On the Limitations of Stochastic Pre-processing Defenses

Anonymous Author(s)
Affiliation
Address
email

Abstract

Defending against adversarial examples remains an open problem. A common belief is that randomness at inference increases the cost of finding adversarial inputs. An example of such a defense is to apply a random transformation to inputs prior to feeding them to the model. In this paper, we empirically and theoretically investigate such stochastic pre-processing defenses and demonstrate that they are flawed. First, we show that most stochastic defenses are weaker than previously thought; they lack sufficient randomness to withstand even standard attacks like projected gradient descent. This casts doubt on a long-held assumption that stochastic defenses invalidate attacks designed to evade deterministic defenses and force attackers to integrate the Expectation over Transformation (EOT) concept. Second, we show that stochastic defenses confront a trade-off between adversarial robustness and model invariance; they become less effective as the defended model acquires more invariance to their randomization. Future work will need to decouple these two effects. Our code is available in the supplementary material.

1 Introduction

Machine learning models are vulnerable to adversarial examples [4, 29], where an adversary can add imperceptible perturbations to the input of a model and change its prediction [5, 22]. Their discovery has motivated a wide variety of defense approaches [6, 8, 10, 25, 34, 35] along with the evaluation of their adversarial robustness [2, 24, 32]. Current evaluations mostly rely on adaptive attacks [2, 32], which require significant modeling and computational efforts. However, even when the attack succeeds, such evaluations may not always reveal the fundamental weaknesses of an examined defense. Without awareness of the underlying weaknesses, subsequent defenses may still conduct inadvertently weak adaptive attacks; this leads to overestimated robustness.

One popular class of defenses that demonstrates the above is the stochastic pre-processing defense, which relies on applying randomized transformations to inputs to provide robustness [10, 35]. Despite existing attack techniques designed to handle randomness [2, 3], there is an increasing effort to improve these defenses through a larger randomization space or more complicated transformations. For example, BaRT [24] employs 25 transformations, where the parameters of each transformation are further randomized. Due to the complexity of this defense, it was only broken recently (three years later) by Sitawarin et al. [28] with a complicated adaptive attack. Still, it is unclear how future defenses can avoid the pitfalls of existing defenses, largely because these pitfalls remain unknown.

In this paper, we investigate stochastic pre-processing defenses and explain their limitations both empirically and theoretically. First, we revisit previous stochastic pre-processing defenses and explain why such defenses are broken. We show that most stochastic defenses are not sufficiently randomized to invalidate standard attacks designed for deterministic defenses. Second, we study recent stochastic defenses that exhibit more randomness and show that they also face key limitations. In particular,

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

we identify a trade-off between their robustness and the model's invariance to their transformations. These defenses achieve a notion of robustness that results from reducing the model's invariance to the applied transformations. We outline our findings below. These findings suggest future work to find new ways of using randomness that decouples these two effects.

Most stochastic defenses lack sufficient randomness. Although Athalye et al. [2] and Tramèr et al. [32] have demonstrated the ineffectiveness of several stochastic defenses with techniques like Expectation over Transformation (EOT) [3], it remains unclear whether and why EOT is required to break them. A commonly accepted explanation is that EOT computes the "correct gradients" of models with randomized components [2, 32], yet the necessity of such correct gradients has not been explicitly discussed. To fill this gap, we examine a long-held assumption that stochastic defenses invalidate standard attacks designed for deterministic defenses.

Specifically, we revisit stochastic pre-processing defenses previously broken by EOT and examine their robustness *without* applying EOT. Interestingly, we find that most stochastic defenses lack sufficient randomness to withstand even standard attacks (that do not integrate any strategy to capture model randomness) like projected gradient descent (PGD) [22]. We then conduct a systematic evaluation to show that applying EOT is only beneficial when the defense is sufficiently randomized. Otherwise, standard attacks already perform well and the randomization's robustness is overestimated.

Trade-off between adversarial robustness and model invariance. When stochastic pre-processing defenses do have sufficient randomness, they must fine-tune the model using augmented training data to preserve utility in the face of randomness added. We characterize this procedure by the model's *invariance* to the applied defense, where we identify a trade-off between the model's robustness (provided by the defense) and its invariance to the applied defense. Stochastic pre-processing defenses become less effective when their defended model acquires more invariance to their transformations.

On the theoretical front, we present a theoretical setting where this trade-off provably exists. We show from this trade-off that stochastic pre-processing defenses provide robustness by inducing variance on the defended model, and must take back such variance to recover utility. We verify this trade-off with empirical evaluations on realistic datasets, models, and defenses. We observe that robustness drops when the defended model is fine-tuned on data processed by its defense to acquire higher invariance.

2 Related Work

Stochastic Defenses. Defending against adversarial examples remains an open problem, where a common belief is that inference-time randomness increases the cost of finding adversarial inputs. Early examples of such stochastic defenses include input transformations [10] and rescaling [35]. These defenses were broken by Athalye et al. [2] using techniques like EOT [3] to capture randomness. After that, more stochastic defenses were proposed but with inadvertently weak evaluations [23, 25, 34, 36], which were found ineffective by Tramèr et al. [32]. Subsequent stochastic defenses resort to larger randomization space like BaRT [24], which was only broken recently by Sitawarin et al. [28]. Randomized smoothing [6, 16, 17] leverages randomness to certify robustness. In this work, instead of designing adaptive attacks for individual defenses, we focus on the general stochastic pre-processing defenses and demonstrate their limitations.

Trade-offs for Adversarial Robustness. The trade-offs associated with adversarial robustness have been widely discussed in the literature. Prior work identified trade-offs between robustness and accuracy [33, 37] and even between the robustness against different types of adversarial examples [12, 31]. Recent work also investigated the trade-off between the model's robustness and invariance to input transformations, such as circular shifts [27] and rotations [13]. These trade-offs characterize a standalone model's own property – the model itself is less robust to adversarial examples when it becomes more invariant to certain transformations, without any defense. Our setting, however, is orthogonal to such analysis; the model we consider is protected by a stochastic pre-processing defense, and we aim to characterize the performance of the added defense to the model.

3 Preliminaries

Notation. Let $f : \mathcal{X} \to \mathbb{R}^C$ denote the classifier with pre-softmax outputs, where $\mathcal{X} = [0, 1]^d$ is the input space with d dimensions and C is the number of classes. We then consider a stochastic

pre-processing defense $t_{\boldsymbol{\theta}} : \mathcal{X} \to \mathcal{X}$, where $\boldsymbol{\theta}$ is the random variable drawn from some randomization space Θ that parameterizes the defense. The defended classifier can be written as $f_{\boldsymbol{\theta}}(\boldsymbol{x}) \coloneqq f(t_{\boldsymbol{\theta}}(\boldsymbol{x}))$.

Let $F(\boldsymbol{x}) \coloneqq \arg\max_{i \in \mathcal{Y}} f_i(\boldsymbol{x})$ denote the classifier that returns the predicted label, where f_i is the output of the i-th class and $\mathcal{Y} = [C]$ is the label space. Similarly, we use $F_{\boldsymbol{\theta}}$ and $f_{\boldsymbol{\theta},i}$ to denote the prediction and class-output of the stochastic classifier $f_{\boldsymbol{\theta}}$. Since this classifier returns varied outputs for a fixed input, it determines the final prediction by aggregating n independent inferences with strategies like majority vote. We discuss these strategies and the choice of n in Appendix A.1.

Adversarial Examples. Given an image $\boldsymbol{x} \in \mathcal{X}$ and a classifier F, the adversarial example $\boldsymbol{x}' \coloneqq \boldsymbol{x} + \boldsymbol{\delta}$ is visually similar to $\boldsymbol{x}$ but either misclassified (i.e., $F(\boldsymbol{x}') \neq F(\boldsymbol{x})$) or classified as a target class y' chosen by the attacker (i.e., $F(\boldsymbol{x}') = y'$). Attack algorithms generate adversarial examples by searching for $\boldsymbol{\delta}$ such that $\boldsymbol{x}'$ fools the classifier while minimizing $\boldsymbol{\delta}$ under some distance metrics; for instance, the ℓ_p norm constraint $\|\boldsymbol{\delta}\|_p \leq \epsilon$ for a perturbation budget ϵ.

Projected Gradient Descent (PGD). PGD [22] is one of the most established attacks to evaluate adversarial example defenses. Given a benign example $\boldsymbol{x}^0$ and its ground-truth label y, each iteration of the untargeted PGD attack (with ℓ_∞ norm budget ϵ) can be formulated as

$$\boldsymbol{x}^{i+1} \leftarrow \boldsymbol{x}^i + \alpha \cdot \operatorname{sgn}\big\{\nabla \mathcal{L}\big(f_{\boldsymbol{\theta}}(\boldsymbol{x}^i), y\big)\big\}, \tag{1}$$

where α is the step size, $\mathcal{L}$ is the loss function, and each iteration is projected to the ℓ_∞ ball around $\boldsymbol{x}^0$ of radius ϵ. We use PGD-k to denote the PGD attack with k steps. We outline formulations for other settings and norms in Appendix A.2.

Expectation over Transformation (EOT). Since the classifier $f_{\boldsymbol{\theta}}$ is stochastic, the defense evaluation literature [2, 32] argues that attacks should target the *expectation* of the gradient using Expectation over Transformation (EOT) [3], which reformulates the PGD attack as

$$\boldsymbol{x}^{i+1} \leftarrow \boldsymbol{x}^i + \alpha \cdot \operatorname{sgn}\Big\{\mathbb{E}_{\boldsymbol{\theta} \sim \Theta}\Big[\nabla \mathcal{L}\big(f_{\boldsymbol{\theta}}(\boldsymbol{x}^i), y\big)\Big]\Big\} \approx \boldsymbol{x}^i + \alpha \cdot \operatorname{sgn}\Big\{\frac{1}{m}\sum_{j=1}^{m} \nabla \mathcal{L}\big(f_{\boldsymbol{\theta}_j}(\boldsymbol{x}^i), y\big)\Big\}, \tag{2}$$

where m is the number of samples to estimate the expectation and $\boldsymbol{\theta}_j \overset{\text{iid}}{\sim} \Theta$ are sampled parameters for the defense. We use EOT-m to denote the EOT technique with m samples at each PGD step.

In addition, for a fair comparison among attacks with different PGD steps and EOT samples, we quantify the attack's strength by its total number of gradient computations. For example, attacks using PGD-k and EOT-m will have strength $k \times m$. Although white-box attacks are typically not constrained in this way, it allows for a fair comparison when attacks have finite computing resources (e.g., when EOT is not parallelizable). We discuss more about this quantification in Appendix A.3.

4 Most Stochastic Defenses Lack Sufficient Randomness

Athalye et al. [2] and Tramèr et al. [32] demonstrate adaptive evaluation of stochastic defenses with the application of EOT. However, it remains unclear why EOT is required to break these stochastic defenses. While a commonly accepted explanation is that EOT computes the "correct gradients" of models with randomized components [2, 32], the necessity of such correct gradients has not been explicitly discussed.To fill this gap, we revisit stochastic defenses previously broken by EOT and examine their robustness *without* applying EOT. Interestingly, we find that applying EOT is mostly *unnecessary* when evaluating existing stochastic defenses.

Case Study: Random Rotation. We start with a simple stochastic defense that randomly rotates the input image for $\boldsymbol{\theta} \in [-90, 90]$ degrees (chosen at uniform) before classification. This defense is representative for most pre-processing defenses [10, 24, 35]. We evaluate this defense on 1,000 ImageNet images with PGD-k and EOT-m under the constraint $k \times m = 50$, as discussed in Section 3. All attacks use maximum ℓ_∞ perturbation $\epsilon = 8/255$ with step size chosen from $\alpha \in \{1/255, 2/255\}$. The results are shown in Table 1, where

Table 1: Evaluation of the random rotation with PGD-k and EOT-m.

Attacks	k	m	Success Rate
Untargeted	10	5	100%
	50	1	100%
Targeted	10	5	99.0%
	50	1	99.0%

Table 2: The missing ablation study of adaptive evaluations of stochastic defenses in the literature. Notations: attack iterations k, EOT samples m, learning rate α, number of gradient queries $k \times m$. The details of these defenses and their evaluation settings are in Appendix B.

Defenses	Original Adaptive Evaluation (w/ EOT)					Our Ablation Study (w/o EOT)				
	k	m	α	$k \times m$	Success Rate	k	m	α	$k \times m$	Success Rate
Guo et al. [10]	1,000	30	0.1	30,000	100%	1,000	1	0.001	1,000	99.0%
Xie et al. [35]	1,000	30	0.1	30,000	100%	200	1	0.1	200	100%
Dhillon et al. [8]	500	10	0.1	5,000	100%	500	1	0.1	500	100%
Xiao et al. [34]	100	1,000	0.01	100,000	100%	40,000	1	0.1/255	40,000	98.4%
Roth et al. [25]	100	40	0.2/255	4,000	100%	4,000	1	0.1/255	4,000	96.1%

PGD-50 performs equally well as PGD-10 combined with EOT-5. This observation suggests that *some stochastic defenses are already breakable without applying EOT*, casting doubt on a long-held assumption that stochastic defenses simply invalidate attacks designed for deterministic defenses.

Comprehensive Evaluations. We then extend the above case study to other stochastic defenses evaluated in the literature. Specifically, we replicate the (untargeted) adaptive evaluation of stochastic defenses from Athalye et al. [2] and Tramèr et al. [32] with their official implementation. We only change the attack's hyper-parameters (e.g., number of iterations and learning rate) and disable EOT by setting its number of samples to one ($m = 1$), which avoids potential implementation flaws if removed from the source code. The comparison between evaluations with and without applying EOT is summarized in Table 2, which serves as a missing ablation study of adaptive evaluations in the literature. The experimental settings are identical within each row (detailed in Appendix B).

Interestingly, we find it *unnecessary* to break these defenses with EOT, as long as the standard attack runs for more iterations with a smaller learning rate. For such defenses, standard iterative attacks already contain an *implicit expectation* across iterations to capture the limited randomness. This observation implies that most stochastic defenses lack sufficient randomness to withstand even standard attacks designed for deterministic defenses. Therefore, increasing randomness becomes a promising approach to enhancing stochastic defenses, as adopted by recent defenses [6, 24]. Note that this ablation study only aims to inspire potential ways of enhancing stochastic defenses; it does not invalidate EOT for stronger adaptive evaluations of stochastic defenses.

5 Trade-offs between Robustness and Invariance

When stochastic pre-processing defenses *do have* sufficient randomness, they must ensure that the utility of the defended model is preserved in the face of randomness. To achieve high utility, existing defenses mostly rely on augmentation invariance through *trained invariance* [21]. In such a case, the invariance is achieved by applying the defense's randomness to the training data so as to guide the model in learning their transformations. For defenses based on stochastic pre-processor $t_{\boldsymbol{\theta}}$, each data sample from the dataset gets augmented with $t_{\boldsymbol{\theta}}$ sampled from the randomization space Θ, and the risk is minimized over such augmented data.

The defended classifier $F_{\boldsymbol{\theta}}(\boldsymbol{x}) \coloneqq F(t_{\boldsymbol{\theta}}(\boldsymbol{x}))$ is invariant under the randomization space Θ if

$$F(t_{\boldsymbol{\theta}}(\boldsymbol{x})) = F(\boldsymbol{x}), \quad \forall\, \boldsymbol{\theta} \in \Theta, \boldsymbol{x} \in \mathcal{X}. \tag{3}$$

As we can observe from the definition, invariance has direct implications on the performance of stochastic pre-processing defenses. If the classifier is invariant under the defense's randomization space Θ as is defined in Equation (3), then the defense should not work – computing the model and its gradients over randomization $\boldsymbol{\theta} \in \Theta$ is the same as if $t_{\boldsymbol{\theta}}$ was not applied at all. This observation suggests a direct coupling between invariance and performance of the defense: the more invariant, hence performant, the model is under a given randomization space, the less protection such a defense would provide. In this section, we present a simple theoretical setting where this coupling provably exists; detailed computations and illustrations of this setting are in Appendix C.1.

Binary Classification Task. We consider a class-balanced dataset $\mathcal{D}$ consisting of input-label pairs (x, y) with $y \in \{-1, +1\}$ and $x|y \sim \mathcal{N}(y, 1)$, where $\mathcal{N}(\mu, \sigma^2)$ is a normal distribution with mean μ and variance σ^2. Moreover, an ℓ_∞-bounded adversary perturbs the input with a small ϵ to fool the classifier. We quantify the classifier's robustness by its robust accuracy, i.e., the ratio of correctly classified samples that remain correct after being perturbed by the adversary. Note that while we use $x + \epsilon$ to denote the perturbed input, its actual value can take within the range $[x - \epsilon, x + \epsilon]$.

Undefended Classification. We start with the optimal linear classifier $F(x) \coloneqq \operatorname{sgn}(x)$ without any defense. This classifier attains robust accuracy

$$\Pr\big[F(x+\epsilon) = y \mid F(x) = y\big] = \frac{\Pr\big[F(x+\epsilon) = y \wedge F(x) = y\big]}{\Pr\big[F(x) = y\big]} = \frac{\Phi(1-\epsilon)}{\Phi(1)}, \tag{4}$$

where Φ is the cumulative distribution function of $\mathcal{N}(0,1)$.

Defended Classification. We then try to improve adversarial robustness by introducing a stochastic pre-processing defense $t_\theta(x) \coloneqq x + \theta$, where $\theta \sim \mathcal{N}(1, \sigma^2)$ is the random variable parameterizing the defense. This defense characterizes common pre-processing defenses that enforce randomness while shifting the input distribution. Here, the processed input follows a shifted distribution $t_\theta(x) \sim \mathcal{N}(y+1, 1+\sigma^2)$. As a result, the defended classifier $F_\theta(x) = \operatorname{sgn}(x+\theta)$ has robust accuracy

$$\Pr\big[F_\theta(x+\epsilon) = y \mid F_\theta(x) = y\big] = \frac{\Pr\big[F_\theta(x+\epsilon) = y \wedge F_\theta(x) = y\big]}{\Pr\big[F_\theta(x) = y\big]} = \frac{\Phi'(-\epsilon) + \Phi'(2-\epsilon)}{\Phi'(0) + \Phi'(2)}, \tag{5}$$

where $\Phi'(x) \coloneqq \Phi(x/\sqrt{1+\sigma^2})$ is the cumulative distribution function of $\mathcal{N}(0, 1+\sigma^2)$. At this point, we have not fit the classifier on processed inputs. Due to its lack of invariance, the defended classifier has low utility yet higher robust accuracy than the undefended one in Equation (4).

Defended Classification (Trained Invariance). As discussed above, one critical step of stochastic pre-processing defenses is to preserve the defended model's utility by minimizing the risk over augmented data $t_\theta(x)$, which leads to a new defended classifier $F^+_\theta(x) = \operatorname{sgn}(x+\theta-1)$. As a result, this new defended classifier achieves higher invariance with robust accuracy

$$\Pr\big[F^+_\theta(x+\epsilon) = y \mid F^+_\theta(x) = y\big] = \frac{\Pr\big[F^+_\theta(x+\epsilon) = y \wedge F^+_\theta(x) = y\big]}{\Pr\big[F^+_\theta(x) = y\big]} = \frac{\Phi'(1-\epsilon)}{\Phi'(1)}, \tag{6}$$

which is less robust than the previous less-invariant classifier F_θ in Equation (5). However, one may observe that this classifier, though loses some robustness compared with F_θ, is still more robust than the original undefended classifier F in Equation (4). This part of robustness comes from the changed data distribution due to the defense's randomness. It shows that we have not achieved perfect invariance to the defense's randomness, thus gaining some robustness at the cost of utility.

Defended Classification (Perfect Invariance). Furthermore, these defenses usually leverage majority vote to obtain stable predictions, which finally produces a perfectly invariant defended classifier

$$F^*_\theta(x) = \operatorname{sgn}\left\{\frac{1}{n}\sum_{i=1}^{n} F^+_{\theta_i}(x)\right\} = \operatorname{sgn}\left\{\frac{1}{n}\sum_{i=1}^{n} \operatorname{sgn}(x+\theta_i-1)\right\} \to \operatorname{sgn}(x) = F(x), \tag{7}$$

where $\theta_i \overset{\text{iid}}{\sim} \mathcal{N}(1, \sigma^2)$ are sampled parameters. In such a case, the defended classifier reduces to the original undefended classifier with the original robust accuracy:

$$\Pr\big[F^*_\theta(x+\epsilon) = y \mid F^*_\theta(x) = y\big] = \Pr\big[F(x+\epsilon) = y \mid F(x) = y\big] = \frac{\Phi(1-\epsilon)}{\Phi(1)}. \tag{8}$$

Summary. The above theoretical setting illustrates how stochastic pre-processing defenses first induce variance on the binary classifier we consider to provide adversarial robustness in Equation (5), and how they finally take back such variance in Equations (6) and (8) to recover utility. In Appendix C.2, we extend this coupling to a general trade-off where we can control the invariance. This trade-off demonstrates that stochastic pre-processing defenses provide robustness by explicitly reducing the model's invariance to added randomized transformations.

6 Experiments

Our experiments are designed to answer the following two questions.

Q1: What properties make applying EOT beneficial when evaluating stochastic defenses?

We show that applying EOT is only beneficial when the defense is sufficiently randomized; otherwise standard attacks already perform well and leave no room for EOT to improve.

Q2: What is the limitation of stochastic defenses when they do have sufficient randomness?

We show a trade-off between the stochastic defense's robustness and the model's invariance to the defense itself. Such defenses become less effective when the defended model achieves higher invariance to their randomness, as required to preserve utility under the defense.

6.1 Experimental Settings

Datasets & Models. We conduct all experiments on ImageNet [26] and ImageNette [9]. For ImageNet, our test data consists of 1,000 images randomly sampled from the validation set. ImageNette is a ten-class subset of ImageNet, and we test on its validation set. We adopt various ResNet [11] models. For defenses with low randomness, we evaluate them on ImageNet with a pre-trained ResNet-50 with Top-1 accuracy 75.9%. For defenses with higher randomness (thus requiring fine-tuning), we switch to ImageNette and a pre-trained ResNet-34 with Top-1 accuracy 96.9% to reduce the training cost like previous work [28]. These models are fine-tuned on the training data processed by tested defenses. As a special case, we also evaluate randomized smoothing on ImageNet using the ResNet-50 models from Cohen et al. [6]. More details of datasets and models can be found in Appendices D.1 and D.2.

Defenses & Metrics. We focus on stochastic defenses allowing us to increase randomness: randomized smoothing [6] and BaRT [24]. For randomized smoothing, we vary the variance of the added Gaussian noise. For BaRT, we vary the number κ of applied randomized transformations. Note that we have evaluated other stochastic defenses and discussed their low randomness in Section 4. We measure the defense's performance by the defended model's *benign accuracy* and the attack's *success rate*, all evaluated with majority vote over $n = 500$ predictions. The attack's success rate is the ratio of samples that do not satisfy the attack's objective prior to the attack but satisfy it after the attack. For example, we discard samples that were misclassified before being perturbed in untargeted attacks. Details of the evaluated defenses can be found in Appendix D.3.

Attacks. We evaluate defenses with standard PGD combined with EOT and focus on the ℓ_∞-bounded adversary with a perturbation budget $\epsilon = 8/255$ in both untargeted and targeted settings. We only conduct adaptive evaluations, where the defense is included in the attack loop with non-differentiable components captured by BPDA [2]. We also utilize AutoPGD [7] to avoid selecting the step size when it is computationally expensive to repeat some experiments. More details of the attack's setting and implementation can be found in Appendix D.4.

6.2 Q1: Evaluate the Benefits of Applying EOT under Different Settings

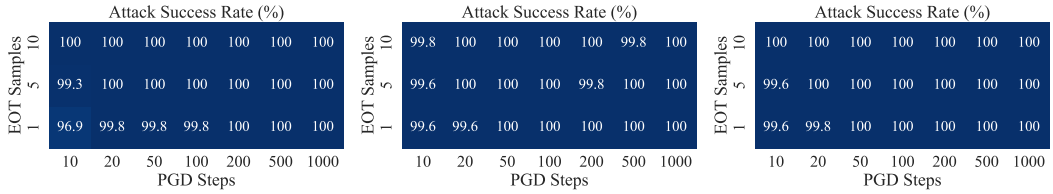
In Section 4, we showed that standard attacks are sufficient to break most stochastic defenses due to their lack of randomness. Here, we aim to understand what properties make applying EOT beneficial when evaluating stochastic defenses. We design a systematic evaluation of stochastic defenses with different level of randomness and check if applying EOT improves the attack.

Stochastic Defenses with Low Randomness. We start with BaRT's noise injection defense, which perturbs the input image with noise of distributions and parameters chosen at random. While this defense has low randomness, it yields meaningful results. We evaluate this defense with various combinations of PGD and EOT. The performance of untargeted and targeted attacks is shown in Figure 1. We test multiple step sizes and summarize their best results (discussed in Appendix E.1).

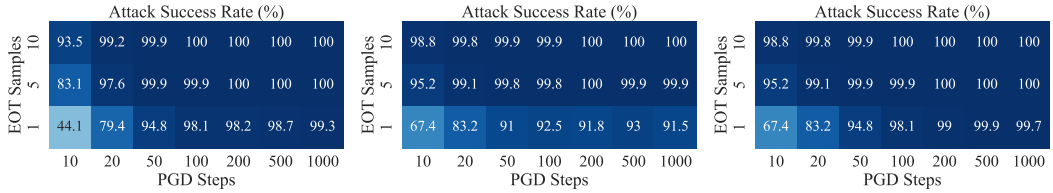
In this case, standard PGD attacks are already good enough when the defense has insufficient randomness, leaving no space for improvements from EOT. In Figure 1f, both (1) PGD-10 combined with EOT-10 and (2) PGD-100 combined with EOT-1 have near 100% success rates. This result is consistent with our observations in Section 4 in both untargeted and targeted settings[1].

Stochastic Defenses with Higher Randomness. We then examine the randomized smoothing defense that adds Gaussian noise to the input image. Although this defense was originally proposed for certifiable adversarial robustness, we adopt it to evaluate how randomness affects the benefits of applying EOT. Similarly, we evaluate this defense with PGD and EOT of different settings with a focus on the *targeted* attack. The results are shown in Figure 2.

[1]The only caveat is that targeted attacks are more likely to benefit from EOT, as their objectives are stricter and may have better performance with gradients of higher precision.

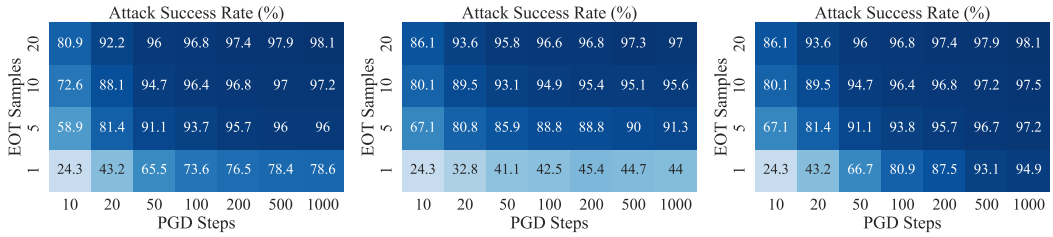


(a) Untargeted (step size = 1/255) (b) Untargeted (step size = 2/255) (c) Untargeted (step size = **best**)

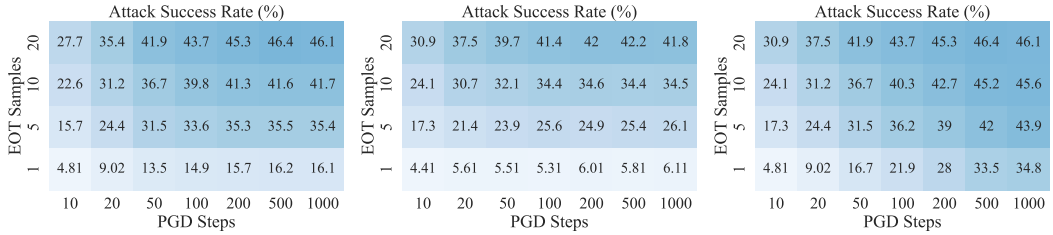


(d) Targeted (step size = 1/255) (e) Targeted (step size = 2/255) (f) Targeted (step size = **best**)

Figure 1: Evaluation of BaRT's noise injection defense on ImageNet. Standard PGD without applying EOT (i.e., applying EOT-1) is already good enough, leaving limited space for EOT to improve.



(a) $\sigma = 0.25$ (step size = 2/255) (b) $\sigma = 0.25$ (step size = 4/255) (c) $\sigma = 0.25$ (step size = **best**)



(d) $\sigma = 0.50$ (step size = 2/255) (e) $\sigma = 0.50$ (step size = 4/255) (f) $\sigma = 0.50$ (step size = **best**)

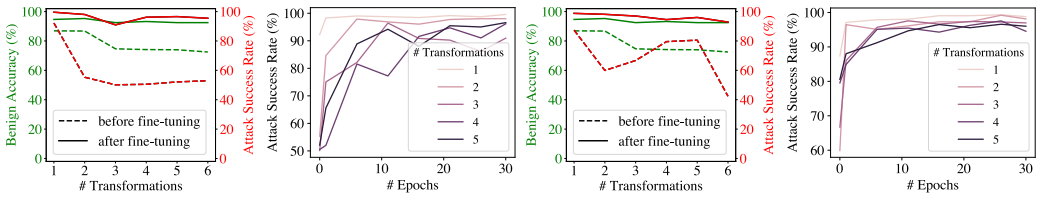
Figure 2: Evaluation of randomized smoothing on ImageNet (targeted attacks). PGD performs well on lower variance ($\sigma = 0.25$) if running for more steps. For a larger variance ($\sigma = 0.50$), applying EOT starts to improve the attack significantly (for a fixed number of gradient computations).

We observe that EOT starts to improve the attack when the defense has a higher level of randomness. For a fixed number of PGD steps, applying EOT significantly improves the attack in most of the settings. For a fixed attack strength (i.e., number of gradient computations), applying EOT always outperforms standalone PGD. In Figure 2f, for example, PGD-100 combined with EOT-10 is 5.5% higher than PGD-1,000 with EOT-1 (40.3% vs. 34.8%).

Takeaways. Applying EOT is only beneficial when the defense has sufficient randomness, such as randomized smoothing with $\sigma = 0.5$. This observation suggests that stochastic defenses only make standard attacks suboptimal when they have sufficient randomness. However, most existing stochastic defenses did not achieve this criterion, as we showed in Section 4.

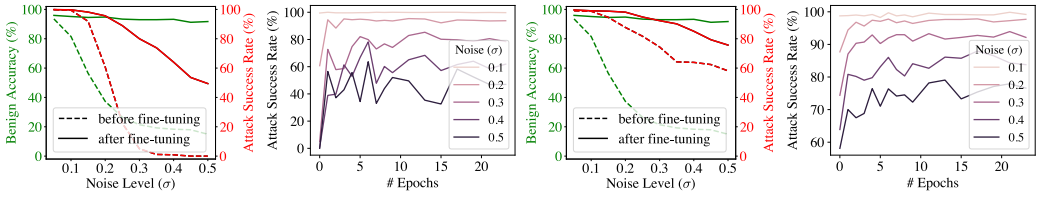
6.3 Q2: Evaluate the Trade-off between Robustness and Invariance

In Section 5, we present a theoretical setting where the trade-off between robustness and invariance provably exists; stochastic defenses become less robust when the defended model achieves higher



(a) Targeted Attacks (view by size) (b) Targeted Attacks (view by epochs) (c) Untargeted Attacks (view by size) (d) Untargeted Attacks (view by epochs)

Figure 3: Performance of the BaRT defense on ImageNette with different numbers of transformations before and after fine-tuning the model. While the model achieves higher invariance, the defense becomes nearly ineffective[2], as evident from the top solid red curves in (a) and (c).



(a) Targeted Attacks (view by levels) (b) Targeted Attacks (view by epochs) (c) Untargeted Attacks (view by levels) (d) Untargeted Attacks (view by epochs)

Figure 4: Performance of the randomized smoothing defense on ImageNette with different noise levels before and after fine-tuning the model. While the model achieves higher invariance, the defense becomes less effective[3], as evident from the gap between dashed and solid red curves in (a) and (c).

invariance to their randomness. Here, we demonstrate this trade-off on realistic datasets, models, and defenses. In particular, we choose defenses with sufficient randomness (achieved in different ways) and compare their performance when being applied to models of different levels of invariance, where the invariance is achieved by applying the defense's randomness to the training data so as to guide the model in learning their transformations.

Randomness through Transformations. We first examine the BaRT defense, which pre-processes input images with κ randomly composited stochastic transformations. It represents defenses aiming to increase randomness through diverse input transformations. Since our objective is to demonstrate the trade-off, it suffices to evaluate a subset of BaRT with $\kappa \leq 6$ transformations; this also avoids the training cost of evaluating the original BaRT with $\kappa = 25$. Figure 3 shows the performance of this defense with models before and after fine-tuning on its processed training data.

In Figures 3a and 3c, we first observe that fine-tuning indeed increases the model's invariance to the applied defense's randomness; the utility's dashed green curves are improved to the solid green curves beyond 90%. However, as the model achieves higher invariance, the defense becomes nearly ineffective; the attack's dashed red curves boost to the solid red curves near 100%. The same attack's effectiveness throughout the fine-tuning procedure further verifies this observation, as shown in Figures 3b and 3d. It shows a clear trade-off between the defense's robustness and the model's invariance. That is, stochastic defenses start to lose robustness when their defended models achieve higher invariance to their transformations.

Randomness through Noise Levels. We then examine the randomized smoothing defense that adds Gaussian noise to the input image. Unlike BaRT's diverse transformations, randomized smoothing increases randomness directly through the added noise's variance σ^2. This allows us to rigorously increase the randomness without unexpected artifacts like non-differentiable components. We evaluate the performance of this defense ($\sigma \leq 0.5$) with models before and after fine-tuning on training data perturbed with designated Gaussian noise. The results are shown in Figure 4.

[2]The defense may not grow stronger with more transformations, which is a drawback of BaRT that we will discuss in Appendix D.3. Yet, our evaluations focus on the fact that solid curves are above the dashed curves.

[3]One may also observe a trade-off between robustness and *utility* by examining the curve's horizontal trend. However, we focus on the trade-off between robustness and *invariance*, which manifests in the vertical gap.

In Figures 4a and 4c, fine-tuning improves the model's invariance, but the defense also becomes significantly weaker during this process. For example, the targeted attack is nearly infeasible when the model is variant to the large noise ($\sigma \geq 0.3$), yet is significantly more effective when the model becomes invariant. The fine-tuning process in Figures 4b and 4d also verifies that stochastic defenses become weaker when their defended models become more invariant to their randomness.

Takeaways. For both the BaRT and the randomized smoothing defense, we observe a clear trade-off between the defense's robustness and the model's invariance to randomness. In particular, we find that stochastic defenses lose adversarial robustness when their defended models achieve higher invariance to their randomness. Our finding implies that such defenses would become ineffective when their defended models are perfectly invariant to their randomness.

7 Discussions

In this section, we discuss several questions that arose from our study of the pre-processing defense. The limitations and potential negative societal impacts of this work are discussed in Appendix F.

What do pre-processing defenses really do? We show that existing stochastic pre-processing defenses do not introduce inherent robustness to the prediction task. Instead, they shift the input distribution through randomness and transformations, which results in variance and introduces errors during prediction. The observed "robustness", in an unusual meaning for this literature, is a result of these errors. This is fundamentally different from the inherent robustness provided by adversarial training [22]. Although defenses like adversarial training still cost accuracy [33, 37], they do not intentionally introduce errors like stochastic pre-processing defenses.

However, stochastic pre-processing defenses *do make* the attack harder when the adversary has only limited knowledge of the defense's transformations, e.g., in a low-query setting. In such a case, the defense practically introduces noise to the attack's optimization procedure, making it difficult for a low-query adversary to find adversarial examples that consistently cross the probabilistic decision boundary. In contrast, our theoretical analysis in Section 5 considers a powerful adversary with full knowledge of the defense's randomization space; hence it can optimize the adversarial example directly towards the defended model's decision boundary in expectation.

How should we utilize randomness? Stochastic defenses should rely on randomness that exploits the properties of the prediction task. For example, some speech and vision problems are inherently divisible into orthogonal subproblems (e.g., performing classification on portions of the speech spectrum for keyword spotting [1]), where randomness can apply to the space of these subproblems. Such defenses decouple robustness and invariance because (1) the models perform well on each independent subproblem, and (2) attacks on one subproblem do not transfer to the other subproblems. As such, an attacker has to target all possible subproblems, which reduces its effective attack budget.

What are implications for adaptive attackers? Our findings demonstrate that an adaptive attacker needs to consider the spectrum of available standard attack algorithms, instead of just focusing on a given attack algorithm because of the defense's design. As we discover in this paper, EOT can be unnecessary for seemingly immune stochastic defenses, yet its application to break these said defenses gives a false impression about their security against weak attackers. When evaluating the robustness of a defense, the adaptive attack should start by tuning standard approaches, before resorting to more involved attack strategies. This approach helps us to identify the minimally capable attack that breaks the defense and develop a better understanding of the defense's fundamental weaknesses.

8 Conclusion

In this paper, we investigate stochastic pre-processing defenses and explain their limitations both empirically and theoretically. We show that most stochastic pre-processing defenses are weaker than previously thought, and recent defenses that indeed exhibit more randomness still face a trade-off between their robustness and the model's invariance to their transformations. While defending against adversarial examples remains an open problem and designing proper adaptive evaluations is arguably challenging, we demonstrate that stochastic pre-processing defenses are fundamentally flawed in their current form. Our findings suggest that future work will need to find new ways of using randomness that decouples the robustness and invariance.

References

[1] Shimaa Ahmed, Ilia Shumailov, Nicolas Papernot, and Kassem Fawaz. Towards more robust keyword spotting for voice assistants. In *31st USENIX Security Symposium*, 2022.

[2] Anish Athalye, Nicholas Carlini, and David A. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 274–283. PMLR, 2018. URL http://proceedings.mlr.press/v80/athalye18a.html.

[3] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 284–293. PMLR, 2018. URL http://proceedings.mlr.press/v80/athalye18b.html.

[4] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Srndic, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In Hendrik Blockeel, Kristian Kersting, Siegfried Nijssen, and Filip Zelezný, editors, *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III*, volume 8190 of *Lecture Notes in Computer Science*, pages 387–402. Springer, 2013. doi: 10.1007/978-3-642-40994-3_25. URL https://doi.org/10.1007/978-3-642-40994-3_25.

[5] Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pages 39–57, 2017. doi: 10.1109/SP.2017.49. URL https://doi.org/10.1109/SP.2017.49.

[6] Jeremy M. Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified adversarial robustness via randomized smoothing. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 1310–1320. PMLR, 2019. URL http://proceedings.mlr.press/v97/cohen19c.html.

[7] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 2206–2216. PMLR, 2020. URL http://proceedings.mlr.press/v119/croce20b.html.

[8] Guneet S. Dhillon, Kamyar Azizzadenesheli, Zachary C. Lipton, Jeremy Bernstein, Jean Kossaifi, Aran Khanna, and Animashree Anandkumar. Stochastic activation pruning for robust adversarial defense. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=H1uR4GZRZ.

[9] FastAI. The imagenette dataset. URL https://github.com/fastai/imagenette.

[10] Chuan Guo, Mayank Rana, Moustapha Cissé, and Laurens van der Maaten. Countering adversarial images using input transformations. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=SyJ7ClWCb.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90. URL https://doi.org/10.1109/CVPR.2016.90.

[12] Jörn-Henrik Jacobsen, Jens Behrmann, Richard S. Zemel, and Matthias Bethge. Excessive invariance causes adversarial vulnerability. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=BkfbpsAcF7.

[13] Sandesh Kamath, Amit Deshpande, Subrahmanyam Kambhampati Venkata, and Vineeth N Balasubramanian. Can we have it all? on the trade-off between spatial and adversarial robustness of neural networks. In M. Ranzato, A. Beygelzimer, K. Nguyen, P. S. Liang, J. W. Vaughan, and Y. Dauphin, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 27462–27474. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/e6ff107459d435e38b54ad4c06202c33-Paper.pdf.

[14] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

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

[16] Mathias Lécuyer, 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 2019, San Francisco, CA, USA, May 19-23, 2019*, pages 656–672. IEEE, 2019. doi: 10.1109/SP.2019.00044. URL https://doi.org/10.1109/SP.2019.00044.

[17] Bai Li, Changyou Chen, Wenlin Wang, and Lawrence Carin. Certified adversarial robustness with additive noise. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 9459–9469, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/335cd1b90bfa4ee70b39d08a4ae0cf2d-Abstract.html.

[18] Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with warm restarts. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=Skq89Scxx.

[19] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

[20] James Lucas, Shengyang Sun, Richard S. Zemel, and Roger B. Grosse. Aggregated momentum: Stability through passive damping. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=Syxt5oC5YQ.

[21] Clare Lyle, Mark van der Wilk, Marta Kwiatkowska, Yarin Gal, and Benjamin Bloem-Reddy. On the benefits of invariance in neural networks, 2020.

[22] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

[23] Tianyu Pang, Kun Xu, and Jun Zhu. Mixup inference: Better exploiting mixup to defend adversarial attacks. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=ByxtC2VtPB.

[24] Edward Raff, Jared Sylvester, Steven Forsyth, and Mark McLean. Barrage of random transforms for adversarially robust defense. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6528–6537. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00669. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Raff_Barrage_of_Random_Transforms_for_Adversarially_Robust_Defense_CVPR_2019_paper.html.

[25] Kevin Roth, Yannic Kilcher, and Thomas Hofmann. The odds are odd: A statistical test for detecting adversarial examples. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 5498–5507. PMLR, 2019. URL http://proceedings.mlr.press/v97/roth19a.html.

[26] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y. URL https://doi.org/10.1007/s11263-015-0816-y.

[27] Vasu Singla, Songwei Ge, Basri Ronen, and David Jacobs. Shift invariance can reduce adversarial robustness. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 1858–1871. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/0e7c7d6c41c76b9ee6445ae01cc0181d-Paper.pdf.

[28] Chawin Sitawarin, Zachary Golan-Strieb, and David Wagner. Demystifying the adversarial robustness of random transformation defenses. In *The AAAI-22 Workshop on Adversarial Machine Learning and Beyond*, 2021. URL https://openreview.net/forum?id=p4SrFydw05.

[29] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6199.

[30] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2818–2826. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.308. URL https://doi.org/10.1109/CVPR.2016.308.

[31] Florian Tramèr, Jens Behrmann, Nicholas Carlini, Nicolas Papernot, and Jörn-Henrik Jacobsen. Fundamental tradeoffs between invariance and sensitivity to adversarial perturbations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 9561–9571. PMLR, 2020. URL http://proceedings.mlr.press/v119/tramer20a.html.

[32] Florian Tramèr, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/11f38f8ecd71867b42433548d1078e38-Abstract.html.

[33] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=SyxAb30cY7.

[34] Chang Xiao, Peilin Zhong, and Changxi Zheng. Enhancing adversarial defense by k-winners-take-all. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=Skgvy64tvr.

[35] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan L. Yuille. Mitigating adversarial effects through randomization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=Sk9yuql0Z.

[36] Yuzhe Yang, Guo Zhang, Zhi Xu, and Dina Katabi. Me-net: Towards effective adversarial robustness with matrix estimation. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 7025–7034. PMLR, 2019. URL http://proceedings.mlr.press/v97/yang19e.html.

[37] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 7472–7482. PMLR, 2019. URL http://proceedings.mlr.press/v97/zhang19p.html.

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]
   (b) Did you describe the limitations of your work? [Yes] See Appendix F.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Appendix F.
   (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] We outline all details for theoretical results in Appendix C.
   (b) Did you include complete proofs of all theoretical results? [Yes] We include complete proofs in Appendix C.
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] Our code is available in the supplementary material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] We specify brief settings in Section 6.1 and provide complete details in Appendix D.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] We study stochastic defenses, whose prediction procedure is already a majority vote over multiple times.
   (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 Appendix D.
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]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A] We only included our own code for evaluation.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] We only used public datasets as discussed in Section 6.1.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] We did not use such data.
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]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]