Combating Adversaries with Anti-Adversaries

Motasem Alfarra[1], Juan C. Pérez[1], Ali Thabet[2], Adel Bibi[3], Philip H.S. Torr[3], and Bernard Ghanem[1]

[1] King Abdullah University of Science and Technology (KAUST), [2] Facebook Reality Labs, [3] University of Oxford
motasem.alfarra@kaust.edu.sa

Abstract

Deep neural networks are vulnerable to small input perturbations known as adversarial attacks. Inspired by the fact that these adversaries are constructed by iteratively minimizing the confidence of a network for the true class label, we propose the anti-adversary layer, aimed at countering this effect. In particular, our layer generates an input perturbation in the opposite direction of the adversarial one and feeds the classifier a perturbed version of the input. Our approach is training-free and theoretically supported. We verify the effectiveness of our approach by combining our layer with both nominally and robustly trained models and conduct large-scale experiments from black-box to adaptive attacks on CIFAR10, CIFAR100, and ImageNet. Our layer significantly enhances model robustness while coming at no cost on clean accuracy.[1]

Introduction

Deep Neural Networks (DNNs) are vulnerable to small input perturbations known as adversarial attacks (Szegedy et al. 2013; Goodfellow, Shlens, and Szegedy 2015). In particular, a classifier f, which correctly classifies x, can be fooled by a small adversarial perturbation δ into misclassifying $(x + \delta)$ even though x and $(x + \delta)$ are indistinguishable to the human eye. Such perturbations can compromise trust in DNNs, hindering their use in safety- and security-critical applications, *e.g.* self-driving cars (Sitawarin et al. 2018). While there have been extensive efforts aimed at training DNNs that are robust to adversarial attacks, assessing the robustness of defenses remains an elusive task. This difficulty is due to the following reasons. (**i**) The robustness of models varies according to the information an attacker is assumed to know, *e.g.* training data, gradients, logits, *etc.*, which, for ease, dichotomously categorizes adversaries as being black- or white-box. Consequently, this categorization results in difficulties when comparing defenses tailored to a specific type of adversaries. For instance, several defenses crafted for robustness against white-box adversaries were later broken by their weaker black-box counterparts (Papernot et al. 2016a; Brendel, Rauber, and Bethge 2018). (**ii**) In addition, empirically-evaluated robustness can be overestimated if fewer efforts are invested into *adaptively* constructing a stronger attack (Tramer et al. 2020; Carlini et al. 2019). The lack of reliable assessments has been responsible for a false sense of security, as several thought-to-be-strong defenses against white-box adversaries were later broken with better carefully-crafted adaptive attacks (Athalye, Carlini, and Wagner 2018). The few defenses that have stood the test of time usually come at the expense of costly training and performance degradation on clean samples (Tsipras et al. 2019). Even worse, while most of these defenses are meant to resist white-box attacks, little effort has been invested into resisting the black-box counterparts, which may be more common and practical (Byun, Go, and Kim 2021), as online APIs such as IBM Watson and Azure tend to abstain from disclosing information about the inner workings of their models.

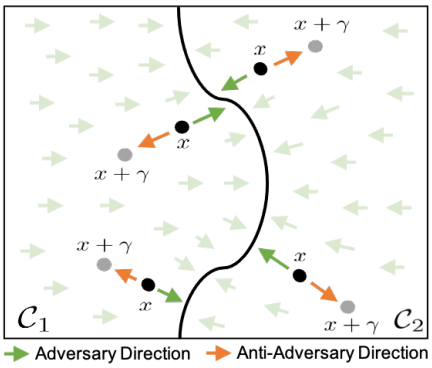

Figure 1: **Anti-adversary classifier.** The flow field of adversarial perturbations is shown in light green for both classes $\mathcal{C}_1$ and $\mathcal{C}_2$. The anti-adversary we construct pulls a given point x to $(x + \gamma)$ by moving in the direction *opposite* to that of the adversary flow field (orange arrows).

In this work, we propose a simple, generic, and training-free layer that improves the robustness of both nominally- and robustly-trained models. Specifically, given a base classifier $f : \mathbb{R}^n \to \mathcal{Y}$, which maps $\mathbb{R}^n$ to labels in the set $\mathcal{Y}$, and an input x, our layer constructs a data- and model-dependent perturbation γ in the *anti-adversary* direction, *i.e.* the direction that maximizes the base classifier's confidence on the pseudo-label $f(x)$, as illustrated in Figure 1. The new sample $(x+\gamma)$ is then fed to the base classifier f in lieu of x. We

[1]Official code: https://github.com/MotasemAlfarra/Combating-Adversaries-with-Anti-Adversaries

dub this complete approach as the *anti-adversary* classifier g. By conducting an extensive robustness assessment of our classifier g on several datasets and under the full spectrum of attacks, from black-box –arguably the most realistic– and white-box, to adaptive attacks, we find across-the-board improvements in robustness over all base classifiers f.

Contributions. (**i**) We propose an anti-adversary layer to improve the adversarial robustness of base classifiers. Our proposed layer comes at marginal computational overhead and virtually no impact on clean accuracy. Moreover, we provide theoretical insights into the robustness enhancement that our layer delivers. (**ii**) We demonstrate empirically under black-box attacks that our layer positively interacts with both nominally trained and state-of-the-art robust models, *e.g.* TRADES (Zhang et al. 2019), ImageNet-Pre (Hendrycks, Lee, and Mazeika 2019), MART (Wang et al. 2019), HYDRA (Sehwag et al. 2020), and AWP (Wu, Xia, and Wang 2020), on CIFAR10, CIFAR100 (Krizhevsky and Hinton 2009) and ImageNet (Krizhevsky, Sutskever, and Hinton 2012). Our results show that the anti-adversary layer not only improves robustness against a variety of black-box attacks (Ilyas, Engstrom, and Madry 2019; Ilyas et al. 2018; Andriushchenko et al. 2020), but also that this improvement comes at no cost on clean accuracy and does not require retraining. (**iii**) We further evaluate our approach on a challenging setting, in which the attacker is granted full access to the anti-adversary classifier, *i.e.* white-box attacks. Under this setup, we equip the five aforementioned defenses with our classifier and test them under the strong AutoAttack benchmark (Croce and Hein 2020b). Our experiments report across-the-board average improvements of 19% and 11% on CIFAR10 and CIFAR100, respectively.

Related Work

Adversarial Attacks. Evaluating network robustness dates back to the works of (Szegedy et al. 2013; Goodfellow, Shlens, and Szegedy 2015), where it was shown that small input perturbations, dubbed as adversarial attacks, can change network predictions. Follow-up methods present a variety of ways to construct such attacks, which are generally categorized as black-box, white-box and adaptive attacks. Black-box attackers either carry out zeroth order optimization to maximize a suitably defined loss function (Guo et al. 2019; Uesato et al. 2018), or learn offline adversaries that transfer well across networks (Papernot et al. 2017; Bhagoji et al. 2018). On the other hand, and less practical, white-box attackers are assumed to have the full knowledge of the network, *e.g.* parameters, gradients, architecture, and training data, among others (Moosavi-Dezfooli, Fawzi, and Frossard 2016; Madry et al. 2018). Despite that, previously proposed attackers from this family often construct adversaries solely based on network predictions and gradients with respect to the input (Carlini and Wagner 2017; Croce and Hein 2020a). Although this results in an overestimation of the worst-case robustness for networks, it has now become the *de facto* standard for benchmarking robustness (Croce and Hein 2020b). It was recently demonstrated that several networks, which were shown to be robust in the white-box setting, were susceptible to weaker black-box attacks (Dong et al. 2020). Consequently, there has been significant interest for *adaptive attacks*, *i.e.* specifically tailored adversaries exploiting complete knowledge of the network (not only predictions and gradients), for a reliable worst-case robustness assessment (Tramer et al. 2020; Athalye, Carlini, and Wagner 2018). However, while worst-case robustness is of interest through adaptive attacks, it may not be of practical relevance. We argue that a proper robustness evaluation should cover the full spectrum of attackers from black-box to adaptive attacks; thus, in this paper, we evaluate our method over such spectrum: black-box, white-box, and adaptive attacks. In particular, we use Bandits (Ilyas, Engstrom, and Madry 2019), NES (Ilyas et al. 2018) and Square (Andriushchenko et al. 2020) for black-box attacks, AutoAttack (Croce and Hein 2020b) which ensembles the APGD, ADLR, FAB (Croce and Hein 2020a), and Square attacks for the white-box evaluation, and tailor an adaptive attack specific to our proposed approach for a worst-case robustness evaluation.

Defenses Against Adversaries. Given the security concerns that adversarial vulnerability brings, a stream of works developed models that are not only accurate but also robust against adversarial attacks. From the black-box perspective, several defenses have shown their effectiveness in defending against such attacks (Rakin, He, and Fan 2018). For example, injecting Gaussian noise into activation maps during both training and testing (Liu et al. 2017) was shown to successfully defend against a variety of black-box attacks (Dong et al. 2020). Moreover, SND (Byun, Go, and Kim 2021) showed that small input perturbations can enhance the robustness of pretrained models against black-box attacks. However, the main drawback of randomized methods is that they can be bypassed by Expectation Over Transformation (EOT) (Athalye et al. 2018). Once an attacker accesses the gradients, *i.e.* white-box attackers, the robust accuracy of such defenses drastically decreases. Thus, a stream of works built models that resist white-box attacks. While several approaches were proposed, such as regularization (Cisse et al. 2017) and distillation (Papernot et al. 2016b), Adversarial Training (AT) (Madry et al. 2018) remains among the most effective. Moreover, recent works showed that AT can be enhanced by combining it with pretraining (Hendrycks, Lee, and Mazeika 2019), exploiting unlabeled data (Carmon et al. 2019), or concurrently, conducting transformations at test time (Pérez et al. 2021). Further improvements were obtained by introducing regularizers, such as TRADES (Zhang et al. 2019) and MART (Wang et al. 2019), or combining AT with network pruning, as in HYDRA (Sehwag et al. 2020), or weight perturbations (Wu, Xia, and Wang 2020). While these methods improve the robustness, they require expensive training and degrade clean accuracy. In this work, we show how our proposed anti-adversary layer enhances the performance of nominally trained models against realistic black-box attacks and even outperforms the strong SND defense. We show that equipping robust models with our anti-adversary layer significantly improves their robustness against black- and white-box attacks, in addition to showing worst-case robustness improvements under adaptive attacks.

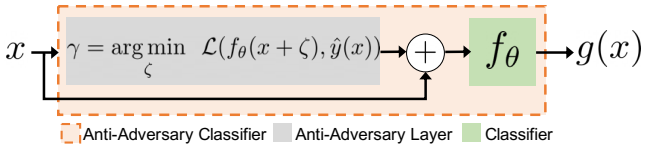

Figure 2: **The Anti-Adversary classifier.** Our anti-adversary layer generates γ for each x and f_θ, and feeds $(x+\gamma)$ to f_θ, resulting in our anti-adversary classifier g.

Methodology

Motivation. Adversarial directions are the ones that maximize a loss function in the input, *i.e.* move an input x closer to the decision boundary, resulting in minimizing the prediction's confidence on the correct label. In this work, we leverage this fact by prepending a layer to a trained model to generate a new input $(x+\gamma)$, which moves x far from the decision boundary, thus hindering the capacity of attackers to successfully tailor adversaries. Before detailing our approach, we start with preliminaries and notations.

Preliminaries and Notation

We use $f_\theta : \mathbb{R}^n \to \mathcal{P}(\mathcal{Y})$ to denote a classifier, *e.g.* a neural network, parameterized by θ, where $\mathcal{P}(\mathcal{Y})$ refers to the probability simplex over the set $\mathcal{Y} = \{1, 2, \ldots, k\}$ of k labels. For an input x, an attacker constructs a small perturbation δ (*e.g.* $\|\delta\|_p \leq \epsilon$) such that $\arg\max_i f_\theta^i(x+\delta) \neq y$, where y is the true label for x. In particular, one popular approach to constructing δ is by solving the following constrained problem with a suitable loss function $\mathcal{L}$:

$$\max_{\delta} \ \mathcal{L}(f_\theta(x+\delta), y) \qquad \text{s.t. } \|\delta\|_p \leq \epsilon. \tag{1}$$

Depending on the information about f_θ given to the attacker when solving Problem (1), the adversary δ can generally be categorized into one of three types. (**i**) **Black-box:** Only function evaluations f_θ are available when solving (1). (**ii**) **White-box:** Full access of the classifier f_θ, e.g. $\nabla_x f_\theta$, is granted when solving (1). (**iii**) **Adaptive:** The attacker is tailored specifically to break the classifier f_θ. That is to say, unlike white-box attacks that can be generic methods for all defenses, adaptive attacks are handcrafted to break specific defenses with full knowledge of f_θ, providing a better assessment of worst-case robustness for the classifier f_θ.

Anti-Adversary Layer

Analogous to the procedure used for constructing an adversary by solving (1), we propose, given a classifier, to prepend a layer that perturbs its input so as to maximize the classifier's prediction confidence at this input, hence the term *anti-adversary*. Formally, given a classifier f_θ, our proposed anti-adversary classifier g (prepending f_θ with an anti-adversary layer) is given as follows:

$$\begin{aligned} &g(x) = f_\theta(x+\gamma), \\ &\text{s.t. } \gamma = \arg\min_{\zeta} \ \mathcal{L}(f_\theta(x+\zeta), \hat{y}(x)), \end{aligned} \tag{2}$$

where $\hat{y}(x) = \arg\max_i f_\theta^i(x)$ is the predicted label. Note that our proposed anti-adversary classifier g is agnostic to the choice of f_θ. Moreover, it does not require retraining f_θ, unlike previous works (Xie et al. 2018; Byun, Go, and Kim 2021) that add random perturbations to the input, further hurting clean accuracy. This is because instances that are correctly classified by f_θ, *i.e.* instances where $y = \arg\max_i f_\theta^i(x)$, will be (by construction as per optimization (2)) classified correctly by g. As such, our anti-adversary layer only increases the confidence of the top prediction of $f_\theta(x)$. Also, we observe that our novel layer aligns with the recent advances in deep declarative models (Gould, Hartley, and Campbell 2019; Amos and Kolter 2017; Chen et al. 2018; Bibi et al. 2019), where the output activations of a given layer are solutions to optimization problems or differential equations. We illustrate our approach in Figure 2.

Theoretical Motivation for Robustness

Since the anti-adversary classifier g perturbs inputs towards locations far from decision boundaries, we argue that g can theoretically enjoy better robustness than f_θ. In particular, we study robustness under the realistic black-box adversary setting of solving the unconstrained version of Problem (1). We analyze the robustness of both g and f_θ under the celebrated SimBA attack (Guo et al. 2019) due to its simplicity and popularity. We show that SimBA requires a larger number of queries (forward passes) to fool g than to fool f_θ, *i.e.* g is more robust than f_θ. First, we show an equivalency between SimBA and Stochastic Three Points (STP) (Bergou, Gorbunov, and Richtárik 2020), a recently proposed derivative-free optimization algorithm. All proofs are left for the **Appendix**.

Proposition 1. *Let SimBA (Guo et al. 2019) with a budget of $2B$ queries select a random direction $q \in Q$, with replacement, thus updating the iterates $x^{k+1} \leftarrow x^k$ by selecting the direction among $\{\epsilon q, -\epsilon q\}$[2] with the maximum $\mathcal{L}$. Then, SimBA is equivalent to STP with B iterations.*

Therefore, when $\mathcal{L}$ is L-smooth, *i.e.* $\|\nabla_x \mathcal{L}(f_\theta(x+\delta), y) - \nabla_x \mathcal{L}(f_\theta(x), y)\| \leq L\|\delta\|$, we can find a lower bound for the number of queries B required by SimBA to maximize $\mathcal{L}$ to a certain precision.

Corollary 1. *Let $\mathcal{L}$ be L-smooth, bounded above by $\mathcal{L}(f_\theta(x^*), y)$, and the steps of SimBA satisfy $0 < \epsilon < \rho/nL$ while sampling directions from the Cartesian canonical basis (Q is an identity matrix here). Then, so long as:*

$$B > \frac{\mathcal{L}(f_\theta(x^*), y) - \mathcal{L}(f_\theta(x^0), y)}{(\frac{\rho}{n} - \frac{L}{2}\epsilon)\epsilon} = K_{f_\theta},$$

we have that $\min_{k=1,2,\ldots,B} \mathbb{E}\left[\|\nabla \mathcal{L}(f_\theta(x^k), y)\|_1\right] < \rho$.

Corollary 1 quantifies the minimum query budget K_{f_θ} required by SimBA to maximize $\mathcal{L}(f_\theta(x), y)$, reaching a specific solution precision ρ measured in gradient norm. Note that SimBA requires 2 queries (evaluating f_θ at $x^k \pm \epsilon q$) before sampling a new direction q from Q; thus, with a budget

[2]Dropping the conditional break for loop, which is originally introduced in SimBA for computational reasons, in Algorithm (1) in (Guo et al. 2019) and evaluating on both $\pm\epsilon q$.

Algorithm 1: Anti-adversary classifier g

Function `AntiAdversaryForward(`f_θ, x, α, K`)`**:**

Initialize: $\gamma^0 = 0$

$\hat{y}(x) = \arg\max_i f^i_\theta(x)$

for $k = 0 \dots K-1$ **do**

$\gamma^{k+1} = \gamma^k - \alpha \operatorname{sign}(\nabla_{\gamma^k} \mathcal{L}(f_\theta(x+\gamma^k), \hat{y}))$

end

return $f_\theta(x+\gamma^K)$

of $2B$, SimBA performs a total of B new updates to x^k with iterates ranging from $k=1$ to $k=B$. To compare the robustness of f_θ to our anti-adversary classifier g described in Eq. (2), we derive K_g, *i.e.* the minimum query budget necessary for SimBA to achieve a similar gradient norm precision ρ when maximizing $\mathcal{L}(g(x), y)$. For ease of purposes, we analyze K_g when the anti-adversary layer in g solves the minimization Problem (2) with one iteration of STP with learning rate ϵ_g. Next, we show that SimBA requires a larger query budget to maximize $\mathcal{L}(g(x), y)$ as opposed to $\mathcal{L}(f_\theta(x), y)$, hence implying that g enjoys improved robustness.

Theorem 1. *Let the assumptions in Proposition 1 and Corollary 1 hold. Then, the anti-adversary classifier g described in Eq.* (2)*, where γ is computed with a single STP update in the same direction q as SimBA but with a learning rate $\epsilon_g = (1-c)\epsilon$ with $c < 1$, is more robust against SimBA attacks than f_θ. In particular, $\forall\ c \leq 0$, SimBA fails to construct adversaries for g (*i.e. $K_g = \infty$*). Moreover, for $c \in (0,1)$, the improved robustness factor for g is:*

$$G(c) := \frac{K_g}{K_{f_\theta}} = \frac{\frac{\rho}{n} - \frac{L\epsilon}{2}}{(\frac{\rho}{n} - \frac{L\epsilon}{2}c)c} > 1. \tag{3}$$

Theorem 1 demonstrates that for any choice of $c < 1$, and under certain assumptions, g is more robust than f_θ under SimBA attacks. In the case where the anti-adversary layer employs a larger learning rate ϵ_g than that of SimBA (ϵ), *i.e.* $c \leq 0$, then SimBA attacks will never alter the prediction of g, since $K_g = \infty$. On the other hand, when ϵ_g (the learning rate of the anti-adversary) is smaller than the learning rate of SimBA, *i.e.* $c \in (0,1)$, SimBA will be successful in altering the prediction of g but with a larger number of queries compared to f_θ, that is, g is more robust than f_θ under SimBA attacks. This outcome is captured by the improved robustness factor G, which is a strictly decreasing function in $c \in (0,1)$ and lower bounded by 1.

In general, we hypothesize that the stronger the anti-adversary layer solver for Problem (2) is, the more robust g is against all attacks (including white-box and particularly against black-box attacks)[3]. To that end, and throughout the paper, the anti-adversary layer solves Problem (2) with K signed gradient descent iterations, zero initialization, and $\mathcal{L}$ being the cross-entropy loss. Algorithm 1 summarizes the forward pass of g. Next, we empirically validate improvements in robustness over the full spectrum of adversaries.

[3]We leave to the **Appendix** a version of Theorem 1, where we derive the improved robustness factor under the white-box setting with the anti-adversary layer solving Eq. (2) using gradient descent.

Experiments

Evaluating robustness is an elusive problem, as it is ill-defined without establishing the information available to the attacker (1) for constructing the adversary δ. Prior works usually evaluate robustness under the adaptive, black-box or white-box settings. Here, we argue that robustness should be evaluated over the *complete* spectrum of adversaries. In particular, we underscore that, while adaptive attacks can provide a worst-case robustness assessment, such assessment may be uninteresting for real deployments. For example, when the worst-case robustness of classifiers results in a draw, this tie can be broken by considering their robustness in the black-box setting, as this property increases its desirability for real-world deployment.

Thus, we validate the effectiveness of our proposed anti-adversary classifier g by evaluating robustness under adversaries from the full spectrum. (**i**) We first compare the robustness of f_θ against our proposed anti-adversary classifier g with popular black-box attacks (Bandits (Ilyas, Engstrom, and Madry 2019), NES (Ilyas et al. 2018) and Square (Andriushchenko et al. 2020)). We consider both cases when f_θ is nominally and robustly trained. Not only do we observe significant robustness improvements over f_θ with virtually no drop in clean accuracy, but we also outperform recently-proposed defenses, such as SND (Byun, Go, and Kim 2021). (**ii**) We further conduct experiments in the more challenging white-box setting with AutoAttack (Croce and Hein 2020b) (in particular against the strong attacks APGD, ADLR (Croce and Hein 2020b), and FAB (Croce and Hein 2020a)), when f_θ is trained robustly with TRADES (Zhang et al. 2019), ImageNet-Pre (Hendrycks, Lee, and Mazeika 2019), MART (Wang et al. 2019), HYDRA (Sehwag et al. 2020), and AWP (Wu, Xia, and Wang 2020). (**iii**) We analyze robustness performance under tailored adaptive attacks, demonstrating that the worst-case performance is lower bounded by the robustness of f_θ. In all experiments, we do *not* retrain f_θ after prepending our anti-adversary layer. We set $K = 2$ and $\alpha = 0.15$ whenever Algorithm 1 is used, unless stated otherwise. (**iv**) Finally, we ablate the effect of the learning rate α and the number of iterations K on the robustness gains.

Robustness under Black-Box Attacks

We start by studying how prepending our proposed anti-adversary layer to a classifier f_θ can induce robustness gains against black-box attacks. This is a realistic setting as several commercially-available APIs, *e.g.* BigML, only allow access to model predictions, and thus, they can only be targeted with black-box adversaries.

Robustness when f_θ is Nominally Trained. We conduct experiments with ResNet18 (He et al. 2016) on CIFAR10 (Krizhevsky and Hinton 2009) and ResNet50 on ImageNet (Deng et al. 2009). We compare our anti-adversary classifier g against f_θ in terms of clean and robust test accuracy when subjected to two black-box attacks. In particular, we use the Bandits and NES attacks with query budgets of $5k$ and $10k$, and report results in Table 1. In addition, we compare against a recently proposed approach for robustness through input

Table 1: **Robustness of nominally trained models against black-box attacks:** We present the robustness of a nominally trained model against Bandits and NES, and how robustness is enhanced when equipping the model with SND (Byun, Go, and Kim 2021) and our anti-adversary layer (Anti-Adv). We perform all attacks with both $5k$ and $10k$ queries. Results shown are accuracy measured in $\%$ where bold numbers correspond to best performance. Our approach outperforms SND by a significant margin across datasets, attacks, and number of queries.

	CIFAR10					ImageNet				
	Clean	Bandits		NES		Clean	Bandits		NES	
		5K	10K	5K	10K		5K	10K	5K	10K
Nominal Training	93.7	24.0	17.2	5.8	4.8	79.2	65.2	58.2	22.4	21.0
+ SND (Byun, Go, and Kim 2021)	92.9	84.5	84.3	30.3	25.5	79.2	72.8	73.2	65.4	60.2
+ Anti-Adv	93.7	**85.5**	**86.4**	**77.0**	**72.7**	79.2	**73.6**	**74.4**	**67.2**	**66.0**

Table 2: **Equipping robustly trained models with Anti-Adv on CIFAR10 and CIFAR100 against black-box attacks.** We report clean accuracy (%) and robust accuracy against *Bandits*, *NES* and *Square attack* where bold numbers correspond to largest accuracy in each experiment. Our layer provides across the board improvements on robustness against all attacks without affecting clean accuracy.

CIFAR10	Clean	Bandits	NES	Square
TRADES	85.4	64.7	74.7	53.1
+ Anti-Adv	85.4	**84.6**	**83.0**	**71.7**
ImageNet-Pre	88.7	68.4	78.1	62.4
+ Anti-Adv	88.7	**88.1**	**86.4**	**78.5**
MART	87.6	72.0	79.5	64.9
+ Anti-Adv	87.6	**86.5**	**85.3**	**78.0**
HYDRA	90.1	69.8	79.2	65.0
+ Anti-Adv	90.1	**89.4**	**87.7**	**78.8**
AWP	88.5	71.5	80.1	66.2
+ Anti-Adv	88.5	**87.4**	**86.9**	**80.7**
CIFAR100	Clean	Bandits	NES	Square
ImageNet-Pre	59.0	40.6	47.7	34.6
+ Anti-Adv	58.9	**58.2**	**55.3**	**42.4**
AWP	59.4	39.8	47.3	34.7
+ Anti-Adv	59.4	**57.7**	**53.8**	**46.4**

randomization (SND (Byun, Go, and Kim 2021)). We set $\sigma = 0.01$ for SND, as it achieves the best performance. Following common practice (Byun, Go, and Kim 2021), and due to the expensive nature of evaluating Bandits and NES, all test accuracy results in Table 1 are reported on 1000 and 500 instances of CIFAR10 and ImageNet, respectively. For this experiment, we set $\alpha = 0.01$ in Algorithm 1. Note that SND, the closest work to ours, outperforms the best performing defense in the black-box settings benchmarked in (Dong et al. 2020).

As shown in Table 1, nominally trained models f_θ are not robust: their clean accuracies on CIFAR10 and ImageNet drop from 93.7% and 79.2%, respectively, to 4.8% and 21% when under black-box attacks. Moreover, while SND improves robustness significantly over f_θ, *e.g.* to 25.5% on CIFAR10 and to 60.2% on ImageNet, our proposed anti-adversary consistently outperforms SND across attacks, budget queries, and datasets. For instance, under the limited $5k$ query budget, our anti-adversary classifier outperforms SND by 1% and 46.7% on CIFAR10 against Bandits and NES. The robustness improvements over SND increase even when attacks have a larger budget of $10k$: on ImageNet our anti-adversary outperforms SND by 1.2% against Bandits and by 5.8% against NES. Further, we note that this improvement comes at *no cost* on clean accuracy. In summary, Table 1 provides strong evidence suggesting that our proposed anti-adversary classifier improves the black-box robustness of a nominally trained f_θ, outperforming the recent SND. In addition, this performance improvement does not hurt clean accuracy nor requires retraining f_θ.

Robustness when f_θ is Robustly Trained. We have provided evidence that our anti-adversary layer can improve black-box robustness of nominally trained f_θ. Here, we investigate whether our anti-adversary layer can also improve robustness in the more challenging setting when f_θ is already robustly trained. This is an interesting setup as f_θ could have been trained robustly against white-box attacks and then deployed in practice where only function evaluations are available to the attacker (1), and hence only black-box robustness is of importance. Here we show we can improve black-box robustness with our proposed anti-adversary layer over five state-of-the-art robustly trained f_θ: TRADES, IN-Pret, MART, HYDRA, and AWP on the CIFAR10 and CIFAR100 datasets. Similar to the previous experimental setup, and due to computational cost, we report robust accuracy on 1000 test set instances against Bandits and NES. However, for the more computationally-efficient Square attack, we report robust accuracy on the full test set.

Table 2 reports the black-box robust accuracies of robustly-trained f_θ on CIFAR10 and CIFAR100, respectively. We highlight the highest scores in bold. In line with our previous observations, prepending our anti-adversary layer to f_θ has no impact on clean accuracy. More importantly, although f_θ is robustly trained and thus already enjoys large black-box robust accuracy, our proposed anti-adversary layer can boost its robustness further by an impressive $\sim 15\%$. For instance, even for the top-performing

Table 3: **Equipping robustly trained models with Anti-Adv on CIFAR10 and CIFAR100 against white-box attacks.** We report clean accuracy (%) and robust accuracy against *APGD*, *ADLR*, *FAB* and *AutoAttack* where bold numbers correspond to largest accuracy in each experiment. The last column summarizes the improvement on the AutoAttack benchmark. We observe strong results on all models and attacks when adding our anti-adversary layer, with improvements close to 19% all around.

CIFAR10	Clean	APGD	ADLR	FAB	AutoAttack	Improvement
TRADES	84.92	55.31	53.12	53.55	53.11	18.60
+ Anti-Adv	84.88	**77.20**	**77.05**	**83.38**	**71.71**	
ImageNet-Pre	87.11	57.65	55.32	55.69	55.31	20.70
+ Anti-Adv	87.11	**78.76**	**79.02**	**85.07**	**76.01**	
MART	87.50	62.18	56.80	57.34	56.75	20.01
+ Anti-Adv	87.50	**81.07**	**80.54**	**86.52**	**76.76**	
HYDRA	88.98	60.13	57.66	58.42	57.64	18.75
+ Anti-Adv	88.95	**80.37**	**81.42**	**87.92**	**76.39**	
AWP	88.25	63.81	60.53	60.98	60.53	18.68
+ Anti-Adv	88.25	**80.65**	**81.47**	**87.06**	**79.21**	
CIFAR100	Clean	APGD	ADLR	FAB	AutoAttack	Improvement
ImageNet-Pre	59.37	33.45	29.03	29.34	28.96	11.72
+ Anti-Adv	58.42	**47.63**	**45.29**	**53.57**	**40.68**	
AWP	60.38	33.56	29.16	29.48	29.15	10.42
+ Anti-Adv	60.38	**44.21**	**40.32**	**50.76**	**39.57**	

f_θ (trained with AWP with a robust accuracy of 66.2% on CIFAR10), our anti-adversary layer improves robustness by 14.5%, reaching 80.7%. Similarly, for CIFAR100, the anti-adversary layer improves the worst-case black-box robustness of AWP by 11.7%. Overall, our anti-adversary layer consistently improves black-box robust accuracy against all attacks for all robust training methods on both CIFAR10 and CIFAR100.

SND + Robustly Trained f_θ. Although SND (Byun, Go, and Kim 2021) does not report performance on robustly-trained models, we experiment with equipping AWP-trained models with SND. We observe that SND significantly degrades both clean and robust accuracies of AWP: employing SND on top of AWP drops clean accuracy from 88.5% to 70.0%, while its robust accuracy (under Square attack) drops from 66.2% to 59.1%. These results suggest that our proposed anti-adversary layer is superior to SND.

Other Black-Box Defenses We compare against Random Self-Ensemble (RSE) (Liu et al. 2017) on CIFAR10 and find that it underperforms in comparison to our approach, both in clean accuracy, with 86.7%, and in robust accuracy, with 78.8% and 85.5% under NES and Bandits, respectively. While RSE is more robust than robustly trained models against black-box attacks, equipping such models with our anti-adversary layer outperforms RSE, as illustrated, for instance, by the HYDRA+Anti-Adv row in Table 2.

Section Summary. Our proposed anti-adversary layer can improve state-of-the-art robust accuracy in the realistic black-box setting when combined with robustly trained f_θ, while coming at no cost to clean accuracy. The robust black-box accuracy improvements are consistent across classifiers f_θ, both with regular or robust training.

Robustness under White-Box Attacks

In this setting, the attacker (1) has complete knowledge about the classifier. This challenging setup is less realistic compared to the black-box setting. Nonetheless, it is still an interesting measure of overall robustness when more information is accessible to the attacker (1). Various prior works (Xie et al. 2019; Zhang and Wang 2019) report robustness performance only in this setting by reporting accuracy under PGD (Madry et al. 2018) or AutoAttack.

Similar to the previous section, we experiment on CIFAR10 and CIFAR100 and assess how prepending our anti-adversary layer to robustly trained classifiers f_θ can enhance the classifiers' robustness. We report the full test robust white-box accuracy against the gradient-based attacks from AutoAttack, *i.e.* APGD, ADLR and FAB, and also measure the accuracy under AutoAttack, defined as the worst-case accuracy across these four attacks (three white-box attacks in addition to Square attacks) under $\epsilon = {}^8\!/_{255}$ in (1). We underscore that the AutoAttack ensemble is currently the standard for benchmarking defenses, *i.e.* it is the *de facto* strongest attack in this setting.

In Table 3, we report robust accuracies on CIFAR10 and CIFAR100, respectively, and highlight the strongest performance in bold. We first observe that our anti-adversary layer improves robust accuracy by an impressive $\sim 19\%$ on average against AutoAttack. In particular, for AWP, the strongest defense we consider, adversarial robustness increases from 60.53% to an astounding 79.21%. We further observe similar results for CIFAR100: Table 3 shows that the anti-

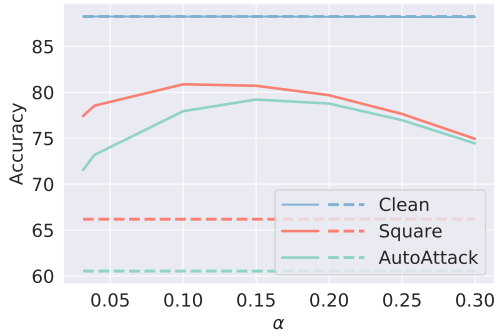

Figure 3: **Effect of varying α on clean and robust accuracy for AWP+Anti-Adv on CIFAR10.** Dashed lines depict AWP's performance. Our layer provides substantial improvements on robust accuracy with different choices of α and with no effect on clean accuracy.

adversary layer adds an average improvement of $\sim 11\%$. For instance, the adversarial robustness of ImageNet-Pre increases from 28.96% to over 40%. The improvement is consistent across all defenses on CIFAR100, with a worst-case drop in clean accuracy of 1%. We also compare our approach against SND under this setup (as the experiments in SND (Byun, Go, and Kim 2021) do not study its interaction with robust training). Notably, equipping AWP with SND comes at a notable drop in clean accuracy (from 88.25% to 70.03%) along with a drastic drop in robust accuracy (from 60.53% to 27.04%) under AutoAttack on CIFAR10.

Section Summary. Our experiments suggest that, even in the challenging setting where the attacker (1) is granted access to the gradients, our anti-adversary layer still proves to provide benefits to all defenses. For both CIFAR10 and CIFAR100, the anti-adversary layer seamlessly provides vast improvements in adversarial robustness.

Adaptive Attacks: Worst-Case Performance

Here, we analyze the worst-case robustness of our proposed anti-adversary classifier g. In particular, and under the *least* realistic setting, we assume that our anti-adversary classifier g is fully transparent to the attacker (1) when tailoring an adversary. Following the recommendations in (Tramer et al. 2020), we explore various directions to construct an attack, such as Expectation Over Transformation (EOT) (Athalye et al. 2018; Tramer et al. 2020). However, since our anti-adversary layer is deterministic, as illustrated in Algorithm 1, EOT is ineffective for improving the gradient estimate. Nevertheless, we note that the anti-adversary layer depends on the pseudo-label assigned by f_θ to the original instance x, *i.e.* $\hat{y}(x) = \arg\max_i f^i_\theta(x)$. Therefore, an attacker with access to g's internal structure can first design an adversary δ such that $\hat{y}(x+\delta) \neq y$ with $\|\delta\|_p \leq \epsilon$ following (1), where y is x's label. If δ is successfully constructed in this way, it will cause both f_θ and g to produce different predictions for x and $(x+\delta)$. Thus, in the least realistic adversary setting, the set of adversaries that fools f_θ fools g as well. Accordingly, we argue that the worst-case robust accuracy for g under adaptive attacks is lower bounded by the robust accuracy of the base classifier f_θ. While, as noted in previous sections, our anti-adversary layer boosts robust accuracy over all tested datasets and classifiers f_θ (nominally or robustly trained), the worst-case robustness under the least realistic setting (adaptive attacks) is lower bounded by the robustness of f_θ. This highlights our motivation that prepending our layer is of a great value to existing robust models due to its simplicity and having no cost on clean accuracy.

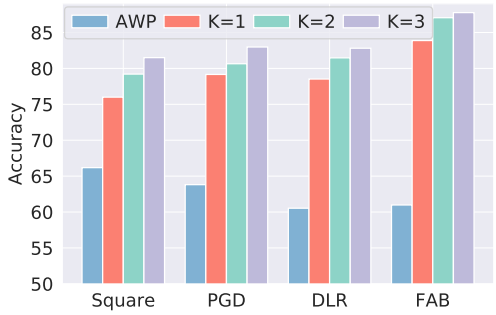

Figure 4: **Effect of varying K on robust accuracy for AWP+Anti-Adv on CIFAR10.** The better the solver for (2) is, the larger the robustness gains that our layer provides.

Ablations

Our proposed Algorithm 1 has two main parameters: the learning rate, α, and the number of iterations, K. We ablate both to assess their effect on robustness. All experiments are conducted on a robustly-trained f_θ (with AWP). First, we fix $K = 2$ and vary α in the set $\{{}^8\!/_{255}, {}^{10}\!/_{255}, 0.1, 0.15, 0.2, 0.25, 0.3\}$. In Figure 3, we compare f_θ to our anti-adversary classifier g in terms of clean and robust accuracies under a black-box (Square) and a white-box (AutoAttack) attacks. As shown in blue, the effect of α on clean accuracy is almost non-existent. On the other hand, while the robust accuracy varies with α, the robustness gain of g over f_θ is always $\geq 10\%$ for all α values. Next, we study the effect of varying $K \in \{1, 2, 3\}$ while fixing $\alpha = 0.15$. Results in Figure 4 show that all choices of K lead to significant improvement in robustness against all attacks, with $K = 3$ performing best. This confirms our claim that the better the solver for (2), the better the robustness performance of our anti-adversary classifier. Note that while one could further improve the robustness gains by increasing K, this improvement comes at the expense of more computations. It is worthwhile to mention that the cost of computing the anti-adversary is $(K+1)$ forward and K backward passes, which is marginal for small values of K. Finally, we leave more ablations, the implementation details, and the rest of our experimental results to the **Appendix**.

Conclusion

We present the anti-adversary layer, a novel training-free and theoretically supported defense against adversarial attacks. Our layer provides significant improvements in network robustness against black- and white-box attacks.

Acknowledgement. This publication is based upon work supported by the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research (OSR) under Award No. OSR-CRG2019-4033. We would also like to thank Humam Alwassel for the help and discussion.

References

Amos, B.; and Kolter, J. Z. 2017. OptNet: Differentiable Optimization as a Layer in Neural Networks. In *International Conference on Machine Learning (ICML)*.

Andriushchenko, M.; Croce, F.; Flammarion, N.; and Hein, M. 2020. Square Attack: a query-efficient black-box adversarial attack via random search. In *European Conference on Computer Vision (ECCV)*.

Athalye, A.; Carlini, N.; and Wagner, D. 2018. Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples. In *International Conference on Machine Learning (ICML)*.

Athalye, A.; Engstrom, L.; Ilyas, A.; and Kwok, K. 2018. Synthesizing Robust Adversarial Examples. In *International Conference on Machine Learning (ICML)*.

Bergou, E. H.; Gorbunov, E.; and Richtárik, P. 2020. Stochastic Three Points Method for Unconstrained Smooth Minimization. In *SIAM Journal on Optimization*.

Bhagoji, A. N.; He, W.; Li, B.; and Song, D. 2018. Practical Black-box Attacks on Deep Neural Networks using Efficient Query Mechanisms. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Bibi, A.; Ghanem, B.; Koltun, V.; and Ranftl, R. 2019. Deep Layers as Stochastic Solvers. *International Conference on Learning Representations (ICLR)*.

Brendel, W.; Rauber, J.; and Bethge, M. 2018. Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models. In *International Conference on Learning Representations (ICLR)*.

Byun, J.; Go, H.; and Kim, C. 2021. Small Input Noise is Enough to Defend Against Query-based Black-box Attacks. *https://openreview.net/forum?id=6HlaJSlQFEj*.

Carlini, N.; Athalye, A.; Papernot, N.; Brendel, W.; Rauber, J.; Tsipras, D.; Goodfellow, I.; Madry, A.; and Kurakin, A. 2019. On Evaluating Adversarial Robustness. *arXiv:1902.06705*.

Carlini, N.; and Wagner, D. 2017. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*.

Carmon, Y.; Raghunathan, A.; Schmidt, L.; Duchi, J. C.; and Liang, P. S. 2019. Unlabeled data improves adversarial robustness. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Chen, R. T. Q.; Rubanova, Y.; Bettencourt, J.; and Duvenaud, D. 2018. Neural Ordinary Differential Equations. In *Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates Inc.

Cisse, M.; Bojanowski, P.; Grave, E.; Dauphin, Y.; and Usunier, N. 2017. Parseval networks: Improving robustness to adversarial examples. *International Conference on Machine Learning (ICML)*.

Croce, F.; and Hein, M. 2020a. Minimally distorted Adversarial Examples with a Fast Adaptive Boundary Attack. In *International conference on machine learning (ICML)*.

Croce, F.; and Hein, M. 2020b. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning (ICML)*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition (CVPR)*.

Dong, Y.; Fu, Q.-A.; Yang, X.; Pang, T.; Su, H.; Xiao, Z.; and Zhu, J. 2020. Benchmarking Adversarial Robustness on Image Classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and harnessing adversarial examples. *International Conference on Learning Representations (ICLR)*.

Gould, S.; Hartley, R.; and Campbell, D. 2019. Deep Declarative Networks: A New Hope. *CoRR*, abs/1909.04866.

Guo, C.; Gardner, J.; You, Y.; Wilson, A. G.; and Weinberger, K. 2019. Simple Black-box Adversarial Attacks. In *International Conference on Machine Learning (ICML)*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hendrycks, D.; Lee, K.; and Mazeika, M. 2019. Using Pre-Training Can Improve Model Robustness and Uncertainty. *International Conference on Machine Learning (ICML)*.

Ilyas, A.; Engstrom, L.; Athalye, A.; and Lin, J. 2018. Black-box adversarial attacks with limited queries and information. In *International Conference on Machine Learning. (ICML)*.

Ilyas, A.; Engstrom, L.; and Madry, A. 2019. Prior Convictions: Black-box Adversarial Attacks with Bandits and Priors. In *International Conference on Learning Representations (ICLR)*.

Krizhevsky, A.; and Hinton, G. 2009. Learning multiple layers of features from tiny images. In *University of Toronto, Canada*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Liu, X.; Cheng, M.; Zhang, H.; and Hsieh, C. 2017. Towards Robust Neural Networks via Random Self-ensemble. *CoRR*, abs/1712.00673.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representations (ICLR)*.

Moosavi-Dezfooli, S.-M.; Fawzi, A.; and Frossard, P. 2016. Deepfool: a simple and accurate method to fool deep neural networks. In *IEEE conference on computer vision and pattern recognition (CVPR)*.

Papernot, N.; McDaniel, P.; Goodfellow, I.; Jha, S.; Celik, Z. B.; and Swami, A. 2016a. Practical black-box attacks against deep learning systems using adversarial examples. *arXiv:1602.02697*.

Papernot, N.; McDaniel, P.; Goodfellow, I.; Jha, S.; Celik, Z. B.; and Swami, A. 2017. Practical Black-Box Attacks against Machine Learning. In *Asia Conference on Computer and Communications Security*.

Papernot, N.; McDaniel, P.; Wu, X.; Jha, S.; and Swami, A. 2016b. Distillation as a defense to adversarial perturbations against deep neural networks. In *IEEE Symposium on Security and Privacy (SP)*.

Pérez, J. C.; Alfarra, M.; Jeanneret, G.; Rueda, L.; Thabet, A.; Ghanem, B.; and Arbeláez, P. 2021. Enhancing Adversarial Robustness via Test-time Transformation Ensembling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Rakin, A. S.; He, Z.; and Fan, D. 2018. Parametric Noise Injection: Trainable Randomness to Improve Deep Neural Network Robustness against Adversarial Attack.

Sehwag, V.; Wang, S.; Mittal, P.; and Jana, S. 2020. HYDRA: Pruning Adversarially Robust Neural Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Sitawarin, C.; Bhagoji, A. N.; Mosenia, A.; Chiang, M.; and Mittal, P. 2018. DARTS: Deceiving Autonomous Cars with Toxic Signs. *arXiv:1802.06430*.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Tramer, F.; Carlini, N.; Brendel, W.; and Madry, A. 2020. On Adaptive Attacks to Adversarial Example Defenses. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Tsipras, D.; Santurkar, S.; Engstrom, L.; Turner, A.; and Madry, A. 2019. Robustness May Be at Odds with Accuracy. In *International Conference on Learning Representations (ICLR)*.

Uesato, J.; O'Donoghue, B.; Kohli, P.; and van den Oord, A. 2018. Adversarial Risk and the Dangers of Evaluating Against Weak Attacks. In *International Conference on Machine Learning (ICML)*.

Wang, Y.; Zou, D.; Yi, J.; Bailey, J.; Ma, X.; and Gu, Q. 2019. Improving adversarial robustness requires revisiting misclassified examples. In *International Conference on Learning Representations (ICLR)*.

Wu, D.; Xia, S.-T.; and Wang, Y. 2020. Adversarial Weight Perturbation Helps Robust Generalization. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Xie, C.; Wang, J.; Zhang, Z.; Ren, Z.; and Yuille, A. 2018. Mitigating Adversarial Effects Through Randomization. In *International Conference on Learning Representations (ICLR)*.

Xie, C.; Wu, Y.; Maaten, L. v. d.; Yuille, A. L.; and He, K. 2019. Feature denoising for improving adversarial robustness. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, H.; and Wang, J. 2019. Defense Against Adversarial Attacks Using Feature Scattering-based Adversarial Training. *CoRR*, abs/1907.10764.

Zhang, H.; Yu, Y.; Jiao, J.; Xing, E. P.; Ghaoui, L. E.; and Jordan, M. I. 2019. Theoretically Principled Trade-off between Robustness and Accuracy. In *International Conference on Machine Learning (ICML)*.