al.* [27] propose PGD-adversarial training which trains a model with adversarial examples generated by Projected Gradient Descent (PGD) [27]. Unlike other defenses that are ineffective against adaptive attacks, it is well known that PGD-AT provides strong defense against a variety of white-box attacks.

3. Analysis

3.1. Our Approach

To defend against query-based black-box attacks, we add Gaussian noise with a sufficiently small σ to the input as follows.

$$f_{\boldsymbol{\eta}}(\boldsymbol{x}) = f(\boldsymbol{x}+\boldsymbol{\eta}), \text{ where } \boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}) \text{ and } \sigma \ll 1. \tag{3}$$

For an adversary, since the exact value of $\boldsymbol{\eta}$ is unknown, there are multiple possible output values for any $\boldsymbol{x}$, so $f_{\boldsymbol{\eta}}$ is a random process. In what follows, we will explain how this transform introduces tremendous difficulty in both gradient estimation and local search in query-based black-box attacks.

3.2. Defense Against Optimization-based Attacks

In this subsection, we will explain how small Gaussian input noise can disturb the gradient estimation in optimization-based attacks. We first look at defense against score-based attacks and then deal with decision-based attacks.

The core of optimization-based attacks is an accurate estimation of $\nabla \ell(\boldsymbol{x})$, which needs to be approximated with finite difference because of the black-box setting. For instance, the gradient can be estimated as $\tilde{\boldsymbol{g}}$ by Random Gradient-Free method [29] as follows.

$$\tilde{\boldsymbol{g}} = \frac{1}{B}\sum_{i=0}^{B} \boldsymbol{g}_i,$$
$$\text{where } \boldsymbol{g}_i = \frac{\ell(\hat{\boldsymbol{x}}_t + \beta \boldsymbol{u}) - \ell(\hat{\boldsymbol{x}}_t)}{\beta}\boldsymbol{u} \text{ and } \boldsymbol{u} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}). \tag{4}$$

Conceptually, by introducing small Gaussian noise into input, $\tilde{\boldsymbol{g}}$ can greatly differ from the true gradient $\nabla \ell$ as shown in Fig. 1.

To illustrate it more formally, let us represent $\boldsymbol{\eta}$ by replacing it with $\boldsymbol{\eta}(\boldsymbol{x})$ to clarify $\boldsymbol{\eta}$ depends on both time and $\boldsymbol{x}$. Suppose $f^*_{\boldsymbol{\eta}(\boldsymbol{x})}$ is a sample function of the random process $f_{\boldsymbol{\eta}(\boldsymbol{x})}(\boldsymbol{x})$ at some time. Then, this function is noisy with regard to $\boldsymbol{x}$ because of $\boldsymbol{\eta}(\boldsymbol{x})$. We also assume that ℓ^* is derived from $f^*_{\boldsymbol{\eta}(\boldsymbol{x})}$, then unless $\mathrm{Var}[\ell^*(\boldsymbol{x}+\boldsymbol{u})]$ is extremely small, ℓ^* is discontinuous and non-differentiable, and thus, $\nabla \ell^*$ does not exist. Therefore, the estimated gradient using finite differences does not converge to the target gradient $\nabla \ell$. For example, simplifying the problem, if f is a one-dimensional function $\mathbb{R} \to \mathbb{R}$ and sampled $\eta(1.000) = 0.08$ and $\eta(1.001) = 0.03$ then, the sampled function values $f(x+\eta(x))$ at $x = 1.000$ and $x = 1.001$ become $f(1.080)$ and $f(0.9701)$, respectively. Therefore, if the variance of f is large, the sample function becomes more noisy.

In decision-based attacks, $\hat{\boldsymbol{x}}_t$ is likely to be in the vicinity of the decision boundary. Therefore, even small noise can move $\hat{\boldsymbol{x}}_t$ across the boundary so that the output is changed. The estimated gradient through erroneous predictions hinders the generation of adversarial examples. We illustrate the working principle of SND against decision-based attacks in supplementary material. Besides, the binary search algorithm, which is widely used to calculate the distance to the decision boundary [7, 11], can make a larger error due to $\boldsymbol{\eta}$. Therefore, algorithms such as HSJA, which assume that $\hat{\boldsymbol{x}}$ is near the decision boundary, are likely to work incorrectly.

3.3. Defense Against Local Search-based Attacks

Local search-based attacks try to update the image in the direction that decreases the adversarial objective loss. However, since the output is unreliable due to noise, this becomes similar to random motion. Suppose an adversary recognizes that the attack objective loss decreases for $\hat{\boldsymbol{x}}_t + \boldsymbol{q}$, where $\boldsymbol{q}$ is a perturbation, and updates $\hat{\boldsymbol{x}}_{t+1}$ as $\hat{\boldsymbol{x}}_t + \boldsymbol{q}$. However, since the actually evaluated input of f is $\hat{\boldsymbol{x}}_t + \boldsymbol{q} + \boldsymbol{\eta}$, where $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$, the attack objective loss might increase at the originally intended input $\hat{\boldsymbol{x}}_t + \boldsymbol{q}$. This prediction error makes the attack algorithm stuck in the iterative process and prevents generating adversarial examples.

3.4. Adaptive Attacks on Small Noise Defense

Solid research for adversarial defense requires evaluating the defense ability against adaptive attacks that exactly know the working principle of the defense [6]. Athalye *et al.* [1] show that randomization-based defense such as R&P can be circumvented through the Expectation Over Transform (EOT) technique [2]. The EOT technique approximates the gradient at each gradient descent step by averaging gradients w.r.t. several samples transformed by randomization-based defenses. In the black-box setting, the gradients w.r.t. an input cannot be obtained at once, and it should be estimated using the finite-difference by giving several queries. Therefore, EOT-based adaptive attacks in black-box settings average outputs for each input to get a reliable prediction for accurate gradient estimation.

The above way of averaging the noisy output (i.e., expectation) is one of two primary techniques for handling noise in derivative-free optimization [38]. The other primary technique is threshold selection which stores a reliable candidate solution set (in our framework, candidate adversarial examples) that makes minimal losses. It accepts a new solution for the candidate set when the evaluated loss is less than the recorded smallest loss by the threshold. This principle of threshold selection can be adapted to the gradient estimation step in query-based attacks. Since decision-based attacks try to reduce the size of adversarial perturbation, selecting a candidate set with a sufficiently large

ResNet-20 on CIFAR-10		ResNet-50 on ImageNet	
Defense	Clean Accuracy (%)	Defense	Clean Accuracy (%)
Baseline	91.34	Baseline	76.13
SND ($\sigma = 0.001$)	91.33 ± 0.02	SND ($\sigma = 0.001$)	76.10 ± 0.02
SND ($\sigma = 0.01$)	90.57 ± 0.09	SND ($\sigma = 0.01$)	75.47 ± 0.03
SND ($\sigma = 0.02$)	87.56 ± 0.18	SND ($\sigma = 0.02$)	73.91 ± 0.02
RSE	83.40 ± 0.15	PGD-AT	57.9
PNI	85.15 ± 0.18	R&P	74.26 ± 0.07

Table 1: Comparison of clean accuracy. For randomization-based methods, we denote the mean and standard deviation of clean accuracy in 5 repetitive experiments with different random seeds.

threshold is identical to increasing σ of random perturbations in the gradient estimation like Eq. 4 to ignore the disturbance of small input noise. However, increasing the perturbation size can amplify the gradient estimation error by itself.

Therefore, following the suggestion of [1, 38], we design expectation-based adaptive attacks against SND. In the following, we shed light on the difficulty of evading the proposed defense with the adaptive attacks in detail.

In our framework, the input of the function, $\boldsymbol{x} + \boldsymbol{\eta}$, is a Gaussian random process. But the result of the nonlinear function f, $f(\boldsymbol{x} + \boldsymbol{\eta})$, is no longer a Gaussian random process. This makes it very difficult for query-based attacks to bypass SND. For expectation-based adaptive attacks, adversaries can approximate $f(\boldsymbol{x})$ as $\mathbb{E}_{\boldsymbol{\eta}}[f_{\boldsymbol{\eta}}(\boldsymbol{x})]$ by taking the average over multiple queries using the fact that $\mathbb{E}(\boldsymbol{\eta}) = \mathbf{0}$. However, this attempt requires many queries for each iteration and greatly diminishes query efficiency. We note that adversaries should also consider the query efficiency for their adaptive attacks.

In addition, even if a large amount of queries are used, $\mathbb{E}_{\boldsymbol{\eta}}[f_{\boldsymbol{\eta}}(\boldsymbol{x})]$ may be different from $f(\boldsymbol{x})$ because of the nonlinearity of the deep neural networks. With simple examples, we will explain how the expectation value differs from the actual value when Gaussian noise is added to the input of nonlinear functions.

Example 1 (A simple nonlinear function)**.** *Let $F(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{x}$, where $F : \mathbb{R}^d \to \mathbb{R}$, and $F_{\boldsymbol{\eta}}(\boldsymbol{x}) = F(\boldsymbol{x} + \boldsymbol{\eta})$ where $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{I})$. Suppose we estimate $F(\mathbf{0})$ with $\mathbb{E}[F_{\boldsymbol{\eta}}(\mathbf{0})]$. Then, $\mathbb{E}[F_{\boldsymbol{\eta}}(\mathbf{0})] = \mathbb{E}[(\mathbf{0} + \boldsymbol{\eta})^T(\mathbf{0} + \boldsymbol{\eta})] = \mathbb{E}[\boldsymbol{\eta}^T\boldsymbol{\eta}] = d\sigma^2$. Therefore, $\mathbb{E}[F_{\boldsymbol{\eta}}(\mathbf{0})] = d\sigma^2 \neq 0 = F(\mathbf{0})$ and if d is very large (e.g., for an image of size $224\times224\times3$, d=$150,528$), then the estimation error would be high.*

Example 2 (A simple ReLU network case)**.** *Let $ReLU(x) = max(0, x)$ and $F(x) = ReLU(wx + b)$, where $F : \mathbb{R} \to \mathbb{R}$. Let $F_\eta(x) = F(x+\eta)$, where $\eta \sim \mathcal{N}(0, \sigma^2)$ and suppose we estimate $F(x)$ with $\mathbb{E}[F_\eta(x)]$. Then $\mathbb{E}[F_\eta(x)]$ is as follows:*

$$(wx+b)(1-\Phi\left(-\frac{wx+b}{|w|\sigma}\right))+|w|\sigma\phi\left(-\frac{wx+b}{|w|\sigma}\right). \quad (5)$$

Proof. Let $w(x + \eta) + b$ be Y, then Y can be represented with $\mu_y = wx + b$ and $\sigma_y^2 = w^2\sigma^2$ as:

$$Y \sim \mathcal{N}(\mu_y, \sigma_y^2). \quad (6)$$

Then, $F_\eta(x)$ is $\max(0, Y)$ and $\mathbb{E}[F_\eta(x)] = \mathbb{E}[\max(0, Y)]$ can be obtained by the law of total expectation.

$$\begin{aligned}\mathbb{E}[F_\eta(x)] &= \mathbb{E}[\max(0, Y)] \\ &= \mathbb{E}[Y|Y > 0]\Pr(Y > 0) + 0\Pr(Y \leq 0).\end{aligned} \quad (7)$$

Using the truncated normal distribution, we recall the fact as follows:

$$\mathbb{E}[Y|Y > a] = \mu_y + \sigma_y\frac{\phi((a - \mu_y)/\sigma_y)}{1 - \Phi((a - \mu_y)/\sigma_y)}, \quad (8)$$

where $\phi(x) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}x^2)$ and $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. Since $\Pr(Y > 0)) = 1 - \Phi(\frac{\mu_y}{\sigma_y})$, $\mathbb{E}[F_\eta(x)]$ is represented as:

$$\begin{aligned}&\mu_y(1 - \Phi\left(-\frac{\mu_y}{\sigma_y}\right)) + \sigma_y\phi\left(-\frac{\mu_y}{\sigma_y}\right) \\ &= (wx + b)(1 - \Phi\left(-\frac{wx+b}{|w|\sigma}\right)) + |w|\sigma\phi\left(-\frac{wx+b}{|w|\sigma}\right).\end{aligned} \quad (9)$$

Therefore, if the noise is added to the input in the simple ReLU case, there can be a difference between the actual $F(x)$ value and the estimated value by $\mathbb{E}[F_\eta(x)]$. □

From the proof on the simple network, we can expect that the average of the output may have an error with the actual output even in a deep neural network.

Attack method	BA			Sign-OPT			HSJA			GeoDA		
# of queries	2K	5K	10K	2K	5K	10K	2K	5K	10K	2K	5K	10K
Baseline	36.2	69.5	84.6	59.1	88.9	91.2	86.4	89.2	89.2	64.7	71.3	76.5
SND ($\sigma = 0.01$)	14.0	17.8	20.0	21.7	22.3	22.8	16.5	19.9	22.7	12.0	12.1	12.4
SND ($\sigma = 0.001$)	33.0	53.0	61.3	20.6	22.2	23.4	48.1	67.6	81.9	11.7	11.8	12.2
RSE	18.2	19.0	19.8	18.7	18.7	18.7	19.9	22.2	23.4	19.7	19.9	20.5
PNI	15.1	15.4	15.8	18.1	18.1	18.1	17.4	19.2	20.6	19.9	20.2	20.4

Table 2: Evaluation of attack success rates (%) on the CIFAR-10 dataset.

Attack type	Decision-based Attack								
Attack method	Sign-OPT			HSJA			GeoDA		
# of queries	5K	10K	20K	5K	10K	20K	5K	10K	20K
Baseline	36.4%	62.4%	88.0%	64.0%	88.4%	99.6%	50.0%	62.8%	72.0%
	(9.71)	(4.72)	(2.38)	(4.43)	(2.34)	(1.28)	(6.38)	(5.04)	(4.12)
SND (σ=0.01)	**6.8%**	**6.8%**	**7.2%**	**6.4%**	**7.6%**	**8.4%**	**7.2%**	**7.6%**	**7.6%**
	(41.37)	(70.27)	(87.94)	(31.03)	(26.81)	(22.93)	(33.48)	(33.14)	(32.61)
SND (σ=0.001)	**6.8%**	7.6%	8.0%	13.6%	20.4%	32.4%	8.4%	8.8%	9.2%
	(35.15)	(54.33)	(88.29)	(11.60)	(8.86)	(6.75)	(27.44)	(25.88)	(24.41)
PGD-AT	28.8%	30.0%	32.4%	30.4%	33.2%	36.0%	32.4%	34.0%	35.6%
	(30.67)	(24.66)	(19.72)	(20.99)	(17.27)	(13.46)	(14.54)	(13.63)	(12.90)
R&P	13.2%	13.2%	13.2%	13.6%	15.2%	16.0%	14.4%	14.4%	15.2%
	(51.19)	(82.14)	(85.78)	(33.01)	(31.00)	(29.42)	(31.72)	(31.21)	(30.45)
Attack type	Score-based Attack								
Attack method	SimBA			SimBA-DCT			Bandit-TD		
# of queries	5K	10K	20K	5K	10K	20K	5K	10K	20K
Baseline	74.0%	74.4%	74.4%	94.8%	95.2%	95.2%	94.0%	97.2%	98.4%
	(3.89)	(3.99)	(4.02)	(3.12)	(3.14)	(3.14)	(4.70)	(4.70)	(4.70)
SND (σ=0.01)	**8.4%**	**9.2%**	**10.0%**	**8.4%**	**8.8%**	**10.4%**	15.2%	15.2%	16.4%
	(0.52)	(0.55)	(0.57)	(0.56)	(0.58)	(0.60)	(4.74)	(4.74)	(4.74)
SND (σ=0.001)	27.2%	35.6%	50.4%	46.4%	60.4%	68.4%	**7.2%**	**7.6%**	**8.4%**
	(1.84)	(2.14)	(2.43)	(2.22)	(2.44)	(2.58)	(4.83)	(4.83)	(4.83)
PGD-AT	27.6%	27.6%	27.6%	36.0%	36.0%	36.0%	38.8%	45.2%	52.8%
	(5.46)	(7.55)	(10.17)	(5.36)	(6.21)	(6.62)	(3.54)	(3.54)	(3.54)
R&P	26.4%	27.6%	28.0%	27.2%	28.4%	29.2%	32.0%	33.2%	33.6%
	(0.48)	(0.51)	(0.54)	(0.52)	(0.55)	(0.58)	(4.50)	(4.50)	(4.50)

Table 3: Evaluation of attack success rates against defenses on the ImageNet dataset. We denote the average ℓ_2 norm of perturbations in the parenthesis.

4. Experiments and Discussion

4.1. Experimental Settings

In this section, we evaluated the defense ability of SND against eight query-based black-box attacks: BA, Sign-OPT, HSJA, GeoDA, SimBA, SimBA-DCT, Bandit-TD, and Subspace Attack, along with other defense methods: PNI, RSE, R&P, and PGD-AT. We used the CIFAR-10 [23] and ImageNet [13] datasets for our experiments and following previous studies [7, 17, 20], we used ResNet-20 for CIFAR-10 and ResNet-50 [19] for ImageNet as target networks. Following [5], we randomly sampled 1,000 and 250 correctly classified images from the CIFAR-10 test set and the ImageNet validation set for evaluation. We describe detailed experimental settings in supplementary material.

For evaluation metrics, we first define a *successfully attacked image* as an image from which an attack can find an adversarial image within the perturbation budget ϵ and query budget Q. With this definition, we use *attack success rate*, which is the percentage of the number of successfully attacked images over the total number of evaluated images. We note that since we evaluate defense performance, **a lower attack success rate is better**. We measured the ℓ_2 norm of perturbations and set ϵ to 1.0 for the CIFAR-10

dataset, and to 5.0 for the ImageNet dataset. Note that we denote the q^{th} query image as $\hat{\boldsymbol{x}}^q$. $\hat{\boldsymbol{x}}^q$ and $\hat{\boldsymbol{x}}_t$ can be different.

4.2. Evaluation of Clean Accuracy

We first evaluated the clean accuracy of models with defenses on the original test split (10K images) of the CIFAR-10 and validation split (50K images) of the ImageNet dataset. As shown in Table 1, SND hardly reduces the clean accuracy compared to other methods. The accuracy drop caused by SND is not significant at $\sigma \leq 0.01$, which implies that sufficiently small σ hardly affects clean accuracy.

4.3. Evaluation on the CIFAR-10 Dataset

We performed four decision-based attacks against models with defenses, and Table 2 shows the evaluated attack success rates. SND shows competitive defense ability despite having more than 5% higher clean accuracy compared to other defenses. Moreover, due to significant performance drop, RSE and PNI cannot be applied to the models for large-scale image classification with the ImageNet dataset.

4.4. Evaluation on the ImageNet Dataset

We performed six query-based attacks against models with defenses, and Table 3 shows the evaluated attack success rates. When the query budget Q is 20K, the average of the attack success rates over the attacks against the baseline is 87.9%, whereas SND with $\sigma = 0.01$ significantly reduces it to 10.0%. SND with $\sigma = 0.001$ also significantly reduces the average attack success rate to 29.5%, which is comparable to the second-best method, R&P (22.5%).

We also calculated the average ℓ_2 norm of perturbations of query images $||\boldsymbol{x}_0 - \hat{\boldsymbol{x}}^q||_2$ at the predefined query budget Q to show whether the perturbation norm diverges or not. If an attack stops in the middle without requesting Q queries, we used the last query image instead. In decision-based attacks, it can be seen that randomization-based defenses, SND and R&P, significantly increase the perturbation norm as q increases. In SimBA and SimBA-DCT, the perturbation norm is minimal in SND and R&P, which implies that the attacks have significant difficulty in finding a perturbation which decreases y_{c_0}.

4.5. Empirical Evidence for Assumptions of SND

To provide supporting evidence for assumptions of SND in score-based attacks, we calculated $\hat{\sigma} = \sqrt{\mathrm{Var}[\ell(\boldsymbol{x}+\boldsymbol{\eta})]}$, where $\boldsymbol{\eta} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. With $\sigma = 0.01$ and 1,000 test images of the CIFAR-10, we evaluated $\ell(\boldsymbol{x}+\boldsymbol{\eta})$ for 100 iterations for each clean image and calculated the $\hat{\sigma}$ averaged over all images. In our experiment, $\hat{\sigma} = 0.04$ which is small but sufficient to make $\ell(\boldsymbol{x}+\boldsymbol{\eta})$ to be non-differentiable about $\boldsymbol{x}$.

Defense Attack	SND ($\sigma = 0.001$)	SND ($\sigma = 0.01$)
BA	0.134	0.227
Sign-OPT	0.216	0.215
HSJA	0.255	0.189
GeoDA	0.314	0.391
None	0.002	0.021

Table 4: Evaluation of $P(h(\boldsymbol{x}) \neq h(\boldsymbol{x}+\boldsymbol{\eta}))$.

In decision-based attacks, our assumption in Section 3.2 is that if $\hat{\boldsymbol{x}}_t$ is near the decision boundary, small noise can easily move the image across the boundary. We evaluated $P_{mis} := P(h(\boldsymbol{x}) \neq h(\boldsymbol{x}+\boldsymbol{\eta}))$ through experiments. We counted the above mismatch case for all queries during the attack process. With $\sigma = 0.001, 0.01$ and the CIFAR-10 test images, the average of P_{mis} over all the attacks is calculated as 0.22 and 0.25, respectively. In contrast, on clean images, P_{mis} is obtained as 0.002 and 0.021, respectively. Therefore, the results shown in Table 4 support our argument.

4.6. Evaluation of Adaptive Attacks Against SND

As described in Section 3.4, we devised an adaptive attack against SND that takes the expectation of predictions for repetitive T queries. In this experiment, we performed HSJA against SND with $\sigma = 0.01$ on the CIFAR-10 dataset. Since HSJA is a decision-based attack, we regard the most predicted class in T queries as the expected class. We measured the attack success rate and P_{mis} according to the query budget, and the adaptive attack clearly shows a higher attack success rate than the baseline ($T = 1$), as shown in Table 5. On the same query budget, however, the adaptive attack shows a lower attack success rate (e.g., 22.7% (T=1) > 18.2% (T=5) at Q=10K, and 29.3% (T=5) > 27.3% (T=10) at Q=50K). Therefore, the expectation-based adaptive attack has limitations due to the restricted query budget. Moreover, even if T increases, P_{mis} does not decrease and this reinforces our argument in Section 3.4. We also applied the adaptive attack to BA, SO, and GeoDA on CIFAR-10 and two score-based attacks (SimBA-DCT and Bandit-TD) on ImageNet for comprehensive comparisons. The experimental results are shown in supplementary material.

4.7. Varying σ for Each Inference

So far, we have used a fixed σ for SND. Changing σ for each query may reduce clean accuracy while maintaining the defense ability. From this motivation, we multiplied $\boldsymbol{\eta}$ with k which is randomly sampled between 0 and 1 from the beta distributions with three different settings: (1) Uniformly random (the same as α=β=1) (2) Sampling from a beta distribution with α=β=2 whose probability density

# of queries	2K×T	5K×T	10K×T	P_{mis}
HSJA (T=1)	16.5%	19.9%	22.7%	0.189
HSJA (T=5)	18.2%	23.8%	29.3%	0.321
HSJA (T=10)	21.0%	27.3%	34.9%	0.376
HSJA (T=20)	25.2%	34.0%	46.6%	0.410
# of queries	50K	100K	200K	
HSJA	29.0%	30.1%	31.0%	0.120

Table 5: Attack success rates of the adaptive version of HSJA against SND ($\sigma = 0.01$) with different T.

function (PDF) is ∩-shaped. (3) Sampling from a beta distribution with α=β=0.5 whose PDF is ∪-shaped. We calculated clean accuracy and average ℓ_2 norm of noise for each method. Among the three ways, at $\sigma = 0.01$, α=β=2 is better than the others in terms of the loss of clean accuracy and the defensive ability against Sign-OPT. Detailed results can be found in supplementary material.

4.8. Defense Against Hybrid Black-box Attacks

Since SND protects models by interfering with gradient estimation and the local search of query-based attacks, we do not expect SND is effective against transfer-based attacks as these attacks exploit the transferability of adversarial examples. However, our method is complementary with other defenses, which are mainly effective against transfer-based black-box attacks such as [22] and [36]. Combined with other defense techniques, SND can work well against general black-box attacks, provided that the model's parameters are kept secret to adversaries.

To support this argument, we experimented with Subspace attack [18], a hybrid attack that exploits transferability in query-based attacks. Specifically, the Subspace attack exploits transferability-based priors, which are gradients from local substitute models trained on a small proxy dataset. We used pre-trained ResNet-18 and ResNet-34 [19] as reference models for gradient priors. We performed the attack based on the ℓ_∞ norm because the authors provide parameter settings only for the ℓ_∞ norm.

Detailed experimental results are described in supplementary material, but the results show that SND alone cannot effectively defend against the hybrid attack with gradient priors. However, when SND is combined with PGD-AT, it effectively protects the model and decreases the attack success rate from 100% to 42.4% at Q=20K and σ=0.01. To focus on the defensive ability against gradient estimation, we recalculated the attack success rate without initially misclassified images. Then, the newly obtained attack success rate decreases from 100% to 16.4% at Q=20K. This result implies that SND can be combined with other defenses against transfer-based attacks to achieve strong defense ability against all types of black-box attacks.

5. Related Work

The idea of injecting random noise at the inference stage for improving adversarial robustness is not new [32, 33, 28]. However, we note that small input noise, which is insufficient to prevent white-box attacks, is surprisingly effective in defending models against query-based black-box attacks.

5.1. History-based Detection Methods Against Query-based Black-box Attacks.

To the best of our knowledge, studies that mainly target defending against query-based black-box attacks have not yet been published. However, history-based detection techniques for query-based attacks have been proposed recently [9, 25]. Considering that adversary requires many queries of similar images for finding an adversarial example, they store information about past query images to detect the unusual behavior of query-based attacks.

5.2. Certified Defense With Additive Gaussian Noise.

Li *et al.* [24] analyze the connection between the robustness of models against additive Gaussian noise and adversarial perturbations. They derive the certified bounds on the norm bounded adversarial perturbation, and they propose a new training strategy to improve the certified robustness. Similarly, randomized smoothing [12] creates a smoothed classifier that correctly classifies when Gaussian noise is added to the classifier's input. Cohen *et al.* [12] prove that this smoothed classifier can have ℓ_2 certified robustness for an input. Both SND and the above certified defenses add Gaussian noise to the input. However, the purpose of the addition of noise in the certified defenses is to induce the classifier to gain certified robustness. Whereas SND adds noise to disturb an accurate measurement of the output to defend against query-based black-box attacks at the inference. In addition, the certified defenses use a much larger σ (≥ 0.25) than SND (0.01).

6. Conclusion

In this paper, we highlight that even small Gaussian input noise can effectively neutralize query-based black-box attacks and name this approach Small Noise Defense (SND). Our work suggests that query-based black-box attacks should consider the randomness of the target network as well. We demonstrate its effectiveness against eight query-based attacks with CIFAR-10 and ImageNet datasets. Interestingly, SND is very simple and easy for defenders but difficult for attackers to bypass. SND is readily applicable to pre-trained models by adding only one code line. Due to its simplicity and effectiveness, we hope that SND will be used as a baseline of defense against query-based black-box attacks in the future.

References

[1] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning*, pages 274–283. PMLR, 2018.

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

[3] Siddhant Bhambri, Sumanyu Muku, Avinash Tulasi, and Arun Balaji Buduru. A survey of black-box adversarial attacks on computer vision models. *arXiv*, pages arXiv–1912, 2019.

[4] Battista Biggio, Igino Corona, Davide Maiorca, B Nelson, N Srndic, P Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, ECML PKDD 2013*, volume 8190, pages 387–402. Springer, 2013.

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

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

[9] Steven Chen, Nicholas Carlini, and David Wagner. Stateful detection of black-box adversarial attacks. *arXiv preprint arXiv:1907.05587*, 2019.

[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*, 2018.

[11] 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*, 2019.

[12] Jeremy M Cohen, Elan Rosenfeld, and J Zico Kolter. Certified adversarial robustness via randomized smoothing. *arXiv preprint arXiv:1902.02918*, 2019.

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[14] 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.

[15] Shixiang Gu and Luca Rigazio. Towards deep neural network architectures robust to adversarial examples. *arXiv preprint arXiv:1412.5068*, 2014.

[16] Chuan Guo, Jared S Frank, and Kilian Q Weinberger. Low frequency adversarial perturbation. *arXiv preprint arXiv:1809.08758*, 2018.

[17] 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.

[18] Yiwen Guo, Ziang Yan, and Changshui Zhang. Subspace attack: Exploiting promising subspaces for query-efficient black-box attacks. In *Advances in Neural Information Processing Systems*, pages 3825–3834, 2019.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016.

[20] 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.

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

[22] Mohammad AAK Jalwana, Naveed Akhtar, Mohammed Bennamoun, and Ajmal Mian. Orthogonal deep models as defense against black-box attacks. *IEEE Access*, 8:119744–119757, 2020.

[23] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[24] Bai Li, Changyou Chen, Wenlin Wang, and Lawrence Carin. Certified adversarial robustness with additive noise. In *Advances in Neural Information Processing Systems*, pages 9464–9474, 2019.

[25] 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.

[26] 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.

[27] 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.

[28] Yaniv Nemcovsky, Evgenii Zheltonozhskii, Chaim Baskin, Brian Chmiel, Maxim Fishman, Alex M Bronstein, and Avi Mendelson. Smoothed inference for adversarially-trained models. *arXiv preprint arXiv:1911.07198*, 2019.

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

[30] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pages 506–519, 2017.

[31] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[32] Rafael Pinot, Laurent Meunier, Alexandre Araujo, Hisashi Kashima, Florian Yger, Cedric Gouy-Pailler, and Jamal Atif. Theoretical evidence for adversarial robustness through randomization. *Advances in Neural Information Processing Systems*, 32:11838–11848, 2019.

[33] Rafael Pinot, Laurent Meunier, Florian Yger, Cédric Gouy-Pailler, Yann Chevaleyre, and Jamal Atif. On the robustness of randomized classifiers to adversarial examples. *arXiv preprint arXiv:2102.10875*, 2021.

[34] Ali Rahmati, Seyed-Mohsen Moosavi-Dezfooli, Pascal Frossard, and Huaiyu Dai. Geoda: a geometric framework for black-box adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8446–8455, 2020.

[35] 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.

[36] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations*, 2018.

[37] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2018.

[38] Hong Wang, Hong Qian, and Yang Yu. Noisy derivative-free optimization with value suppression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

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