

---

# Boosting Adversarial Training with Hypersphere Embedding

---

Tianyu Pang\*, Xiao Yang\*, Yinpeng Dong, Kun Xu, Jun Zhu, Hang Su†

Dept. of Comp. Sci. & Tech., Institute for AI, BNRist Center

Tsinghua-Bosch Joint ML Center, THBI Lab, Tsinghua University, Beijing, China

{pty17, yangxiao19, dyp17}@mails.tsinghua.edu.cn

kunxu.thu@gmail.com, {suhangss, dcszj}@mail.tsinghua.edu.cn

## Abstract

Adversarial training (AT) is one of the most effective defenses against adversarial attacks for deep learning models. In this work, we advocate incorporating the hypersphere embedding (HE) mechanism into the AT procedure by regularizing the features onto compact manifolds, which constitutes a lightweight yet effective module to blend in the strength of representation learning. Our extensive analyses reveal that AT and HE are well coupled to benefit the robustness of the adversarially trained models from several aspects. We validate the effectiveness and adaptability of HE by embedding it into the popular AT frameworks including PGD-AT, ALP, and TRADES, as well as the FreeAT and FastAT strategies. In the experiments, we evaluate our methods under a wide range of adversarial attacks on the CIFAR-10 and ImageNet datasets, which verifies that integrating HE can consistently enhance the model robustness for each AT framework with little extra computation.

## 1 Introduction

The adversarial vulnerability of deep learning models has been widely recognized in recent years [4, 24, 67]. To mitigate this potential threat, a number of defenses have been proposed, but most of them are ultimately defeated by the attacks adapted to the specific details of the defenses [2, 8]. Among the existing defenses, **adversarial training (AT)** is a general strategy achieving the state-of-the-art robustness under different settings [60, 70, 79, 82, 83, 85, 93]. Various efforts have been devoted to improving AT from different aspects, including accelerating the training procedure [63, 64, 78, 88] and exploiting extra labeled and unlabeled training data [1, 10, 27, 87], which are conducive in the cases with limited computational resources or additional data accessibility.

In the meanwhile, another research route focuses on boosting the adversarially trained models via imposing more direct supervision to regularize the learned representations. Along this line, recent progress shows that encoding triplet-wise metric learning or maximizing the optimal transport (OT) distance of data batch in AT is effective to leverage the inter-sample interactions, which can promote the learning of robust classifiers [38, 45, 49, 89]. However, optimization on the sampled triplets or the OT distance is usually of high computational cost, while the sampling process in metric learning could also introduce extra class biases on unbalanced data [54, 62].

In this work, we provide a lightweight yet competent module to tackle several defects in the learning dynamics of existing AT frameworks, and facilitate the adversarially trained networks learning more robust features. Methodologically, we augment the AT frameworks by integrating the **hypersphere embedding (HE)** mechanism, which normalizes the features in the penultimate layer and the weights in the softmax layer with an additive angular margin. Except for the generic benefits of HE on

---

\*Equal contribution. † Corresponding author.

Table 1: Formulations of the AT frameworks without ( $\times$ ) or with ( $\checkmark$ ) HE. The notations are defined in Sec. 2.2. We substitute the adversarial attacks in ALP with untargeted PGD as suggested [20].

Strategy	HE	Training objective $\mathcal{L}_T$	Adversarial objective $\mathcal{L}_A$
PGD-AT	$\times$	$\mathcal{L}_{CE}(f(x^*), y)$	$\mathcal{L}_{CE}(f(x'), y)$
	$\checkmark$	$\mathcal{L}_{CE}^m(\tilde{f}(x^*), y)$	$\mathcal{L}_{CE}(\tilde{f}(x'), y)$
ALP	$\times$	$\alpha \mathcal{L}_{CE}(f(x), y) + (1-\alpha) \mathcal{L}_{CE}(f(x^*), y) + \lambda \ \mathbf{W}^\top(z - z^*)\ _2$	$\mathcal{L}_{CE}(f(x'), y)$
	$\checkmark$	$\alpha \mathcal{L}_{CE}^m(\tilde{f}(x), y) + (1-\alpha) \mathcal{L}_{CE}^m(\tilde{f}(x^*), y) + \lambda \ \widetilde{\mathbf{W}}^\top(\tilde{z} - \tilde{z}^*)\ _2$	$\mathcal{L}_{CE}(\tilde{f}(x'), y)$
TRADES	$\times$	$\mathcal{L}_{CE}(f(x), y) + \lambda \mathcal{L}_{CE}(f(x^*), f(x))$	$\mathcal{L}_{CE}(f(x'), f(x))$
	$\checkmark$	$\mathcal{L}_{CE}^m(\tilde{f}(x), y) + \lambda \mathcal{L}_{CE}(\tilde{f}(x^*), \tilde{f}(x))$	$\mathcal{L}_{CE}(\tilde{f}(x'), \tilde{f}(x))$

learning angularly discriminative representations [41, 43, 75, 80], we contribute to the extensive analyses (detailed in Sec. 3) showing that the encoded HE mechanism naturally adapts to AT.

To intuitively explain the main insights, we take a binary classification task as an example, where the cross-entropy (CE) objective equals to maximizing  $\mathcal{L}(x) = (W_0 - W_1)^\top z = \|W_{01}\| \|z\| \cos(\theta)$  on an input  $x$  with label  $y = 0$ . **(i)** If  $x$  is correctly classified, there is  $\mathcal{L}(x) > 0$ , and adversaries aim to craft  $x'$  such that  $\mathcal{L}(x') < 0$ . Since  $\|W_{01}\|$  and  $\|z\|$  are always positive, they cannot alter the sign of  $\mathcal{L}$ . Thus feature normalization (FN) and weight normalization (WN) encourage the adversaries to attack the crucial component  $\cos(\theta)$ , which results in more efficient perturbations when crafting adversarial examples in AT; **(ii)** In a data batch, points with larger  $\|z\|$  will dominate (vicious circle on increasing  $\|z\|$ ), which makes the model ignore the critical component  $\cos(\theta)$ . FN alleviates this problem by encouraging the model to devote more efforts on learning hard examples, and well-learned hard examples will dynamically have smaller weights during training since  $\cos(\theta)$  is bounded; This can promote the worst-case performance under adversarial attacks; **(iii)** When there are much more samples of label 0, the CE objective will tend to have  $\|W_0\| \gg \|W_1\|$  to minimize the loss. WN can relieve this trend and encourage  $W_0$  and  $W_1$  to diversify in directions. This mechanism alleviates the unbalanced label distributions caused by the untargeted or multi-targeted attacks applied in AT [25, 44], where the resulted adversarial labels depend on the semantic similarity among classes; **(iv)** The angular margin (AM) induces a larger inter-class variance and margin under the angular metric to further improve model robustness, which plays a similar role as the margin in SVM.

Our method is concise and easy to implement. To validate the effectiveness, we consider three typical AT frameworks to incorporate with HE, namely, **PGD-AT** [44], **ALP** [31], and **TRADES** [90], as summarized in Table 1. We further verify the generality of our method by evaluating the combination of HE with previous strategies on accelerating AT, e.g., **FreeAT** [64] and **FastAT** [78]. In Sec. 4, we empirically evaluate the defenses on CIFAR-10 [34] and ImageNet [16] under several different adversarial attacks, including the commonly adopted PGD [44] and other strong ones like the feature attack [39], FAB [14], SPSA [72], and NES [29], etc. We also test on the CIFAR-10-C and ImageNet-C datasets with corrupted images to inspect the robustness under general transformations [26]. The results demonstrate that incorporating HE can consistently improve the performance of the models trained by each AT framework, while introducing little extra computation.

## 2 Methodology

In this section, we define the notations, introduce the hypersphere embedding (HE) mechanism, and provide the formulations under the adversarial training (AT) frameworks. Due to the limited space, we extensively introduce the related work in Appendix B, including those on combining metric learning with AT [38, 45, 48, 50, 89] and further present their bottlenecks.

### 2.1 Notations

For the classification task with  $L$  labels in  $[L] := \{1, \dots, L\}$ , a deep neural network (DNN) can be generally denoted as the mapping function  $f(x)$  for the input  $x$  as

$$f(x) = \mathbb{S}(\mathbf{W}^\top z + b), \quad (1)$$

where  $z = z(x; \omega)$  is the extracted feature with model parameters  $\omega$ , the matrix  $\mathbf{W} = (W_1, \dots, W_L)$  and vector  $b$  are respectively the weight and bias in the softmax layer, and  $\mathbb{S}(h) : \mathbb{R}^L \rightarrow \mathbb{R}^L$  is the

softmax function. One common training objective for DNNs is the cross-entropy (CE) loss defined as

$$\mathcal{L}_{\text{CE}}(f(x), y) = -\mathbf{1}_y^\top \log f(x), \quad (2)$$

where  $\mathbf{1}_y$  is the one-hot encoding of label  $y$  and the logarithm of a vector is taken element-wisely. In this paper, we use  $\angle(u, v)$  to denote the angle between vectors  $u$  and  $v$ .

## 2.2 The AT frameworks with HE

Adversarial training (AT) is one of the most effective and widely studied defense strategies against adversarial vulnerability [6, 37]. Most of the AT methods can be formulated as a two-stage framework:

$$\min_{\omega, \mathbf{W}} \mathbb{E} [\mathcal{L}_T(\omega, \mathbf{W} | x, x^*, y)], \text{ where } x^* = \arg \max_{x' \in \mathbf{B}(x)} \mathcal{L}_A(x' | x, y, \omega, \mathbf{W}). \quad (3)$$

Here  $\mathbb{E}[\cdot]$  is the expectation w.r.t. the data distribution,  $\mathbf{B}(x)$  is a set of allowed points around  $x$ ,  $\mathcal{L}_T$  and  $\mathcal{L}_A$  are the training and adversarial objectives, respectively. Since the inner maximization and outer minimization problems are mutually coupled, they are iteratively executed in training until the model parameters  $\omega$  and  $\mathbf{W}$  converge [44]. To promote the performance of the adversarially trained models, recent work proposes to embed pair-wise or triplet-wise metric learning into AT [38, 45, 89], which facilitates the neural networks learning more robust representations. Although these methods are appealing, they could introduce high computational overhead [45], cause unexpected class biases [28], or be vulnerable under strong adversarial attacks [39].

In this paper, we address the above deficiencies by presenting a lightweight yet effective module that integrates the **hypersphere embedding (HE)** mechanism with an AT procedure. Though HE is not completely new, our analysis in Sec. 3 demonstrates that HE naturally adapts to the learning dynamics of AT and can induce several advantages special to the adversarial setting. Specifically, the HE mechanism involves three typical operations including feature normalization (FN), weight normalization (WN), and angular margins (AM), as described below.

Note that in Eq. (1) there is  $\mathbf{W}^\top z = (W_1^\top z, \dots, W_L^\top z)$ , and  $\forall l \in [L]$ , the inner product  $W_l^\top z = \|W_l\| \|z\| \cos(\theta_l)$ , where  $\theta_l = \angle(W_l, z)$ .<sup>1</sup> Then the WN and FN operations can be denoted as

$$\text{WN operation: } \widetilde{W}_l = \frac{W_l}{\|W_l\|}; \text{ FN operation: } \widetilde{z} = \frac{z}{\|z\|}, \quad (4)$$

Let  $\cos \boldsymbol{\theta} = (\cos(\theta_1), \dots, \cos(\theta_L))$  and  $\widetilde{\mathbf{W}}$  be the weight matrix after executing WN on each column vector  $W_l$ . Then, the output predictions of the DNNs with HE become

$$\tilde{f}(x) = \mathbb{S}(\widetilde{\mathbf{W}}^\top \widetilde{z}) = \mathbb{S}(\cos \boldsymbol{\theta}), \quad (5)$$

where no bias vector  $b$  exists in  $\tilde{f}(x)$  [42, 75]. In contrast, the **AM operation** is only performed in the training phase, where  $\tilde{f}(x)$  is fed into the CE loss with a margin  $m$  [76], formulated as

$$\mathcal{L}_{\text{CE}}^m(\tilde{f}(x), y) = -\mathbf{1}_y^\top \log \mathbb{S}(s \cdot (\cos \boldsymbol{\theta} - m \cdot \mathbf{1}_y)). \quad (6)$$

Here  $s > 0$  is a hyperparameter to improve the numerical stability during training [75]. To highlight our main contributions in terms of methodology, we summarize the proposed formulas of AT in Table 1. We mainly consider three representative AT frameworks including **PGD-AT** [44], **ALP** [31], and **TRADES** [90]. The differences between our enhanced versions (with HE) from the original versions (without HE) are colorized. Note that we apply the HE mechanism both on the adversarial objective  $\mathcal{L}_A$  for constructing adversarial examples (the inner maximization problem), and the training objective  $\mathcal{L}_T$  for updating parameters (the outer minimization problem).

## 3 Analysis of the benefits

In this section, we analyze the benefits induced by the mutual interaction between AT and HE under the  $\ell_p$ -bounded threat model [9], where  $\mathbf{B}(x) = \{x' | \|x' - x\|_p \leq \epsilon\}$  and  $\epsilon$  is the maximal perturbation. Detailed proofs for the conclusions below can be found in Appendix A.

---

<sup>1</sup>We omit the subscript of  $\ell_2$ -norm without ambiguity.

### 3.1 Formalized first-order adversary

Most of the adversarial attacks applied in AT belong to the family of first-order adversaries [65], due to the computational efficiency. We first define the vector function  $\mathbb{U}_p$  as

$$\mathbb{U}_p(u) = \arg \max_{\|v\|_p \leq 1} u^\top v, \text{ where } u^\top \mathbb{U}_p(u) = \|u\|_q. \quad (7)$$

Here  $\|\cdot\|_q$  is the dual norm of  $\|\cdot\|_p$  with  $\frac{1}{p} + \frac{1}{q} = 1$  [5]. Specially, there are  $\mathbb{U}_2(u) = \frac{u}{\|u\|_2}$  and  $\mathbb{U}_\infty(u) = \text{sign}(u)$ . If  $u = \nabla_x \mathcal{L}_A$ , then  $\mathbb{U}_p(\nabla_x \mathcal{L}_A)$  is the direction of greatest increase of  $\mathcal{L}_A$  under the first-order Taylor's expansion [33] and the  $\ell_p$ -norm constraint, as stated below:

**Lemma 1.** (*First-order adversary*) Given the adversarial objective  $\mathcal{L}_A$  and the set  $\mathbf{B}(x) = \{x' \mid \|x' - x\|_p \leq \epsilon\}$ , under the first-order Taylor's expansion, the solution for  $\max_{x' \in \mathbf{B}(x)} \mathcal{L}_A(x')$  is  $x^* = x + \epsilon \mathbb{U}_p(\nabla_x \mathcal{L}_A(x))$ . Furthermore, there is  $\mathcal{L}_A(x^*) = \mathcal{L}_A(x) + \epsilon \|\nabla_x \mathcal{L}_A(x)\|_q$ .

According to the one-step formula in Lemma 1, we can generalize to the multi-step generation process of first-order adversaries under the  $\ell_p$ -bounded threat model. For example, in the  $t$ -th step of the iterative attack with step size  $\eta$  [35], the adversarial example  $x^{(t)}$  is updated as

$$x^{(t)} = x^{(t-1)} + \eta \mathbb{U}_p(\nabla_x \mathcal{L}_A(x^{(t-1)})), \quad (8)$$

where the increment of the loss is  $\Delta \mathcal{L}_A = \mathcal{L}_A(x^{(t)}) - \mathcal{L}_A(x^{(t-1)}) = \eta \|\nabla_x \mathcal{L}_A(x^{(t-1)})\|_q$ .

### 3.2 The inner maximization problem in AT

As shown in Table 1, the adversarial objectives  $\mathcal{L}_A$  are usually the CE loss between the adversarial prediction  $f(x')$  and the target prediction  $f(x)$  or  $1_y$ . Thus, to investigate the inner maximization problem of  $\mathcal{L}_A$ , we expand the gradient of CE loss w.r.t.  $x'$  as below:

**Lemma 2.** (*The gradient of CE loss*) Let  $W_{ij} = W_i - W_j$  be the residual vector between two weights, and  $z' = z(x'; \omega)$  be the mapped feature of the adversarial example  $x'$ , then there is

$$\nabla_{x'} \mathcal{L}_{CE}(f(x'), f(x)) = - \sum_{i \neq j} f(x)_i f(x')_j \nabla_{x'} (W_{ij}^\top z'). \quad (9)$$

If  $f(x) = 1_y$  is the one-hot label vector, we have  $\nabla_{x'} \mathcal{L}_{CE}(f(x'), y) = - \sum_{l \neq y} f(x')_l \nabla_{x'} (W_{yl}^\top z')$ .

Lemma 2 indicates that the gradient of CE loss can be decomposed into the linear combination of the gradients on the residual logits  $W_{ij}^\top z'$ . Let  $y^*$  be the predicted label on the finally crafted adversarial example  $x^*$ , where  $y^* \neq y$ . Based on the empirical observations [23, 47], we are justified to assume that  $f(x)_y$  is much larger than  $f(x)_l$  for  $l \neq y$ , and  $f(x')_{y^*}$  is much larger than  $f(x')_l$  for  $l \notin \{y, y^*\}$ . Then we can approximate the linear combination in Eq. (9) with the dominated term as

$$\nabla_{x'} \mathcal{L}_{CE}(f(x'), f(x)) \approx -f(x)_y f(x')_{y^*} \nabla_{x'} (W_{yy^*}^\top z'), \text{ where } W_{yy^*} = W_y - W_{y^*}. \quad (10)$$

Let  $\theta'_{yy^*} = \angle(W_{yy^*}, z')$ , there is  $W_{yy^*}^\top z' = \|W_{yy^*}\| \|z'\| \cos(\theta'_{yy^*})$  and  $W_{yy^*}$  does not depend on  $x'$ . Thus by substituting Eq. (10) into Eq. (8), the update direction of each attacking step becomes

$$\mathbb{U}_p[\nabla_{x'} \mathcal{L}_{CE}(f(x'), f(x))] \approx -\mathbb{U}_p[\nabla_{x'} (\|z'\| \cos(\theta'_{yy^*}))], \quad (11)$$

where the factor  $f(x)_y f(x')_{y^*}$  is eliminated according to the definition of  $\mathbb{U}_p[\cdot]$ . Note that Eq. (11) also holds when  $f(x) = 1_y$ , and the resulted adversarial objective is analogous to the C&W attack [7].

### 3.3 Benefits from feature normalization

To investigate the effects of FN alone, we deactivate the WN operation in Eq. (5) and denote

$$\bar{f}(x) = \mathbb{S}(\mathbf{W}^\top \tilde{z}). \quad (12)$$

Then similar to Eq. (11), we can obtain the update direction of the attack with FN applied as

$$\mathbb{U}_p[\nabla_{x'} \mathcal{L}_{CE}(\bar{f}(x'), \bar{f}(x))] \approx -\mathbb{U}_p[\nabla_{x'} (\cos(\theta'_{yy^*}))]. \quad (13)$$

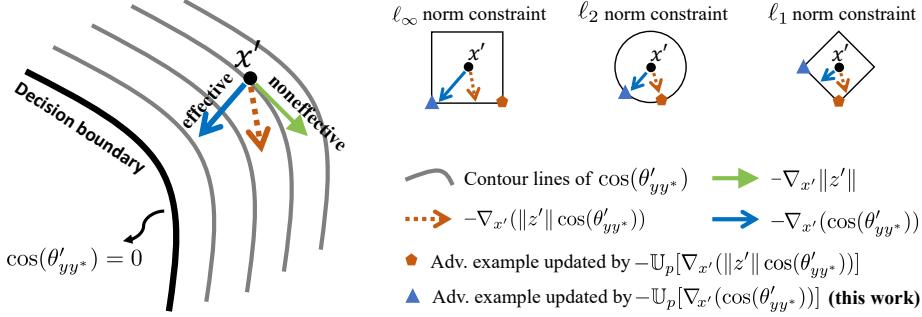


Figure 1: Intuitive illustration in the input space. When applying FN in  $\mathcal{L}_A$ , the adversary can take more effective update steps to move  $x'$  across the decision boundary defined by  $\cos(\theta'_{yy^*}) = 0$ .

**More effective adversarial perturbations.** In the AT procedure, we prefer to craft adversarial examples more efficiently to reduce the computational burden [40, 74, 78]. As shown in Fig. 1, to successfully fool the model to classify  $x'$  into the label  $y^*$ , the adversary needs to craft iterative perturbations to move  $x'$  across the decision boundary defined by  $\cos(\theta'_{yy^*}) = 0$ . Under the first-order optimization, the most effective direction to reach the decision boundary is along  $-\nabla_{x'} (\cos(\theta'_{yy^*}))$ , namely, the direction with the steepest descent of  $\cos(\theta'_{yy^*})$ . In contrast, the direction of  $-\nabla_{x'} \|\boldsymbol{z}'\|$  is nearly tangent to the contours of  $\cos(\theta'_{yy^*})$ , especially in high-dimension spaces, which is noneffective as the adversarial perturbation. Actually from Eq. (1) we can observe that when there is no bias term in the softmax layer, changing the norm  $\|\boldsymbol{z}'\|$  will not affect the predicted labels at all.

By comparing Eq. (11) and Eq. (13), we can find that applying FN in the adversarial objective  $\mathcal{L}_A$  exactly removes the noneffective component caused by  $\|\boldsymbol{z}'\|$ , and encourages the adversarial perturbations to be aligned with the effective direction  $-\nabla_{x'} (\cos(\theta'_{yy^*}))$  under the  $\ell_p$ -norm constraint. This facilitate crafting adversarial examples with fewer iterations and improve the efficiency of the AT progress, as empirically verified in the left panel of Fig. 2. Besides, in Fig. 1 we also provide three instantiations of the  $\ell_p$ -norm constraint. We can see that if we do not use FN, the impact of the noneffective component of  $-\nabla_{x'} \|\boldsymbol{z}'\|$  could be magnified under, e.g., the  $\ell_\infty$ -norm constraint, which could consequently require more iterative steps and degrade the training efficiency.

**Better learning on hard (adversarial) examples.** As to the benefits of applying FN in the training objective  $\mathcal{L}_T$ , we formally show that FN can promote learning on hard examples, as empirically observed in the previous work [57]. In the adversarial setting, this property can promote the worst-case performance under potential adversarial threats. Specifically, the model parameters  $\omega$  is updated towards  $-\nabla_\omega \mathcal{L}_{CE}$ . When FN is not applied, we can use similar derivations as in Lemma 2 to obtain

$$-\nabla_\omega \mathcal{L}_{CE}(f(x), y) = \sum_{l \neq y} f(x)_l \|W_{yl}\| (\cos(\theta_{yl}) \nabla_\omega \|\boldsymbol{z}\| + \|\boldsymbol{z}\| \nabla_\omega \cos(\theta_{yl})), \quad (14)$$

where  $W_{yl} = W_y - W_l$  and  $\theta_{yl} = \angle(W_{yl}, \boldsymbol{z})$ . According to Eq. (14), when we use a mini-batch of data to update  $\omega$ , the inputs with small  $\nabla_\omega \|\boldsymbol{z}\|$  or  $\nabla_\omega \cos(\theta_{yl})$  contribute less in the direction of model updating, which are qualitatively regarded as hard examples [57]. This causes the training process to devote noneffective efforts to increasing  $\|\boldsymbol{z}\|$  for easy examples and consequently overlook the hard ones, which leads to vicious circles and could degrade the model robustness against strong adversarial attacks [58, 66]. As shown in Fig. 3, the hard examples in AT are usually the crafted adversarial examples, which are those we actually expect the model to focus on in the AT procedure. In comparison, when FN is applied, there is  $\nabla_\omega \|\tilde{\boldsymbol{z}}\| = 0$ , then  $\omega$  is updated towards

$$-\nabla_\omega \mathcal{L}_{CE}(\bar{f}(x), y) = \sum_{l \neq y} \bar{f}(x)_l \|W_{yl}\| \nabla_\omega \cos(\theta_{yl}). \quad (15)$$

In this case, due to the bounded value range  $[-1, 1]$  of cosine function, the easy examples will contribute less when they are well learned, i.e., have large  $\cos(\theta_{yl})$ , while the hard examples could later dominate the training. This causes a dynamic training procedure similar to curriculum learning [3].

### 3.4 Benefits from weight normalization

In the AT procedure, we usually apply untargeted attacks [20, 44]. Since we do not explicit assign targets, the resulted prediction labels and feature locations of the crafted adversarial examples will

Table 2: Classification accuracy (%) on **CIFAR-10** under the *white-box* threat model. The perturbation  $\epsilon = 0.031$ , step size  $\eta = 0.003$ . We highlight the best-performance model under each attack.

Defense	Clean	PGD-20	PGD-500	MIM-20	FGSM	DeepFool	C&W	FeaAtt.	FAB
PGD-AT	86.75	53.97	51.63	55.08	59.70	57.26	84.00	52.38	51.23
PGD-AT+ <b>HE</b>	86.19	59.36	57.59	60.19	63.77	61.56	84.07	52.88	<b>54.45</b>
ALP	87.18	52.29	50.13	53.35	58.99	59.40	84.96	49.55	50.54
ALP+ <b>HE</b>	<b>89.91</b>	57.69	51.78	58.63	65.08	<b>65.19</b>	<b>87.86</b>	48.64	51.86
TRADES	84.62	56.48	54.84	57.14	61.02	60.70	81.13	55.09	53.58
TRADES+ <b>HE</b>	84.88	<b>62.02</b>	<b>60.75</b>	<b>62.71</b>	<b>65.69</b>	60.48	81.44	<b>58.13</b>	53.50

Table 3: Validation of combining FastAT and FreeAT with HE and m-HE on **CIFAR-10**. We report the accuracy (%) on clean and PGD, as well as the total training time (min).

Defense	Epo.	Clean	PGD-50	Time
FastAT	30	<b>83.80</b>	46.40	11.38
FastAT+ <b>HE</b>	30	82.58	52.55	11.48
FastAT+m- <b>HE</b>	30	83.14	<b>53.49</b>	11.49
FreeAT	10	77.21	46.14	15.78
FreeAT+ <b>HE</b>	10	76.85	50.98	15.87
FreeAT+m- <b>HE</b>	10	<b>77.59</b>	<b>51.85</b>	15.91

Table 4: Top-1 classification accuracy (%) on **ImageNet** under the *white-box* threat model.

Model	Method	Clean	PGD-10	PGD-50
ResNet-50	FreeAT	60.28	32.13	31.39
	FreeAT+ <b>HE</b>	<b>61.83</b>	<b>40.22</b>	<b>39.85</b>
ResNet-152	FreeAT	65.20	36.97	35.87
	FreeAT+ <b>HE</b>	<b>65.41</b>	<b>43.24</b>	<b>42.60</b>
WRN-50-2	FreeAT	64.18	36.24	35.38
	FreeAT+ <b>HE</b>	<b>65.28</b>	<b>43.83</b>	<b>43.47</b>
WRN-101-2	FreeAT	66.15	39.35	38.23
	FreeAT+ <b>HE</b>	<b>66.37</b>	<b>45.35</b>	<b>45.04</b>

depend on the unbalanced semantic similarity among different classes. For example, the learned features of dogs and cats are usually closer than those between dogs and planes, so an untargeted adversary will prefer to fool the model to predict the cat label on a dog image, rather than the plane label [46]. To understand how the adversarial class biases affect training, assuming that we perform the gradient descent on a data batch  $\mathcal{D} = \{(x^k, y^k)\}_{k \in [N]}$ . Then we can derive that  $\forall l \in [L]$ , the softmax weight  $W_l$  is updated towards

$$-\nabla_{W_l} \mathcal{L}_{CE}(\mathcal{D}) = \sum_{x^k \in \mathcal{D}_l} z^k - \sum_{x^k \in \mathcal{D}} f(x^k)_l \cdot z^k, \quad (16)$$

where  $\mathcal{D}_l$  is the subset of  $\mathcal{D}$  with true label  $l$ . We can see that the weight  $W_l$  will tend to have larger norm when there are more data or easy examples in class  $l$ , i.e., larger  $|\mathcal{D}_l|$  or  $\|z^k\|$  for  $x^k \in \mathcal{D}_l$ . Besides, if an input  $x$  in the batch is adversarial, then  $f(x)_y$  is usually small and consequently  $z$  will have a large effect on the update of  $W_y$ . Since there is  $W_{yy^*}^\top z < 0$ ,  $W_y$  will be updated towards  $W_{y^*}$  both in norm and direction, which causes repetitive oscillation during training.

When applying WN, the update of  $W_l$  will only depend on the averaged feature direction within each class, which alleviates the noneffective oscillation on the weight norm and speed up training [59]. Besides, when FN and WN are both applied, the inner products  $\mathbf{W}^\top z$  in the softmax layer will become the angular metric  $\cos \theta$ , as shown in Eq. (5). Then we can naturally introduce AM to learn angularly more discriminative and robust features [19, 84].

### 3.5 Modifications to better utilize strong adversaries

In most of the AT procedures, the crafted adversarial examples will only be used once in a single training step to update the model parameters [44], which means hard adversarial examples may not have an chance to gradually dominate the training as introduced in Eq. (15). Since  $\nabla_{\omega} \cos(\theta_{yy^*}) = -\sin(\theta_{yy^*}) \nabla_{\omega} \theta_{yy^*}$ , the weak adversarial examples around the decision boundary with  $\theta_{yy^*} \sim 90^\circ$  have higher weights  $\sin(\theta_{yy^*})$ . This makes the model tend to overlook the strong adversarial examples with large  $\theta_{yy^*}$ , which contain abundant information. To be better compatible with strong adversaries, an easy-to-implement way is to directly substitute  $\mathbb{S}(\cos \theta)$  in Eq. (5) with  $\mathbb{S}(-\theta)$ , using the arccos operator. We name this form of embedding as modified HE (**m-HE**), as evaluated in Table 3.

Table 5: Top-1 classification accuracy (%) on **CIFAR-10-C** and **ImageNet-C**. The models are trained on the original datasets CIFAR-10 and ImageNet, respectively. Here ‘mCA’ refers to the mean accuracy averaged on different corruptions and severity. Full version of the table is in Appendix C.6.

Defense	mCA	Blur				Weather				Digital			
		Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Bright	Contra	Elastic	Pixel	JPEG
<b>CIFAR-10-C</b>													
PGD-AT	77.23	81.84	79.69	77.62	80.88	81.32	77.95	61.70	84.05	44.55	80.79	84.76	84.35
PGD-AT+HE	<b>77.29</b>	81.86	79.45	78.17	80.87	80.77	77.98	62.45	83.67	45.11	80.69	84.16	84.10
ALP	77.73	81.94	80.31	78.23	80.97	81.74	79.26	61.51	84.88	45.86	80.91	85.09	84.68
ALP+HE	<b>80.55</b>	80.87	85.23	81.26	84.43	85.14	83.89	68.83	88.33	50.74	84.44	87.44	87.28
TRADES	75.36	79.84	77.72	76.34	78.66	79.52	76.94	59.68	82.06	43.80	78.53	82.65	82.31
TRADES+HE	<b>75.78</b>	80.55	77.61	77.26	79.62	79.23	76.53	61.39	82.33	45.04	79.29	82.50	82.40
<b>ImageNet-C</b>													
FreeAT	28.22	19.15	26.63	25.75	28.25	23.03	23.47	3.71	45.18	5.40	41.76	48.78	52.55
FreeAT+HE	<b>30.04</b>	21.16	29.28	28.08	30.76	26.62	28.35	5.34	49.88	7.03	44.72	51.17	55.05

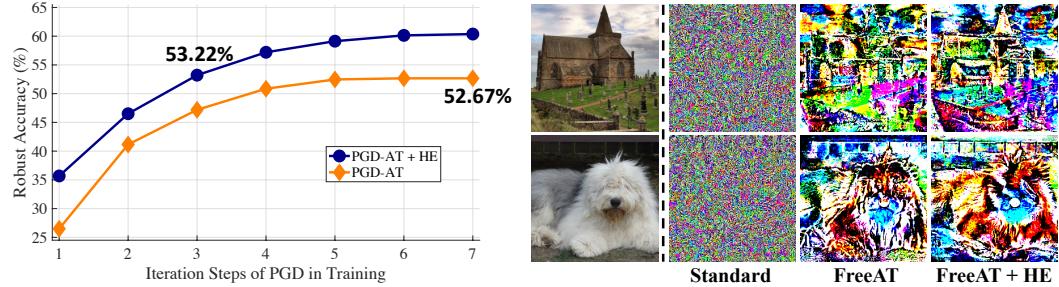


Figure 2: *Left*. Accuracy (%) under PGD-20, where the models are trained by PGD-AT with different iteration steps on **CIFAR-10**. *Right*. Visualization of the adversarial perturbations on **ImageNet**.

## 4 Experiments

**CIFAR-10 [34] setup.** We apply the wide residual network WRN-34-10 as the model architecture [86]. For each AT framework, we set the maximal perturbation  $\epsilon = 8/255$ , the perturbation step size  $\eta = 2/255$ , and the number of iterations  $K = 10$ . We apply the momentum SGD [56] optimizer with the initial learning rate of 0.1, and train for 100 epochs. The learning rate decays with a factor of 0.1 at 75 and 90 epochs, respectively. The mini-batch size is 128. Besides, we set the regularization parameter  $1/\lambda$  as 6 for TRADES, and set the adversarial logit pairing weight as 0.5 for ALP [31, 90]. The scale  $s = 15$  and the margin  $m = 0.2$  in HE, where different  $s$  and  $m$  correspond to different trade-offs between the accuracy and robustness, as detailed in Appendix C.3.

**ImageNet [16] setup.** We apply the framework of free adversarial training (**FreeAT**) in Shafahi et al. [64], which has a similar training objective as PGD-AT and can train a robust model using four GPU workers. We set the repeat times  $m = 4$  in FreeAT. The perturbation  $\epsilon = 4/255$  with the step size  $\eta = 1/255$ . We train the model for 90 epochs with the initial learning rate of 0.1, and the mini-batch size is 256. The scale  $s = 10$  and the margin  $m = 0.2$  in HE.<sup>2</sup>

### 4.1 Performance under white-box attacks

On the CIFAR-10 dataset, we test the defenses under different attacks including FGSM [24], PGD [44], MIM [17], Deepfool [46], C&W ( $l_\infty$  version) [7], feature attack [39], and FAB [14]. We report the classification accuracy in Table 2 following the evaluation settings in Zhang et al. [90]. We denote the iteration steps behind the attacking method, e.g., 10-step PGD as PGD-10. To verify that our strategy is generally compatible with previous work on accelerating AT, we combine HE with the one-step based FreeAT and fast adversarial training (**FastAT**) frameworks [78]. We provide the accuracy and training time results in Table 3. We can see that the operations in HE increase negligible computation, even in the cases pursuing extremely fast training. Besides, we also evaluate embedding **m-HE** (introduced in Sec. 3.5) and find it more effective than HE when combining with PGD-AT, FreeAT and Fast AT that exclusively train on adversarial examples. On the ImageNet dataset, we follow the evaluation settings in Shafahi et al. [64] to test under PGD-10 and PGD-50, as shown in Table 4.

<sup>2</sup>Code is available at [https://github.com/ShawnXYang/AT\\_HE](https://github.com/ShawnXYang/AT_HE).

Table 6: Classification accuracy (%) on the clean test data, and under two benchmark attacks RayS and AutoAttack.

Method	Architecture	Clean	RayS	AA
PGD-AT+HE	WRN-34-10	86.25	57.8	53.16
	WRN-34-20	85.14	59.0	53.74

Table 7: Attacking standardly trained WRN-34-10 with or without FN.

Attack	FN	Acc. (%)
PGD-1	✗	67.09
	✓	<b>62.89</b>
PGD-2	✗	50.37
	✓	<b>33.75</b>

Table 8: Classification accuracy (%) under different *black-box* query-based attacks on **CIFAR-10**.

Method	Iterations	PGD-AT	PGD-AT + HE	ALP	ALP + HE	TRADES	TRADES + HE
ZOO	-	73.47	74.65	72.70	74.90	71.92	74.65
	20	73.64	73.66	73.11	74.45	72.12	72.36
	50	68.93	69.31	68.39	68.28	68.20	68.42
SPSA	80	65.67	65.97	65.14	63.89	65.32	65.62
	20	74.68	74.87	74.41	75.91	73.40	73.47
	50	71.22	71.48	70.81	71.11	70.29	70.53
NES	80	69.16	69.92	68.88	68.53	68.83	69.06

**Ablation studies.** To investigate the individual effects caused by the three components FN, WN, and AM in HE, we perform the ablation studies for PGD-AT on CIFAR-10, and attack the trained models with PGD-20. We get the clean and robust accuracy of 86.43% / 54.46% for PGD-AT+FN, and 87.28% / 53.93% for PGD-AT+WN. In contrast, when we apply both the FN and WN, we can get the result of 86.20% / 58.95%. For TRADES, applying appropriate AM can increase  $\sim 2\%$  robust accuracy compared to only applying FN and WN. Detailed results are included in the Appendix C.3.

**Adaptive attacks.** A generic PGD attack apply the cross-entropy loss on the model prediction as the adversarial objective, as reported in Table 2. To exclude the potential effect of gradient obfuscation [2], we construct an adaptive version of the PGD attack against our methods, which uses the training loss in Eq. (6) with the scalar  $s$  and margin  $m$  as the adversarial objective. In this case, when we apply the adaptive PGD-20 / PGD-500 attack, we will get a accuracy of 55.25% / 52.54% for PGD-AT+HE, which is still higher than the accuracy of 53.97% / 51.63% for PGD-AT (quote from Table 2).

**Benchmark attacks.** We evaluate our enhanced models under two stronger benchmark attacks including RayS [11] and AutoAttack [15] on CIFAR-10. We train WRN models via PGD-AT+HE, with weight decay of  $5 \times 10^{-4}$  [51]. For RayS, we evaluate on 1,000 test samples due to the high computation. The results are shown in Table 6, where the trained WRN-34-20 model achieves the state-of-the-art performance (no additional data) according to the reported benchmarks.

## 4.2 Performance under black-box attacks

**Query-based Black-box Attacks.** ZOO [12] proposes to estimate the gradient at each coordinate  $e_i$  as  $\hat{g}_i$ , with a small finite-difference  $\sigma$ . In experiments, we randomly select one sampled coordinate to perform one update with  $\hat{g}_i$ , and adopt the C&W optimization mechanism based on the estimated gradient. We set  $\sigma$  as  $10^{-4}$  and max queries as 20,000. SPSA [72] and NES [29] can make a full gradient evaluation by drawing random samples and obtaining the corresponding loss values. NES randomly samples from a Gaussian distribution to acquire the direction vectors while SPSA samples from a Rademacher distribution. In experiments, we set the number of random samples  $q$  as 128 for every iteration and  $\sigma = 0.001$ . We show the robustness of different iterations against untargeted score-based ZOO, SPSA, and NES in Table 8, where details on these attacks are in Appendix B.2.

We include detailed experiments on the **transfer-based black-box attacks** in Appendix C.4. As expected, these results show that embedding HE can generally provide promotion under the black-box threat models [9], including the transfer-based and the query-based black-box attacks.

## 4.3 Performance under general-purpose attacks

It has been shown that the adversarially trained models could be vulnerable to rotations [21], image corruptions [22] or affine attacks [69]. Therefore, we evaluate on the benchmarks with distributional shifts: CIFAR-10-C and ImageNet-C [26]. As shown in Table 5, we report the classification accuracy

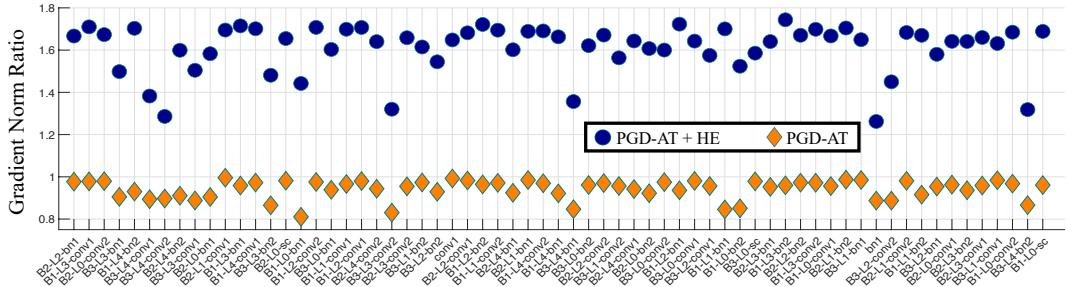


Figure 3: The ratios of  $\mathbb{E}(\|\nabla_{\omega} \mathcal{L}(x^*)\|/\|\nabla_{\omega} \mathcal{L}(x)\|)$  w.r.t. different parameters  $\omega$ , where  $x^*$  and  $x$  are the adversarial example and its clean counterpart. Higher ratio values indicate more attention on the adversarial examples. 'B' refers to block, 'L' refers to layer, 'conv' refers to convolution.

under each corruption averaged on five levels of severity, where the models are WRN-34-10 trained on CIFAR-10 and ResNet-50 trained on ImageNet, respectively. Here we adopt accuracy as the metric to be consistent with other results, while the reported values can easily convert into the corruption error metric [26]. We can find that our methods lead to better robustness under a wide range of corruptions that are not seen in training, which prevents the models from overfitting to certain attacking patterns.

## 4.4 More empirical analyses

As shown in the left panel of Fig. 2, we separately use PGD-1 to PGD-7 to generate adversarial examples in training, then we evaluate the trained models under the PGD-20 attack. Our method requires fewer iterations of PGD to achieve certain robust accuracy, e.g., applying 3-step PGD in PGD-AT+HE is more robust than applying 7-step PGD in PGD-AT, which largely reduces the necessary computation [77]. To verify the mechanism in Fig. 1, we attack a standardly trained WRN-34-10 (no FN applied) model, applying PGD-1 and PGD-2 with or without FN in the adversarial objective. We report the accuracy in Table 7. As seen, the attacks are more efficient with FN, which suggest that the perturbations are crafted along more effective directions.

Besides, previous studies observe that the adversarial examples against robust models exhibit salient data characteristics [30, 61, 68, 71, 92]. So we visualize the untargeted perturbations on ImageNet, as shown in the right panel of Fig. 2. We can observe that the adversarial perturbations produced for our method have sharper profiles and more concrete details, which are better aligned with human perception. Finally in Fig. 3, we calculate the norm ratios of the loss gradient on the adversarial example to it on the clean example. The model is trained for 70 epochs on CIFAR-10 using PGD-AT. The results verify that our method can prompt the training procedure to assign larger gradients on the crafted adversarial examples, which would benefit robust learning.

## 5 Conclusion

In this paper, we propose to embed the HE mechanism into AT, in order to enhance the robustness of the adversarially trained models. We analyze the intriguing benefits induced by the interaction between AT and HE from several aspects. It is worth clarifying that empirically our HE module has varying degrees of adaptability on combining different AT frameworks, depending on the specific training principles. Still, incorporating the HE mechanism is generally conducive to robust learning and compatible with previous strategies, with little extra computation and simple code implementation.

## **Broader Impact**

When deploying machine learning methods into the practical systems, the adversarial vulnerability can cause a potential security risk, as well as the negative impact on the crisis of confidence by the public. To this end, this inherent defect raises the requirements for reliable, general, and lightweight strategies to enhance the model robustness against malicious, especially adversarial attacks. In this work, we provide a simple and efficient way to boost the robustness of the adversarially trained models, which contributes to the modules of constructing more reliable systems in different tasks.

## Acknowledgements

This work was supported by the National Key Research and Development Program of China (No.2020AAA0104304), NSFC Projects (Nos. 61620106010, 62076147, U19B2034, U1811461), Beijing Academy of Artificial Intelligence (BAAI), Tsinghua-Huawei Joint Research Program, a grant from Tsinghua Institute for Guo Qiang, Tiangong Institute for Intelligent Computing, and the NVIDIA NVAIL Program with GPU/DGX Acceleration.

## References

- [1] Jean-Baptiste Alayrac, Jonathan Uesato, Po-Sen Huang, Alhussein Fawzi, Robert Stanforth, and Pushmeet Kohli. Are labels required for improving adversarial robustness? In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 12192–12202, 2019.
- [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.
- [3] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *International Conference on Machine Learning (ICML)*, pages 41–48. ACM, 2009.
- [4] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 387–402. Springer, 2013.
- [5] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.
- [6] Wieland Brendel, Jonas Rauber, Alexey Kurakin, Nicolas Papernot, Behar Veliqi, Sharada P Mohanty, Florian Laurent, Marcel Salathé, Matthias Bethge, Yaodong Yu, et al. Adversarial vision challenge. In *The NeurIPS’18 Competition*, pages 129–153. Springer, 2020.
- [7] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy (S&P)*, 2017.
- [8] Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *ACM Workshop on Artificial Intelligence and Security (AISec)*, 2017.
- [9] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019.
- [10] Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, Percy Liang, and John C Duchi. Unlabeled data improves adversarial robustness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.
- [11] Jinghui Chen and Quanquan Gu. Rays: A ray searching method for hard-label adversarial attack. *arXiv preprint arXiv:2006.12792*, 2020.
- [12] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *ACM Workshop on Artificial Intelligence and Security (AISec)*. ACM, 2017.
- [13] Cody Coleman, Deepak Narayanan, Daniel Kang, Tian Zhao, Jian Zhang, Luigi Nardi, Peter Bailis, Kunle Olukotun, Chris Ré, and Matei Zaharia. Dawnbench: An end-to-end deep learning benchmark and competition. *Training*, 100(101):102, 2017.
- [14] Francesco Croce and Matthias Hein. Minimally distorted adversarial examples with a fast adaptive boundary attack. In *International Conference on Machine Learning (ICML)*, 2020.
- [15] 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.

- [16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.
- [17] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [18] Yinpeng Dong, Qi-An Fu, Xiao Yang, Tianyu Pang, Hang Su, Zihao Xiao, and Jun Zhu. Benchmarking adversarial robustness. *arXiv preprint arXiv:1912.11852*, 2019.
- [19] Gamaleldin Elsayed, Dilip Krishnan, Hossein Mobahi, Kevin Regan, and Samy Bengio. Large margin deep networks for classification. In *Advances in neural information processing systems (NeurIPS)*, pages 842–852, 2018.
- [20] Logan Engstrom, Andrew Ilyas, and Anish Athalye. Evaluating and understanding the robustness of adversarial logit pairing. *arXiv preprint arXiv:1807.10272*, 2018.
- [21] Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. A rotation and a translation suffice: Fooling cnns with simple transformations. In *International Conference on Machine Learning (ICML)*, 2019.
- [22] Justin Gilmer, Nicolas Ford, Nicolas Carlini, and Ekin Cubuk. Adversarial examples are a natural consequence of test error in noise. In *International Conference on Machine Learning (ICML)*, 2019.
- [23] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. <http://www.deeplearningbook.org>.
- [24] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015.
- [25] Sven Gowal, Jonathan Uesato, Chongli Qin, Po-Sen Huang, Timothy Mann, and Pushmeet Kohli. An alternative surrogate loss for pgd-based adversarial testing. *arXiv preprint arXiv:1910.09338*, 2019.
- [26] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)*, 2019.
- [27] Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. In *International Conference on Machine Learning (ICML)*, 2019.
- [28] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*, pages 84–92. Springer, 2015.
- [29] 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.
- [30] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Anish Athalye, Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.
- [31] Harini Kannan, Alexey Kurakin, and Ian Goodfellow. Adversarial logit pairing. *arXiv preprint arXiv:1803.06373*, 2018.
- [32] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.
- [33] Konrad Königsberger. Analysis 2 springer verlag, 2004.
- [34] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

- [35] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *The International Conference on Learning Representations (ICLR) Workshops*, 2017.
- [36] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. In *International Conference on Learning Representations (ICLR)*, 2017.
- [37] Alexey Kurakin, Ian Goodfellow, Samy Bengio, Yinpeng Dong, Fangzhou Liao, Ming Liang, Tianyu Pang, Jun Zhu, Xiaolin Hu, Cihang Xie, et al. Adversarial attacks and defences competition. *arXiv preprint arXiv:1804.00097*, 2018.
- [38] Pengcheng Li, Jinfeng Yi, Bowen Zhou, and Lijun Zhang. Improving the robustness of deep neural networks via adversarial training with triplet loss. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2019.
- [39] Daquan Lin. <https://github.com/Line290/FeatureAttack>, 2019.
- [40] Guanxiong Liu, Issa Khalil, and Abdallah Khreishah. Using single-step adversarial training to defend iterative adversarial examples. *arXiv preprint arXiv:2002.09632*, 2020.
- [41] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Spherenetface: Deep hypersphere embedding for face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 212–220, 2017.
- [42] Weiyang Liu, Yan-Ming Zhang, Xingguo Li, Zhiding Yu, Bo Dai, Tuo Zhao, and Le Song. Deep hyperspherical learning. In *Advances in neural information processing systems (NeurIPS)*, pages 3950–3960, 2017.
- [43] Weiyang Liu, Zhen Liu, Zhehui Chen, Bo Dai, Tuo Zhao, and Le Song. Deep hyperspherical defense against adversarial perturbations. *Submission to ICLR*, 2018.
- [44] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018.
- [45] Chengzhi Mao, Ziyuan Zhong, Junfeng Yang, Carl Vondrick, and Baishakhi Ray. Metric learning for adversarial robustness. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 478–489, 2019.
- [46] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2574–2582, 2016.
- [47] Tianyu Pang, Chao Du, Yinpeng Dong, and Jun Zhu. Towards robust detection of adversarial examples. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4579–4589, 2018.
- [48] Tianyu Pang, Chao Du, and Jun Zhu. Max-mahalanobis linear discriminant analysis networks. In *International Conference on Machine Learning (ICML)*, 2018.
- [49] Tianyu Pang, Kun Xu, Chao Du, Ning Chen, and Jun Zhu. Improving adversarial robustness via promoting ensemble diversity. In *International Conference on Machine Learning (ICML)*, 2019.
- [50] Tianyu Pang, Kun Xu, Yinpeng Dong, Chao Du, Ning Chen, and Jun Zhu. Rethinking softmax cross-entropy loss for adversarial robustness. In *International Conference on Learning Representations (ICLR)*, 2020.
- [51] Tianyu Pang, Xiao Yang, Yinpeng Dong, Hang Su, and Jun Zhu. Bag of tricks for adversarial training. *arXiv preprint arXiv:2010.00467*, 2020.
- [52] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016.

- [53] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against deep learning systems using adversarial examples. *arXiv preprint arXiv:1602.02697*, 2016.
- [54] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al. Deep face recognition. In *The British Machine Vision Conference (BMVC)*, 2015.
- [55] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 8024–8035, 2019.
- [56] Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural networks*, 12(1):145–151, 1999.
- [57] Rajeev Ranjan, Carlos D Castillo, and Rama Chellappa. L2-constrained softmax loss for discriminative face verification. *arXiv preprint arXiv:1703.09507*, 2017.
- [58] Leslie Rice, Eric Wong, and J Zico Kolter. Overfitting in adversarially robust deep learning. *arXiv preprint arXiv:2002.11569*, 2020.
- [59] Tim Salimans and Durk P Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 901–909, 2016.
- [60] Hadi Salman, Jerry Li, Ilya Razenshteyn, Pengchuan Zhang, Huan Zhang, Sebastien Bubeck, and Greg Yang. Provably robust deep learning via adversarially trained smoothed classifiers. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 11289–11300, 2019.
- [61] Shibani Santurkar, Andrew Ilyas, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Image synthesis with a single (robust) classifier. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1260–1271, 2019.
- [62] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015.
- [63] Leo Schwinn and Björn Eskofier. Fast and stable adversarial training through noise injection. *arXiv preprint arXiv:2002.10097*, 2020.
- [64] Ali Shafahi, Mahyar Najibi, Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.
- [65] Carl-Johann Simon-Gabriel, Yann Ollivier, Leon Bottou, Bernhard Schölkopf, and David Lopez-Paz. First-order adversarial vulnerability of neural networks and input dimension. In *International Conference on Machine Learning (ICML)*, 2019.
- [66] Chuanbiao Song, Kun He, Liwei Wang, and John E Hopcroft. Improving the generalization of adversarial training with domain adaptation. In *International Conference on Learning Representations (ICLR)*, 2019.
- [67] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014.
- [68] Guanhong Tao, Shiqing Ma, Yingqi Liu, and Xiangyu Zhang. Attacks meet interpretability: Attribute-steered detection of adversarial samples. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 7717–7728, 2018.
- [69] Florian Tramèr and Dan Boneh. Adversarial training and robustness for multiple perturbations. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5858–5868, 2019.

- [70] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations (ICLR)*, 2018.
- [71] 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.
- [72] Jonathan Uesato, Brendan O’Donoghue, Aaron van den Oord, and Pushmeet Kohli. Adversarial risk and the dangers of evaluating against weak attacks. In *International Conference on Machine Learning (ICML)*, 2018.
- [73] BS Vivek and R Venkatesh Babu. Regularizers for single-step adversarial training. *arXiv preprint arXiv:2002.00614*, 2020.
- [74] S Vivek B and R Venkatesh Babu. Single-step adversarial training with dropout scheduling. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [75] Feng Wang, Xiang Xiang, Jian Cheng, and Alan Loddon Yuille. Normface:  $\ell_2$  hypersphere embedding for face verification. In *ACM International Conference on Multimedia (ACM MM)*, pages 1041–1049. ACM, 2017.
- [76] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5265–5274, 2018.
- [77] Yisen Wang, Xingjun Ma, James Bailey, Jinfeng Yi, Bowen Zhou, and Quanquan Gu. On the convergence and robustness of adversarial training. In *International Conference on Machine Learning (ICML)*, pages 6586–6595, 2019.
- [78] Eric Wong, Leslie Rice, and J. Zico Kolter. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations (ICLR)*, 2020.
- [79] Tong Wu, Liang Tong, and Yevgeniy Vorobeychik. Defending against physically realizable attacks on image classification. In *International Conference on Learning Representations (ICLR)*, 2020.
- [80] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3733–3742, 2018.
- [81] Cihang Xie and Alan Yuille. Intriguing properties of adversarial training at scale. In *International Conference on Learning Representations (ICLR)*, 2020.
- [82] Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan Yuille, and Kaiming He. Feature denoising for improving adversarial robustness. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [83] Xiaogang Xu, Hengshuang Zhao, and Jiaya Jia. Dynamic divide-and-conquer adversarial training for robust semantic segmentation. *arXiv preprint arXiv:2003.06555*, 2020.
- [84] Ziang Yan, Yiwen Guo, and Changshui Zhang. Deep defense: Training dnns with improved adversarial robustness. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 419–428, 2018.
- [85] Ziqing Yang, Yiming Cui, Wanxiang Che, Ting Liu, Shijin Wang, and Guoping Hu. Improving machine reading comprehension via adversarial training. *arXiv preprint arXiv:1911.03614*, 2019.
- [86] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *The British Machine Vision Conference (BMVC)*, 2016.
- [87] Runtian Zhai, Tianle Cai, Di He, Chen Dan, Kun He, John Hopcroft, and Liwei Wang. Adversarially robust generalization just requires more unlabeled data. *arXiv preprint arXiv:1906.00555*, 2019.

- [88] Dinghuai Zhang, Tianyuan Zhang, Yiping Lu, Zhanxing Zhu, and Bin Dong. You only propagate once: Accelerating adversarial training via maximal principle. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.
- [89] Haichao Zhang and Jianyu Wang. Defense against adversarial attacks using feature scattering-based adversarial training. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1829–1839, 2019.
- [90] 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.
- [91] Richard Zhang. Making convolutional networks shift-invariant again. In *International Conference on Machine Learning (ICML)*, 2019.
- [92] Tianyuan Zhang and Zhanxing Zhu. Interpreting adversarially trained convolutional neural networks. In *International Conference on Machine Learning (ICML)*, 2019.
- [93] Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Thomas Goldstein, and Jingjing Liu. Freelb: Enhanced adversarial training for language understanding. In *International Conference on Learning Representations (ICLR)*, 2020.

## A Proofs

### A.1 Proof of Lemma 1

**Lemma 1.** Given a loss function  $\mathcal{L}$  and under the first-order Taylor expansion, the solution of

$$\max_{\|x' - x\|_p \leq \epsilon} \mathcal{L}(x')$$

is  $x^* = x + \epsilon \mathbb{U}_p(\nabla \mathcal{L}(x))$ . Furthermore, there is  $\mathcal{L}(x^*) = \mathcal{L}(x) + \epsilon \|\nabla \mathcal{L}(x)\|_q$ , where  $\|\cdot\|_q$  is the dual norm of  $\|\cdot\|_p$ .

*Proof.* We denote  $x' = x + \epsilon v$ , where  $\|v\|_p \leq 1$ . Then we know that  $\|x' - x\|_p \leq \epsilon$ . Under the first-order Taylor expansion, there is

$$\begin{aligned} \max_{\|x' - x\|_p \leq \epsilon} \mathcal{L}(x') &= \max_{\|v\|_p \leq 1} [\mathcal{L}(x) + \epsilon v^\top \nabla \mathcal{L}(x)] \\ &= \mathcal{L}(x) + \epsilon \max_{\|v\|_p \leq 1} v^\top \nabla \mathcal{L}(x). \end{aligned}$$

According to the definition of the dual norm [5], there is  $\max_{\|v\|_p \leq 1} v^\top \nabla \mathcal{L}(x) = \|\nabla \mathcal{L}(x)\|_q$ , where where  $\|\cdot\|_q$  is the dual norm of  $\|\cdot\|_p$ . Thus we prove that  $\mathcal{L}(x^*) = \mathcal{L}(x) + \epsilon \|\nabla \mathcal{L}(x)\|_q$  and  $x^* = x + \epsilon \mathbb{U}_p(\nabla \mathcal{L}(x))$ .  $\square$

### A.2 Proof of Lemma 2

**Lemma 2.** By derivations, there is

$$\nabla_{x'} \mathcal{L}_{CE}(f(x'), f(x)) = - \sum_{i \neq j} f(x)_i f(x')_j \nabla_{x'}(W_{ij}^\top z'),$$

where  $W_{ij} = W_i - W_j$ ,  $z' = z(x'; \omega)$ . When  $f(x) = 1_y$ , we have  $\nabla_{x'} \mathcal{L}_{CE}(f(x'), y) = - \sum_{l \neq y} f(x')_l \nabla_{x'}(W_{yl}^\top z')$ .

*Proof.* By derivations, there is

$$\begin{aligned} & -\nabla_{x'} \mathcal{L}_{CE}(f(x'), f(x)) \\ &= \nabla_{x'} (f(x)^\top \log f(x')) \\ &= \sum_{i \in [L]} f(x)_i \nabla_{x'} \log(f(x')_i) \\ &= \sum_{i \in [L]} f(x)_i \nabla_{x'} \log \left( \frac{\exp(W_i^\top z')}{\sum_{j \in [L]} \exp(W_j^\top z')} \right) \\ &= \sum_{i \in [L]} f(x)_i \nabla_{x'} \left( W_i^\top z' - \log \left( \sum_{j \in [L]} \exp(W_j^\top z') \right) \right) \\ &= \sum_{i \in [L]} f(x)_i \left( \nabla_{x'}(W_i^\top z') - \sum_{j \in [L]} f(x')_j \nabla_{x'}(W_j^\top z') \right) \\ &= \sum_{i \in [L]} f(x)_i \left( \sum_{j \neq i} f(x')_j \nabla_{x'}(W_{ij}^\top z') \right) \\ &= \sum_{i \neq j} f(x)_i f(x')_j \nabla_{x'}(W_{ij}^\top z'). \end{aligned}$$

Specially, when  $f(x) = 1_y$ , we can obtain based on the above formulas that

$$\nabla_{x'} \mathcal{L}_{CE}(f(x'), y) = - \sum_{l \neq y} f(x')_l \nabla_{x'}(W_{yl}^\top z').$$

$\square$

## B Related work

In this section, we extensively introduce the related work in the adversarial setting, including the adversarial threat models (Sec. B.1), the adversarial attacks (Sec. B.2), the adversarial training strategy (Sec. B.3), and some recent work on combining metric learning with adversarial training (Sec. B.4).

### B.1 Adversarial threat models

Now we introduce different threat models in the adversarial setting following the suggestions in Carlini et al. [9]. Specifically, a threat model includes a set of assumptions about the adversary’s goals, capabilities, and knowledge.

Adversary’s goals could be simply fooling the classifiers to misclassify, which is referred to as *untargeted mode*. On the other hand, the goals can be more aggressive to make the model misclassify from a source class into a target class, which is referred to as *targeted mode*.

Adversary’s capabilities describe the constraints imposed on the attackers. For the  $\ell_p$  bounded threat models, adversarial examples require the perturbation  $\delta$  to be bounded by a preset threshold  $\epsilon$  under  $\ell_p$ -norm, i.e.,  $\|\delta\|_p \leq \epsilon$ .

Adversary’s knowledge tells what knowledge the adversary is assumed to own. Typically, there are four settings when evaluating a defense method:

- *Oblivious adversaries* are not aware of the existence of the defense  $D$  and generate adversarial examples based on the unsecured classification model  $F$  [8].
- *White-box adversaries* know the scheme and parameters of  $D$ , and can design adaptive methods to attack both the model  $F$  and the defense  $D$  simultaneously [2].
- *Black-box adversaries* have no access to the parameters of the defense  $D$  or the model  $F$  with varying degrees of black-box access [17].
- *General-purpose adversaries* apply general transformations or corruptions on the images, which are related to traditional research topics on the input invariances [27, 91].

### B.2 Adversarial attacks

Below we show the details of the attack methods that we test on in our experiments. For clarity, we only introduce the untargeted attacks. The descriptions below mainly adopt from Dong et al. [18].

**FGSM** [24] generates an untargeted adversarial example under the  $\ell_\infty$  norm as

$$\mathbf{x}^{adv} = \mathbf{x} + \epsilon \cdot \text{sign}(\nabla_{\mathbf{x}} \mathcal{L}_{CE}(\mathbf{x}, y)). \quad (17)$$

**BIM** [35] extends FGSM by iteratively taking multiple small gradient updates as

$$\mathbf{x}_{t+1}^{adv} = \text{clip}_{\mathbf{x}, \epsilon}(\mathbf{x}_t^{adv} + \eta \cdot \text{sign}(\nabla_{\mathbf{x}} \mathcal{L}_{CE}(\mathbf{x}_t^{adv}, y))), \quad (18)$$

where  $\text{clip}_{\mathbf{x}, \epsilon}$  projects the adversarial example to satisfy the  $\ell_\infty$  constrain and  $\eta$  is the step size.

**PGD** [44] is similar to BIM except that the initial point  $\mathbf{x}_0^{adv}$  is uniformly sampled from the neighborhood around the clean input  $x$ , which can cover wider diversity of the adversarial space [78].

**MIM** [17] integrates a momentum term into BIM with the decay factor  $\mu = 1.0$  as

$$\mathbf{g}_{t+1} = \mu \cdot \mathbf{g}_t + \frac{\nabla_{\mathbf{x}} \mathcal{L}_{CE}(\mathbf{x}_t^{adv}, y)}{\|\nabla_{\mathbf{x}} \mathcal{L}_{CE}(\mathbf{x}_t^{adv}, y)\|_1}, \quad (19)$$

where the adversarial examples are updated by

$$\mathbf{x}_{t+1}^{adv} = \text{clip}_{\mathbf{x}, \epsilon}(\mathbf{x}_t^{adv} + \alpha \cdot \text{sign}(\mathbf{g}_{t+1})). \quad (20)$$

MIM has good performance as a transfer-based attack, which won the NeurIPS 2017 Adversarial Competition [37]. We set the step size  $\eta$  and the number of iterations identical to those in BIM.

**DeepFool** [46] is also an iterative attack method, which generates an adversarial example on the decision boundary of a classifier with the minimum perturbation. We set the maximum number of

iterations as 100 in DeepFool, and it will early stop when the solution at an intermediate iteration is already adversarial.

**C&W** [7] is a powerful optimization-based attack method, which generates an  $\ell_2$  adversarial example  $\mathbf{x}^{adv}$  by solving

$$\begin{aligned} \arg \min_{\mathbf{x}'} & \left\{ c \cdot \max (Z(\mathbf{x}')_y - \max_{i \neq y} Z(\mathbf{x}')_i, 0) \right. \\ & \left. + \|\mathbf{x}' - \mathbf{x}\|_2^2 \right\}, \end{aligned} \quad (21)$$

where  $Z(\mathbf{x}')$  is the logit output of the classifier and  $c$  is a constant. This optimization problem is solved by an Adam [32] optimizer.  $c$  is found by binary search. The C&W attack can also be applied under the  $\ell_\infty$  threat model with the adversarial loss function  $\max(Z(\mathbf{x}')_y - \max_{i \neq y} Z(\mathbf{x}')_i, 0)$ , using the iterative crafting process.

**ZOO** [12] has been proposed to optimize Eq. (21) in the black-box manner through queries. It estimates the gradient at each coordinate as

$$\hat{g}_i = \frac{\mathcal{L}(\mathbf{x} + \sigma \mathbf{e}_i, y) - \mathcal{L}(\mathbf{x} - \sigma \mathbf{e}_i, y)}{2\sigma} \approx \frac{\partial \mathcal{L}(\mathbf{x}, y)}{\partial x_i}, \quad (22)$$

where  $\mathcal{L}$  is the objective in Eq. (21),  $\sigma$  is a small constant, and  $\mathbf{e}_i$  is the  $i$ -th unit basis vector. In our experiments, we perform one update with  $\hat{g}_i$  at one randomly sampled coordinate. We set  $\sigma = 10^{-4}$  and max queries as 20,000.

**NES** [29] and **SPSA** [72] adopt the update rule in Eq. (18) for adversarial example generation. Although the true gradient is unavailable, NES and SPSA give the full gradient estimation as

$$\hat{\mathbf{g}} = \frac{1}{q} \sum_{i=1}^q \frac{\mathcal{J}(\mathbf{x} + \sigma \mathbf{u}_i, y) - \mathcal{J}(\mathbf{x} - \sigma \mathbf{u}_i, y)}{2\sigma} \cdot \mathbf{u}_i, \quad (23)$$

where we use  $\mathcal{J}(\mathbf{x}, y) = Z(\mathbf{x})_y - \max_{i \neq y} Z(\mathbf{x})_i$  instead of the cross-entropy loss,  $\{\mathbf{u}_i\}_{i=1}^q$  are the random vectors sampled from a Gaussian distribution in NES, and a Rademacher distribution in SPSA. We set  $\sigma = 0.001$  and  $q = 128$  in our experiments, as default in the original papers.

### B.3 Adversarial training

Adversarial training (AT) is one of the most effective strategies on defending adversarial attacks, which dominates the winner solutions in recent adversarial defense competitions [6, 37]. The AT strategy stems from the seminal work of Goodfellow et al. [24], where the authors propose to craft adversarial examples with FGSM and augment them into the training data batch in a mixed manner, i.e., each mini-batch of training data consists of a mixture of clean and crafted adversarial samples. However, FGSM-based AT was shown to be vulnerable under multi-step attacks, where Wong et al. [78] later verify that random initialization is critical for the success of FGSM-based AT. Recent work also tries to solve the degeneration problem of one-step AT by adding regularizers [73]. Another well-known AT strategy using the mixed mini-batch manner is ALP [31], which regularizes the distance between the clean logits and the adversarial ones. But later Engstrom et al. [20] successfully evade the models trained by ALP. As to the mixed mini-batch AT, Xie and Yuille [81] show that using an auxiliary batch normalization for the adversarial part in the data batch can improve the performance of the trained models.

Among the proposed AT frameworks, the most popular one is the PGD-AT [44], which formulates the adversarial training procedure as a min-max problem. Zhang et al. [90] propose the TRADES framework to further enhance the model robustness by an additional regularizer between model predictions, which achieves state-of-the-art performance in the adversarial competition of NeurIPS 2018 [6]. However, multi-step AT usually causes high computation burden, where training a robust model on ImageNet requires tens of GPU workers in parallel [36, 82]. To reduce the computational cost, Shafahi et al. [64] propose the FreeAT strategy to reuse the back-propagation result for crafting the next adversarial perturbation, which facilitate training robust models on ImageNet with four GPUs running for two days.

### B.4 Metric learning + adversarial training

Previous work finds that the adversarial attack would cause the internal representation to shift closer to the "false" class [45, 38]. Based on this observation, they propose to introduce an extra triplet loss

term in the training objective to capture the stable metric space representation, formulated as

$$\begin{aligned} & \mathcal{L}_{\text{trip}}(z(x^*), z(x_p), z(x_n)) \\ &= [D(z(x^*), z(x_p)) - D(z(x^*), z(x_n)) + \alpha]^+, \end{aligned} \quad (24)$$

where  $\alpha > 0$  is a hyperparameter for margin,  $x^*$  (anchor example) is an adversarial counterpart based on the clean input  $x$ ,  $x_p$  (positive example) is a clean image from the same class of  $x$ ;  $x_n$  (negative example) is a clean image from a different class. Here  $D(u, v)$  is a distance function. Mao et al. [45] employ an angular distance as  $D(u, v) = 1 - \cos \angle(u, v)$ ; Li et al. [38] apply the  $\ell_\infty$  distance as  $D(u, v) = \|u - v\|_\infty$ . In the implementation, these methods apply some heuristic strategies to sample triplets, in order to alleviate high computation overhead. For example, Mao et al. [45] select the closest sample in a mini-batch as an approximation to the semi-hard negative example. However, the optimization on sampled triplets is still computationally expensive and could introduce class biases on unbalanced datasets [28].

Zhang and Wang [89] apply a feature-scatter solver for the inner maximization problem of AT, which is different from PGD. Instead of crafting each adversarial example based on its clean counterpart, the feature-scatter solver generate the adversarial examples in batch to utilize inter-sample interactions, via maximizing the optimal transport (OT) distance between the clean and adversarial empirical distributions. In the implementation, they use practical OT-solvers to calculate the OT distance and maximize it w.r.t. the adversarial examples. However, the calculation of the OT distance will increase the computational burden for the AT procedure. Besides, the feature-scatter solver also leads to potential threats for the trained models to be evaded by adaptive attacks, e.g., feature attacks, as discussed before<sup>34</sup>.

## C More empirical results

In this section, we provide more empirical results and setups. In our experiments, we apply NVIDIA P100 / 2080Ti GPUs, as well as the Apex package to execute training for FastAT [13, 78]. On CIFAR-10, all the models are trained by four GPUs in parallel for PGD-AT, ALP, and TRADES.

### C.1 Code references

To ensure that our experiments perform fair comparison with previous work, we largely adopt the public codes and make minimal modifications on them to run the trials. Specifically, we refer to the codes of TRADES<sup>5</sup> [90], FreeAT<sup>6</sup> [64], FastAT<sup>7</sup> [78] and the corrupted datasets<sup>8</sup> from Hendrycks and Dietterich [26]. The codes are mostly based on PyTorch [55].

### C.2 Datasets

The CIFAR-10 dataset [34] consists of 60,000 32x32 colour images in 10 classes, with 6,000 images per class. There are 50,000 training images and 10,000 test images. We perform RandomCrop with 4 padding and RandomHorizontalFlip in training as the data augmentation. The ImageNet (ILSVRC 2012) dataset [16] consists of 1.28 million training images and 50,000 validation images in 1,000 classes. As to the data augmentation, we perform RandomResizedCrop and RandomHorizontalFlip in training; Resize and CenterCrop in test. The image size is 256 and the crop size is 224.

### C.3 Extensive ablation studies

Different choices of the scale  $s$  and the margin  $m$  in HE lead to different trade-offs between the clean accuracy and the adversarial robustness of the trained models, as shown in Table 9. This kind of trade-off is ubiquitous w.r.t. the hyperparameter settings in different AT frameworks [31, 90].

<sup>3</sup><https://github.com/Line290/FeatureAttack>

<sup>4</sup><https://openreview.net/forum?id=Syejj0NYvr&noteId=rkeBhuBMjS>

<sup>5</sup><https://github.com/yaodongyu/TRADES>

<sup>6</sup><https://github.com/mahyarnajibi/FreeAdversarialTraining>

<sup>7</sup>[https://github.com/locuslab/fast\\_adversarial](https://github.com/locuslab/fast_adversarial)

<sup>8</sup><https://github.com/hendrycks/robustness>

Table 9: Classification accuracy (%) on **CIFAR-10**. The training framework is TRADES + HE with different scale  $s$  and margin  $m$ . We report the performance on clean inputs and under PGD-20 attack.

Defense	Scale $s$	Margin $m$	Clean	PGD-20
TRADES + HE	15	0.0	82.53	60.35
	15	0.1	<b>85.00</b>	61.13
	15	0.2	84.88	<b>62.02</b>
	15	0.3	82.99	61.54
	15	0.4	78.05	58.05
	15	0.5	74.71	56.27
	1	0.2	<b>89.34</b>	50.33
	5	0.2	85.70	58.75
	10	0.2	85.30	60.17
	15	0.2	84.88	<b>62.02</b>
	20	0.2	77.67	57.51

#### C.4 Transfer-based black-box attacks

Due to the adversarial transferability [52, 53], the black-box adversaries can construct adversarial examples based on the substitute models and then feed these examples to evade the original models. In our experiments, we apply PGD-AT, ALP, and TRADES to train the substitute models, respectively. To generate adversarial perturbations, we employ the untargeted PGD-20 [44] and MIM-20 [17] attacks, where the MIM attack won both the targeted and untargeted attacking tracking in the adversarial competition of NeurIPS 2017 [37]. In Fig. 4, we show the results of transfer-based attacks against the defense models trained without or with the HE mechanism. As expected, we can see that applying HE can also better defend transfer-based attacks.

Figure 4: Classification accuracy (%) under the *black-box* transfer-based attacks on **CIFAR-10**. The substitute models are PGD-AT, ALP and TRADES separately. \* indicates white-box cases.

PGD-20						MIM-20								
TRADES	PGD-AT	53.93*	66.30	65.83	67.49	68.44	66.58	TRADES	55.08*	66.25	65.86	67.39	68.35	66.64
	ALP	66.17	52.26*	66.12	67.95	69.58	67.42		66.02	53.34*	66.00	67.88	69.24	67.42
	PGD-AT	66.33	66.29	56.48*	67.21	68.89	65.90		66.45	66.51	57.14*	67.32	68.83	66.04
PGD-AT	ALP	TRADES	PGD-AT <sub>+HE</sub>	ALP <sub>+HE</sub>	TRADES <sub>+HE</sub>	PGD-AT	ALP	TRADES	PGD-AT <sub>+HE</sub>	ALP <sub>+HE</sub>	TRADES <sub>+HE</sub>	PGD-AT	ALP	TRADES

#### C.5 Full results of m-HE on CIFAR-10

In Table 10, we evaluate the white-box performance of the combinations of the modified HE (m-HE) with PGD-AT, ALP, and TRADES. We set the parameters with  $s = 15$  and  $m = 0.1$ . We can see that m-HE is more effective than HE when combining with PGD-AT, FreeAT and Fast AT that exclusively train on adversarial examples. In contrast, HE performs better than m-HE when combining with the frameworks training on the mixture of clean and adversarial examples, e.g., ALP and TRADES.

Table 10: Classification accuracy (%) on **CIFAR-10** under the *white-box* threat model. The perturbation  $\epsilon = 0.031$ , step size  $\eta = 0.003$ , following the setting in Zhang et al. [90].

Defense	Clean	PGD-20	PGD-500	MIM-20	FGSM	DeepFool	C&W- $\ell_\infty$
PGD-AT	<b>86.75</b>	53.97	51.63	55.08	59.70	57.26	84.00
PGD-AT + HE	86.19	59.36	57.59	60.19	<b>63.77</b>	<b>61.56</b>	<b>84.07</b>
PGD-AT + m-HE	86.25	<b>59.90</b>	<b>58.46</b>	<b>60.50</b>	63.70	59.47	83.71
ALP	87.18	52.29	50.13	53.35	58.99	59.40	84.96
ALP + HE	<b>89.91</b>	<b>57.69</b>	51.78	<b>58.63</b>	<b>65.08</b>	<b>65.19</b>	<b>87.86</b>
ALP + m-HE	89.23	57.09	<b>53.34</b>	58.04	63.81	60.74	87.21
TRADES	84.62	56.48	54.84	57.14	61.02	<b>60.70</b>	81.13
TRADES + HE	<b>84.88</b>	<b>62.02</b>	<b>60.75</b>	<b>62.71</b>	<b>65.69</b>	60.48	<b>81.44</b>
TRADES + m-HE	84.30	61.83	60.43	62.67	65.49	60.51	80.53

### C.6 Full results on CIFAR-10-C and ImageNet-C

In Table 11 and Table 12 we provide the full classification accuracy results of different defenses on CIFAR-10-C and ImageNet-C [26], respectively. These reports include detailed accuracy under 75 combinations of severity and corruption.

Table 11: Classification accuracy (%) on **CIFAR-10-C**. Full results on different combination of severity and corruption. Here 'S' refers to the severity from 1 to 5, 'P' refers to PGD-AT, 'A' refers to ALP, 'T' refers to TRADES.

Defense	S	Noise			Blur				Weather				Digital			
		Gauss	Shot	Impulse	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Bright	Contra	Elastic	Pixel	JPEG
P	1	85.97	86.3	83.55	86.15	81.92	83.65	82.86	86.12	84.88	84.27	87.15	82.37	82.06	86.23	85.17
	2	84.16	85.82	79.92	84.8	82.06	80.09	82.49	84.4	80.84	76.22	86.68	59.77	82.2	85.62	84.64
	3	81.22	83.08	76.79	82.87	81.55	76.32	81.12	82.39	75.24	65.71	85.7	40.64	81.13	85.33	84.41
	4	79.4	81.2	69.55	80.67	76.18	76.39	80.17	78.03	75.93	51.96	83.87	23.09	79.65	84.27	83.86
	5	77.48	78.02	62.8	74.7	76.72	71.66	77.77	75.69	72.9	30.37	76.86	16.9	78.95	82.37	83.69
P+HE	1	85.37	85.67	83.7	85.63	81.51	83.75	82.67	85.33	84.26	84.1	86.45	82.43	81.69	85.43	85.05
	2	83.69	84.99	80.97	84.5	81.76	80.66	82.33	83.99	80.33	76.84	86.0	61.36	81.92	85.15	84.52
	3	81.14	82.46	78.01	82.65	81.2	76.73	81.11	81.43	75.71	66.59	85.17	41.76	81.2	84.63	84.05
	4	79.74	81.25	71.67	80.83	76.14	77.03	80.09	77.77	76.35	52.75	83.31	23.65	80.01	83.77	83.67
	5	77.76	77.99	66.37	75.58	76.67	72.68	78.16	75.35	73.28	31.96	77.44	16.36	78.61	81.84	83.23
A	1	86.57	86.93	84.13	86.53	83.04	84.12	83.07	86.4	85.77	85.12	87.4	83.21	82.18	86.46	85.79
	2	84.96	86.31	80.48	85.01	82.65	81.04	82.73	84.86	82.19	77.27	87.24	61.34	82.11	86.02	85.09
	3	81.84	83.55	77.23	82.84	81.83	76.91	81.15	82.76	76.57	65.68	86.36	41.96	81.32	85.73	84.63
	4	80.05	81.64	69.53	80.65	77.1	77.09	80.17	78.56	77.54	51.09	84.74	24.92	80.13	84.46	84.23
	5	78.01	78.6	62.97	74.69	76.96	71.99	77.75	76.13	74.26	28.39	78.66	17.85	78.81	82.79	83.68
A+HE	1	88.61	89.27	85.56	89.58	83.58	87.05	86.36	88.71	88.62	88.2	90.1	86.55	85.93	89.37	88.52
	2	85.89	88.25	81.13	88.15	83.79	83.83	85.88	87.0	86.2	82.37	89.9	69.33	85.97	88.72	87.62
	3	81.61	83.99	76.29	86.14	83.69	79.81	84.68	85.71	82.47	73.58	89.13	51.2	84.94	87.98	86.99
	4	78.72	81.42	67.61	84.08	75.73	80.34	83.73	82.52	82.79	61.5	88.06	30.06	83.36	86.92	86.89
	5	76.26	76.64	60.84	78.21	77.56	75.28	81.5	81.76	79.36	38.48	84.46	16.58	82.01	84.19	86.37
T	1	83.74	84.16	81.61	84.01	79.97	81.98	80.69	84.16	83.57	82.53	85.21	79.91	79.47	84.22	83.27
	2	81.84	83.44	78.34	82.61	79.85	78.62	80.04	82.96	79.7	74.42	84.78	57.63	79.8	83.45	82.64
	3	78.63	80.56	74.84	80.58	79.55	75.09	78.9	80.79	73.97	63.06	83.78	39.34	79.05	83.07	82.32
	4	77.09	78.42	67.85	78.62	74.7	75.0	77.9	76.33	75.08	49.91	82.13	24.6	77.48	82.2	81.83
	5	74.8	75.27	61.88	73.4	74.55	71.03	75.78	73.39	72.41	28.5	74.41	17.54	76.86	80.31	81.53
T+HE	1	83.06	83.78	81.12	83.96	79.42	82.13	81.28	83.69	83.22	82.31	85.0	80.2	80.35	83.84	83.24
	2	81.18	83.02	78.41	82.89	79.75	79.39	81.03	82.23	79.21	75.12	84.93	60.38	80.14	83.23	82.64
	3	78.43	80.11	75.4	81.35	79.53	76.19	79.9	80.13	73.93	65.6	83.87	41.97	79.88	83.12	82.37
	4	76.85	78.63	69.59	79.78	74.03	76.62	78.95	76.34	74.51	52.36	82.2	25.59	78.33	82.16	82.07
	5	74.89	75.55	64.55	74.77	75.32	71.98	76.92	73.74	71.79	31.56	75.64	17.07	77.74	80.17	81.66

Table 12: Classification accuracy (%) on **ImageNet-C**. Full results on different combination of severity and corruption. Here 'S' refers to the severity from 1 to 5, 'F' refers to FreeAT.

Defense	S	Noise			Blur				Weather				Digital			
		Gauss	Shot	Impulse	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Bright	Contra	Elastic	Pixel	JPEG
F	1	54.26	53.23	44.00	33.91	42.25	43.97	38.64	42.57	45.83	11.00	56.42	18.29	48.45	53.91	54.74
	2	45.84	43.08	32.60	26.68	34.06	34.70	33.36	26.16	28.50	4.41	53.20	6.59	31.78	52.98	53.92
	3	29.74	28.36	23.48	16.82	25.22	23.83	26.93	22.69	17.22	1.54	47.98	1.38	49.64	50.19	53.34
	4	12.96	10.66	8.91	11.10	19.35	15.20	23.52	11.98	15.51	1.20	39.56	0.40	46.09	45.24	51.57
	5	3.28	4.85	2.73	7.25	12.29	11.05	18.84	11.75	10.28	0.43	28.73	0.34	32.82	41.55	49.19
F+HE	1	55.14	53.27	44.29	38.15	46.08	47.57	41.90	45.92	50.32	15.25	58.73	23.31	51.15	56.17	57.03
	2	43.96	40.02	29.08	30.34	37.97	38.39	36.21	30.21	34.26	6.59	56.47	9.23	34.61	55.40	56.23
	3	24.49	23.14	18.38	18.52	27.98	26.44	29.38	26.83	22.34	2.39	52.39	1.71	52.64	52.65	55.77
	4	9.37	8.37	5.89	11.71	21.45	16.42	25.61	15.22	20.43	1.85	45.56	0.47	48.94	47.82	54.17
	5	2.23	3.95	1.63	7.09	12.93	11.59	20.68	14.89	14.36	0.58	36.21	0.41	36.25	43.81	52.04