Combating Adversaries with Anti-Adversaries

Motasem Alfarra
KAUST

Juan C. Pérez
KAUST

Ali Thabet
Meta AI

Adel Bibi
University of Oxford

Philip H. S. Torr
University of Oxford

Bernard Gahnem
KAUST

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 anti-adversary 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 [11, 23]. While there has been interest in training DNNs that are robust to adversarial attacks, assessing a defense's robustness remains an elusive task. This difficulty is related to: (**i**) Model robustness varies with the information the attacker is assumed to know, *e.g*. training data, gradients, logits, *etc*.— dichotomously categorizing adversaries as black- or white-box. Consequently, this categorization imposes difficulties for comparing defenses tailored to specific types of adversaries. For instance, some defenses crafted for white-box attacks were later broken by black-box attacks [4, 20]. (**ii**) In addition, empirically-evaluated robustness can be overestimated if weaker efforts are invested in *adaptively* constructing attacks [6, 24]. The lack of reliable assessments is responsible for false senses of security, as presumably-strong defenses against white-box attacks were subsequently broken by carefully-crafted adaptive attacks [2]. The few defenses standing the test of time require costly training and degrade clean-sample performance [25]. Even worse, while most defenses are designed to resist white-box attacks, fewer efforts have been invested into resisting black-box attacks, which are more practical [5], *e.g*. public APIs (IBM Watson, Microsoft Azure, *etc*.) may not disclose information about their models' inner workings.

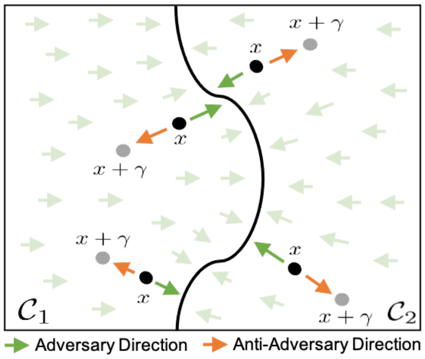

Figure 1. **Anti-adversary classifier.** The flow field of adversarial perturbations is shown in light green for both classes $\mathcal{C}_1$ and $\mathcal{C}_2$. The anti-adversary we construct pulls a given point x to $(x+\gamma)$ by moving in the direction *opposite* to that of the adversary flow.

In this work, we propose a simple, generic, training-free layer that robustifies both nominally and robustly trained DNNs. Given a base classifier $f : \mathbb{R}^n \to \mathcal{Y}$, which maps $\mathbb{R}^n$ to labels in 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 maximizing the classifier's confidence on the pseudo-label $f(x)$ (illustrated in Figure 1). The new sample $(x+\gamma)$ is then fed to f in lieu of x. We dub this approach the *anti-adversary* classifier g. We conduct an extensive robustness assessment of our layer on several datasets and under black-box, white-box, and adaptive attacks, and find across-the-board improvements in robustness over all base classifiers f.

1.1. Related Work

Given the security concerns that adversarial vulnerability brings, a stream of works built models that both accu-

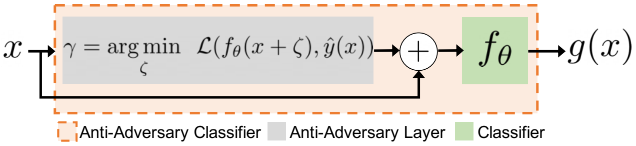
x
$\gamma = \arg\min_{\zeta} \; \mathcal{L}(f_\theta(x+\zeta), \hat{y}(x))$
$+$
f_θ
$g(x)$
Anti-Adversary Classifier
Anti-Adversary Layer
Classifier

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 into our anti-adversary classifier g).

rate and robust against attacks. Various black-box defenses have successfully worked against such attacks [10, 18, 21]. SND [5] showed that small input perturbations can enhance the robustness of pretrained models against black-box attacks. However, the main drawback of such randomized methods is that they can be bypassed via Expectation Over Transformation (EOT) [3]. Once an attacker accesses gradients, *i.e.* white-box attackers, the robust accuracy of such defenses drastically decreases. Adversarial training (AT) [19] and its enhanced versions [7, 13] remain the most effective defenses. This was further improved by incorporating additional regularizers such as TRADES [30] and MART [26], or combining AT with pruning as in HYDRA [22], or perturbing network parameters [27]

2. Methodology

Motivation. Adversary directions maximize a loss function w.r.t. the input, *i.e.* move input x closer to the decision boundary, thereby minimizing the confidence on the predicted label. In this work, we leverage this and prepend a layer to a trained model to generate a new input $(x+\gamma)$, which moves x away from the decision boundary, thus hindering the capacity of adversaries to tailor attacks.

Preliminaries and Notation. We use $f_\theta : \mathbb{R}^n \rightarrow \mathcal{P}(\mathcal{Y})$ to denote a classifier, *e.g.* a DNN parameterized by θ, where $\mathcal{P}(\mathcal{Y})$ is the probability simplex over the set $\mathcal{Y} = \{1, 2, \ldots, k\}$ of k labels. For 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 x's true label. One approach to construct δ is by solving the following constrained problem with a 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 accessible by the attacker when solving (1), the adversary can generally be categorized into one of three types. (i) **Black-box:** Only function evaluations f_θ are available. (ii) **White-box:** Only f_θ and $\nabla_x f_\theta$ are accessible, with no other intermediate layer information available to the attacker. (iii) **Adaptive:** The attacker has full knowledge about the classifier f_θ, including the θ, intermediate-layer gradients, training data, *etc.*

Anti-Adversary Layer. Analogous to constructing an adversary by solving (1), we propose to prepend a layer that perturbs the input to maximize the classifier's prediction's confidence at an input, hence the term *anti-adversary*. Formally, given classifier f_θ, our anti-adversary classifier g (prepending an anti-adversary layer to f_θ) is:

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

where $\hat{y}(x) = \arg\max_i f_\theta^i(x)$ is the predicted label. Note our proposed anti-adversary g is agnostic to the choice of f_θ. Moreover, it does not require retraining f_θ, unlike previous works [5, 28] that add random input perturbations, hurting clean accuracy. This is because correctly-classified instances, *i.e.* instances where $y = \arg\max_i f_\theta^i(x)$, will also be, by construction, classified correctly by g. Thus, our layer only increases the confidence of $f_\theta(x)$'s top prediction. We illustrate our approach in Figure 2. Finally, our layer solves Problem (2) with K signed gradient descent iterations, zero initialization, where $\mathcal{L}$ is the cross-entropy loss. Algorithm 1 summarizes g's forward pass.

Algorithm 1 Anti-adversary classifier g

```
Function AntiAdversaryForward($f_\theta$, $x$, $\alpha$, $K$):
    Initialize: $\gamma^0 = 0$
    $\hat{y}(x) = \arg\max_i f_\theta^i(x)$
    for $k = 0 \ldots K-1$ do
        $\gamma^{k+1} = \gamma^k - \alpha\,\text{sign}(\nabla_{\gamma^k}\mathcal{L}(f_\theta(x+\gamma^k), \hat{y}))$
    end
    return $f_\theta(x+\gamma^K)$
```

Table 1. **Robustness of nominally-trained DNNs against black-box attacks.** We attack nominally-trained DNNs with Bandits and NES (both with $10k$ queries), and find robustness increments when introducing SND [5] and our anti-adversary layer (Anti-Adv). We report accuracy and embolden best performance.

	CIFAR10			ImageNet		
	Clean	Bandits	NES	Clean	Bandits	NES
Nominal Training	93.7	17.2	4.8	79.2	58.2	21.0
+ SND [5]	92.9	84.3	25.5	79.2	73.2	60.2
+ Anti-Adv	93.7	**86.4**	**72.7**	79.2	**74.4**	**66.0**

3. Experiments

We validate the effectiveness of our anti-adversary classifier g by evaluating robustness against various adversaries. **(i)** We compare f_θ's robustness with our anti-adversary classifier g against black-box attacks (Bandits [15], NES [14] and Square [1]) both when f_θ is nominally and robustly trained. We observe significant robustness improvements over f_θ with virtually no drop in clean accuracy, while also outperforming the recently-proposed SND [5] defense. **(ii)** We experiment in the white-box setting with AutoAttack [9] (with APGD, ADLR [9], and FAB [8]), when f_θ is trained robustly with ImageNet-Pre [13] and AWP [27]. In all experiments, we do *not* retrain f_θ after prepending our layer.

Table 2. **Black-box attacks on robust models equipped with Anti-Adv.** We report clean (%) and robust accuracies against *Bandits*, *NES* and *Square attack* on CIFAR10 and CIFAR100. Bold values indicate highest accuracy in each experiment. Our layer provides across-the-board improvements on robustness against all attacks, without affecting clean accuracy.

		Clean	Bandits	NES	Square
CIFAR10	ImageNet-Pre	88.7	68.4	78.1	62.4
	+ Anti-Adv	88.7	**88.1**	**86.4**	**78.5**
	AWP	88.5	71.5	80.1	66.2
	+ Anti-Adv	88.5	**87.4**	**86.9**	**80.7**
CIFAR100	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**

Table 3. **White-box attacks on robust models equipped with Anti-Adv.** We report clean and robust accuracies against *APGD*, *ADLR*, and *AutoAttack* (AA) on CIFAR10 and CIFAR100.

		Clean	APGD	ADLR	AA
CIFAR10	ImageNet-Pre	87.11	57.65	55.32	55.31
	+ Anti-Adv	87.11	**78.76**	**79.02**	**76.01**
	AWP	88.25	63.81	60.53	60.53
	+ Anti-Adv	88.25	**80.65**	**81.47**	**79.21**
CIFAR100	ImageNet-Pre	59.37	33.45	29.03	28.96
	+ Anti-Adv	58.42	**47.63**	**45.29**	**40.68**
	AWP	60.38	33.56	29.16	29.15
	+ Anti-Adv	60.38	**44.21**	**40.32**	**39.57**

3.1. Robustness under Black-Box Attacks

We measure black-box robustness gains when prepending our layer to classifiers. This setting is interesting for commercial APIs that only provide model predictions and so can only be targeted with black-box adversaries.

Robustness of Nominally-Trained f_θ. We experiment with ResNet18 [12] on CIFAR10 [16] and ResNet50 on ImageNet [17]. We compare the clean and robust accuracies of f_θ against SND [5], and our anti-adversary classifier. We conduct $10k$ queries of two black-box attacks, Bandits and NES, and set $\alpha = 0.01$ in Algorithm 1. Following SND [5], we evaluate on 1000 and 500 instances of CIFAR10 and ImageNet, respectively, and report results in Table 1.

As shown in Table 1, nominally trained models are brittle: while clean accuracies on CIFAR10 and ImageNet are 93.7% and 79.2%, respectively, these drop to 4.8% and 21% under black-box attacks. Moreover, while SND improves robustness, *i.e.* to 25.5% and 60.2% on CIFAR10 and ImageNet, our proposed anti-adversary outperforms SND across attacks and datasets. For instance, on ImageNet we outperform SND by 1.2% and 5.8% for Bandits and NES, respectively, while not hurting clean accuracy.

Robustness of Robustly-Trained f_θ. The previous section showed our anti-adversary layer improves black-box robustness of a nominally-trained f_θ. Here, we investigate if our layer can also improve robustness when f_θ is robustly trained. We thus study black-box robustness of our anti-adversary layer with two state-of-the-art robustly trained f_θ (IN-Pret and AWP) on CIFAR10 and CIFAR100. We report robust accuracy for 1000 instances under Bandits and NES attacks. For the more efficient Square attack, we report robust accuracy on the full test set.

In Table 2, we report the black-box robust accuracies on CIFAR10 and CIFAR100. Confirming previous observations, prepending our anti-adversary layer to f_θ does not hurt clean accuracy. More importantly, although f_θ is robustly trained, our anti-adversary layer can still boost robustness by an impressive $\sim$ 15%. For instance, even for the highest worst-case robust accuracy on CIFAR10 (AWP's 66.18%), the anti-adversary improves robustness by 14.53%, reaching 80.71%. Similarly, for CIFAR100 our layer improves the worst-case black-box robustness of AWP by 11.7%. Overall, our layer consistently improves black-box robustness against all attacks and for all robust training methods f_θ on both CIFAR10 and CIFAR100.

3.2. Robustness under White-Box Attacks

In this setting, the attacker (1) only accesses the classifier's outputs and gradients. Note that the attacker ignores the classifier's inner workings and training specifications. While this setup is less realistic than the black-box setting, it is an interesting measure of robustness by providing more information to the attacker, which is related to why most prior works report performance in this category by computing accuracy under PGD [19] or AutoAttack, such as [29].

We prepend our anti-adversary layer to robustly trained classifiers f_θ and assess robustness on CIFAR10 and CIFAR100. We report robust accuracy against three white-box attacks, namely APGD, ADLR and FAB, and measure performance under AutoAttack with $\epsilon = {}^{8}\!/_{255}$ in (1).

In Table 3 we report robust accuracies on CIFAR10 and CIFAR100, respectively. We observe that, on CIFAR10, our anti-adversary layer remarkably improves robust accuracy against AutoAttack. In particular, for the strongest defense we consider, AWP, adversarial robustness increases from 60.53% to an astounding 79.21%. We observe similar results for CIFAR100. In particular, Table 3 shows that the anti-adversary layer induces an average improvement of $\sim$ 11%, where the adversarial robustness of ImageNet-Pre increases from 28.96% to over 40%. The improvement is consistent across defenses on CIFAR100 with a worst-case drop in clean accuracy of 1%. In contrast, integrating SND with AWP 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%) on CIFAR10.

References

[1] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[2] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning (ICML)*, 2018. 1

[3] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *International Conference on Machine Learning (ICML)*, 2018. 2

[4] Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In *International Conference on Learning Representations (ICLR)*, 2018. 1

[5] Junyoung Byun, Hyojun Go, and Changick Kim. Small input noise is enough to defend against query-based black-box attacks. *https://openreview.net/forum?id=6HlaJSlQFEj*, 2021. 1, 2, 3

[6] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness. *arXiv:1902.06705*, 2019. 1

[7] Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. Unlabeled data improves adversarial robustness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 2

[8] Francesco Croce and Matthias Hein. Minimally distorted adversarial examples with a fast adaptive boundary attack. In *International conference on machine learning (ICML)*, 2020. 2

[9] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning (ICML)*, 2020. 2

[10] Yinpeng Dong, Qi-An Fu, Xiao Yang, Tianyu Pang, Hang Su, Zihao Xiao, and Jun Zhu. Benchmarking adversarial robustness on image classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[11] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *International Conference on Learning Representations (ICLR)*, 2015. 1

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[13] Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. *International Conference on Machine Learning (ICML)*, 2019. 2

[14] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *International Conference on Machine Learning. (ICML)*, 2018. 2

[15] Andrew Ilyas, Logan Engstrom, and Aleksander Madry. Prior convictions: Black-box adversarial attacks with bandits and priors. In *International Conference on Learning Representations (ICLR)*, 2019. 2

[16] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. In *University of Toronto, Canada*, 2009. 3

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2012. 3

[18] Xuanqing Liu, Minhao Cheng, Huan Zhang, and Cho-Jui Hsieh. Towards robust neural networks via random self-ensemble. *CoRR*, abs/1712.00673, 2017. 2

[19] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representations (ICLR)*, 2018. 2, 3

[20] Nicolas Papernot, Patrick D. McDaniel, Ian J. Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Practical black-box attacks against deep learning systems using adversarial examples. 2016. 1

[21] Adnan Siraj Rakin, Zhezhi He, and Deliang Fan. Parametric noise injection: Trainable randomness to improve deep neural network robustness against adversarial attack, 2018. 2

[22] Vikash Sehwag, Shiqi Wang, Prateek Mittal, and Suman Jana. Hydra: Pruning adversarially robust neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2

[23] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 1

[24] Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 1

[25] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations (ICLR)*, 2019. 1

[26] Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *International Conference on Learning Representations (ICLR)*, 2019. 2

[27] Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2

[28] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. In *International Conference on Learning Representations (ICLR)*, 2018. 2

[29] Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan L Yuille, and Kaiming He. Feature denoising for improving adversarial robustness. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[30] 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 *International Conference on Machine Learning (ICML)*, 2019. 2