Journal of Computers Vol. 34 No. 1, February 2023, pp. 29-43
doi: 10.53106/199115992023023401003

Improving Adversarial Robustness via Finding Flat Minimum of the Weight Loss Landscape

Jiale Yan, Yang Xu*, Sicong Zhang, Kezi Li, Xiaoyao Xie

Key Laboratory of Information and Computing Science of Guizhou Province, Guizhou Normal University, Guiyang 550001, China
xy@gznu.edu.cn

Received 27 March 2022; Revised 30 June 2022; Accepted 6 July 2022

Abstract. Recent studies have shown that robust overfitting and robust generalization gap are a major trouble in adversarial training of deep neural networks. These interesting problems of robust overfitting and robust generalization gap motivate us to explore more solutions. Inspired by recent research on the idea of smoothness, this paper introduces the latest research work on the Adversarial Model Perturbation (AMP) method of finding the flatter minimum of the weight loss landscape into the adversarial training (AT) framework of deep neural networks to alleviate the robust overfitting and robust generalization gap troubles, called AT-AMP method. The validity of the flat minimum is explained from the perspective of statistical generalization theory. Although the idea is plain, this approach is surprisingly effective. Experiments demonstrate that by incorporating the AMP method into adversarial training framework, we can boost the robust accuracy by 1.14% ~ 5.73%, on three different benchmark datasets SVHN, CIFAR-10, CIFAR-100 and two threat models L_∞ norm constraint and L_2 norm constraint, across diverse types of adversarial training framework such as AT, TRADES, MART, AT with pre-training and RST and diverse white-box and black-box attack, achieving the state-of-the-art performance in adversarial training framework. In addition, we compare several classical regularization and modern deep learning data augmentation tricks for robust overfitting and robust generalization with the AMP method, and the experimental research results consistently indicate that introducing the AMP method achieves advanced adversarial robustness in the adversarial training framework.

Keywords: adversarial example, adversarial training, adversarial robustness, deep neural networks

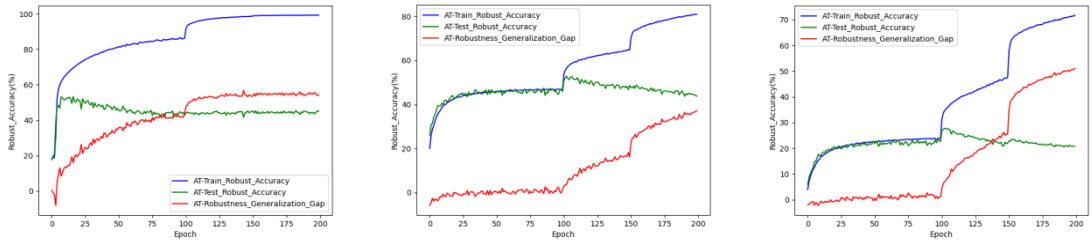
1 Introduction

As the core technology of artificial intelligence, deep neural networks (DNNs) have been extensively used in the majority of scenes and applications. They have achieved state-of-the-art performance in many tasks such as computer vision [1], natural language processing [2], speech recognition [3], autonomous driving [4], medical diagnosis [5], and even surpass human processing ability in some fields. However, it is found that the DNNs are easily fooled by adding human-imperceptible small perturbations to the normal input examples (known as adversarial examples) [6-7], resulting in wrong output, which brings tremendous challenges to the application of the DNNs in security-sensitive systems [4-5]. As the DNNs model is widely used, it is almost everywhere in daily life. Therefore, how to construct a more secure, reliable and robust DNNs model, such as improving model robustness against adversarial examples, becomes more and more urgent.

So far, there have been many defense techniques to improve the adversarial robustness of DNNs [8-11]. Among these defense methods, Aleksander Madry et al. proposed projected gradient descent (PGD) adversarial training (AT) [12] is recognized as the most effective and promising defense method, which trains DNNs to minimize the training loss under the worst input perturbation. Although AT has achieved a certain degree of adversarial robustness, its robustness is far from satisfactory due to the enormous robust generalization gap [13-14]. For instance, on CIFAR-100 dataset [15], PreAct ResNet-18 [16] using PGD-AT under the L_∞ norm constraint achieved 71% training robustness accuracy, but only achieved 27% test robustness accuracy after 200 epochs, as shown in Fig. 1(c). Surprisingly, this gap in robust generalization is as high as 44%, which is completely distinct from the DNNs standard training on normal examples. The standard generalization gap is usually less than 10%. In addition, the recent research by Leslie Rice et al. [13] shows that the AT of DNNs has a property, that is, "robust overfitting" is a dominant phenomenon. Robust overfitting is an important factor that leads to unsatisfactory ad-

* Corresponding Author

versarial robustness of DNNs model. Therefore, how to solve the robust generalization gap and robust overfitting is the key way to go a step further improve the robustness of adversarial training methods.



(a) SVHN (b) CIFAR-10 (c) CIFAR-100

Fig. 1. The curve of the accuracy of PGD-AT on three different benchmark datasets SVHN, CIFAR-10 and CIFAR-100 under the L_∞ norm constraint ($\varepsilon = 8/255$) applying PreAct ResNet-18 model for 200 epochs
(Blue, Green, and Red curves represent the accuracy of train robust, test robust, and robustness generalization gap respectively)

Recalling that weight loss landscape is a commonly used approach to represent the standard generalization gap in standard training scenarios [17-18], however, there have been few investigations into adversarial training framework, among which Prabhu et al. [19] and Yu et al. [20] attempted to use the pre-generated adversarial examples to investigate but were unable to reach the desired conclusions. In this article, we investigate the weight loss landscape under adversarial training framework using on-the-fly produced adversarial examples, and we verify a strong relationship between the flatness of the weight loss landscape and robust generalization gap as well as robust overfitting. A few well-known adversarial training framework refinements, such as AT with pre-training [13], TRADES [21], MART [11], and RST [22], all implicitly flatten the weight loss landscape to alleviate the robust generalization gap and robust overfitting. Motivated by this, we propose to incorporate an explicitly flatten weight loss landscape strategy, Adversarial Model Perturbation (AMP) [23], into adversarial training framework, which directly bounds the flatness of the weight loss landscape to facilitate the DNNs model to find a more flat minimum. Unlike random perturbations [24], the AMP technique can infuse the strongest worst-case weight perturbations, establishing a double perturbation mechanism such as inputs and weights parameter are both adversarially perturbed in the adversarial training framework. The explicit flatness of weight loss landscape AMP approach is universal and can be conveniently introduced into existing improved adversarial training framework method with small computational expense. We have carried out a lot of experimental comparisons, and confirmed that AMP method has effectively improved the adversarial robustness of DNNs models in the adversarial training framework.

The main contributions of this paper are as follows:

(1) We verify the fact that flatter weight loss landscape to facilitate the DNNs model to find a more flat minimum often contributes to smaller robust generalization gap and robust overfitting in adversarial training framework utilizing on-the-fly produced adversarial examples.

(2) We propose to incorporate an explicitly flatten weight loss landscape Adversarial Model Perturbation (AMP) strategy into adversarial training framework, establishing a double perturbation mechanism that infuses the worst-case input and weight parameter perturbations, which directly bounds the flatness of the weight loss landscape to facilitate the DNNs model to find a more flat minimum.

(3) Experiments demonstrate that by incorporating the AMP strategy into adversarial training framework, it can boost the robust accuracy by 1.14% ~ 5.73%, across diverse types of adversarial training approach such as AT, TRADES, MART, AT with pre-training and RST, three different datasets, two threat models and diverse white-box and black-box attacks, achieving the state-of-the-art performance in adversarial training framework. In addition, we explore the optimal hyperparameter for using the AMP method. We compare several classical and modern deep learning tricks for robust overfitting and robust generalization, including regularization and data augmentation, with the AMP method and the experimental research results consistently indicate that AMP acquires advanced adversarial robustness.

The rest of this paper is organized as follows. Section 2 introduces a review of related work. In Section 3, a method is proposed to improve model adversarial robustness in the adversarial training framework. Section 4 presents the experiment setup, experiment results and experiment effect discussion. Finally, Section 5 concludes the paper.

2 Related Work

In this part, we will briefly review these recent adversarial example defense methods and robust generalization and robust overfitting problems in adversarial training.

2.1 Adversarial Defense

A series of previous works have proposed many defense methods to defend adversarial attacks, such as defensive distillation [8], feature denoising [25], input denoising [26], adversarial detection [27], gradient regularization [28], gradient masking [29], model compression [30], activation pruning [31], adversarial training [6-7, 12] and so on. However, many defense methods either provide little improvement in robustness or have been evaded by new attack methods. At present, one of the recognized relatively effective defense approach is AT, which has not been completely attacked up to now and has relatively excellent adversarial robustness [32]. A variety of additional strategies are presented based on adversarial training to improve its performance even more. A brief review is as follows:

AT [12] formalize a min-max problem to seek the model parameter θ to minimize the adversarial loss, and adversarial training approach solves the following optimization problems to improve the robustness of the DNNs model.

$$\min_{\theta} \rho(\theta) \text{ , where } \rho(\theta) = \frac{1}{n}\sum_{i=1}^{n} \max_{\|x_i^{'} - x_i\|_p \le \varepsilon} \ell(f_\theta(x_i^{'}), y_i) \text{ ,} \tag{1}$$

where f_θ is DNNs with parameter θ, n is the quantity of training examples, $x_i^{'}$ is the adversarial example generated under the L_p norm constraint, $\ell(*)$ is the standard loss function, and $\rho(\theta)$ is the adversarial loss.

TRADES [21]. TRADES optimizes an upper bound of adversarial risk that is a trade-off between accuracy and robustness:

$$\rho^{TRADES}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \{CE(f_\theta(x_i), y_i) + \beta \cdot \max KL(f_\theta(x_i) \| f_\theta(x_i^{'}))\} \text{ ,} \tag{2}$$

where KL denotes the Kullback-Leibler divergence, CE denotes the cross-entropy loss, and β is the hyperparameter that controls the trade-off between natural accuracy and robust accuracy.

MART [11]. As a regularizer of adversarial risk, MART includes an explicit differentiation of misclassified examples:

$$\rho^{MART}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \{BCE(f_\theta(x_i^{'}), y_i) + \lambda \cdot KL(f_\theta(x_i) \| f_\theta(x_i^{'})) \cdot (1 - [f_\theta(x_i)]_{y_i})\} \text{ ,} \tag{3}$$

where $BCE(f_\theta(x_i^{'}), y_i) = -\log([f_\theta(x_i^{'})]_{y_i}) - \log(1 - \max_{k \ne y_i}[f_\theta(x_i^{'})]_k)$, λ is a tunable scaling parameter that balances the two parts of the final loss, KL denotes the Kullback–Leibler divergence, $[f_\theta(x_i)]_{y_i}$ denotes the y_i-th element of output vector $f_\theta(x_i)$.

Semi-Supervised Learning (SSL) [11, 22, 33]. SSL-based approaches make use of additional unlabeled data. They begin by training a natural model on the labeled data to produce pseudo labels for unlabeled data. Then, using both labeled and unlabeled data, adversarial loss $\rho(\theta)$ is used to train a robust model:

$$\rho^{SSL}(\theta) = \rho^{labeled}(\theta) + \lambda \cdot \rho^{unlabeled}(\theta) \text{ ,} \tag{4}$$

where λ is the weight applied to unlabeled data. $\rho^{labeled}(\theta)$ and $\rho^{unlabeled}(\theta)$ generally refer to the same adver-

sarial loss. For example, RST in Carmon et al. [22] uses TRADES loss, whereas semi-supervised MART in Wang et al. [11] employs MART loss.

The above work of adversarial defense is based on promising adversarial training framework. They have introduced some new ideas and techniques to further improve the adversarial robustness of the DNNs model under adversarial training framework. The main difference between our ideas and methods and above work methods is that we propose to incorporate an explicitly flatten weight loss landscape the AMP techniques into adversarial training framework, establishing a double perturbation mechanism that infuses the worst-case input and weight parameter perturbations, which directly bounds the flatness of the weight loss landscape to facilitate the DNNs model to find a more flat minimum. However, the DNNs model to find a more flat minimum often contributes to smaller robust generalization gap and robust overfitting in adversarial training framework. In addition, more importantly, our ideas and methods are not related to specific methods and general, which can be incorporated into the above work improved adversarial training framework to further enhance the adversarial robustness of the corresponding methods.

Unfortunately, the method proposed in this paper has the same shortcomings as the above related work improved methods based on adversarial training framework. Although the adversarial robustness accuracy of the DNNs model has been improved, the natural accuracy of the normal examples will be damaged in some scenarios, and the introduction of the double perturbation mechanism proposed in this paper will increase some computational overhead.

2.2 Robust Generalization and Robust Overfitting

It is more difficult to train a DNN with robust generalization on adversarial examples than standard generalization on normal examples [12], and it needs more training data [14] and has higher examples complexity [34]. Preetum Nakkiran et al. [35] show that the model needs large capacity to become more robust. Dimitris Tsipras et al. [36] and Hongyang Zhang et al. [21] prove that the adversarial robustness accuracy may be inherently incompatible with accuracy on natural examples. Recently, Leslie Rice et al. [13] confirmed that robust overfitting is a crucial problem in adversarial training and suggested early stop as an effective mitigation measure. Furthermore, there is a body of work that investigates robust generalization and robust overfitting from the perspective of the loss landscape. There are two kinds of loss landscape in the adversarial training framework: 1) The input loss landscape, which is the change in loss with regard to the input. It represents the change in loss around training examples. By training on adversarially perturbed examples, AT explicitly flattens the input loss landscape. 2) The weight loss landscape, which is the change in loss with regard to the weight. It discloses the loss landscape geometry in the vicinity of model weights. In contrast to the standard training scenario, where various researches have indicated a correlation between the weight loss landscape and their standard generalization gap and overfitting [18, 37-38], whether the correlation occurs in adversarial training framework is currently being investigated.

Different from these studies, we propose to introduce explicitly flatten weight loss landscape in adversarial training framework, establishing a double perturbation mechanism that infuses input and weight parameter perturbations, which directly bounds the flatness of the weight loss landscape to facilitate the DNNs model to find a more flat minimum. A large number of experimental results identify a more flat minimum often contributes to smaller robust generalization gap and robust overfitting in adversarial training framework.

3 Methods

Recently, some research work focuses on using smoothness to improve robustness [39] or using smoothness to improve the generalization performance of standard training models [23]. Inspired by the smoothness of research work and implicitly flattening the weight loss landscape, we pay attention to explicitly flattening the weight loss landscape in the adversarial training framework, establishing a double perturbation mechanism that infuses input and weight parameter perturbations that directly bound the flatness of the weight loss landscape to facilitate the DNN model to find a more flat minimum. Previous work [38] shows that a flat minimum corresponds to a simple model, which can prevent overfitting. A widely accepted and empirically verified viewpoint is that models trained on normal data corresponding to the flat minimum of the weight loss landscape tends to be better generalized [40-41]. Overall, intuition and experience motivate us to do this research. In this paper, we explore the robust model corresponding to finding the flat minimum of the weight loss landscape under the adversarial training framework, applying on-the-fly generated adversarial examples, and identify a strong correlation between the flat minimum of the weight loss landscape and robust overfitting as well as robust generalization gap. Some research work im-

plicitly flattens the weight loss landscape to alleviate the robust generalization gap and improve robust overfitting. For the first time, we introduce an explicit AMP method to flatten the weight loss landscape in the adversarial training framework, which directly bounds the flatness of the weight loss landscape to facilitate the DNN model to find a more flat minimum. Details are as follows.

Assume that under a classification setting, where the task is to seek a classifier $f_\theta : \mathcal{X} \to \mathcal{Y}$ maps the input space $\mathcal{X}$ to the label space $\mathcal{Y}$, the function f_θ as a deep neural network is parameterized by θ, θ comes from the weight space Θ, and for each training sample $(x, y) \in \mathcal{X} \times \mathcal{Y}$, $\ell(f(x), y; \theta)$ is the loss function (cross-entropy loss).

Under the traditional empirical risk minimization (ERM) criterion [42], a deep neural network is trained by using a training set $\mathcal{D}$ to minimize the empirical risk loss of equation (5), which is abbreviated as ERM loss.

$$\mathcal{L}_{ERM}(\theta) = \frac{1}{|\mathcal{D}|} \sum_{(x,y)\in\mathcal{D}} \ell(f(x), y; \theta) \ . \tag{5}$$

However, it is well known that ERM loss training is subject to overfitting [42], and the learned parameters can never be well generalized on unknown data.

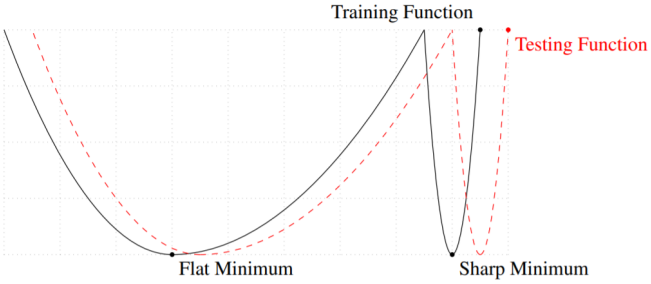


Fig. 2. Two kinds of minima: flat and sharp [38]

The AMP is a state-of-the-art method based on the principle of finding a flat minimum in ERM, which can be proved theoretically to be capable of flattening the weight loss landscape, which directly bounds the flatness of the weight loss landscape to facilitate the DNN model to find a more flat minimum. The AMP method does not minimize the traditional empirical risk loss but minimizes an AMP loss, as shown in Equation (6), which is described in detail below.

$$\mathcal{L}_{AMP}(\theta) = \max_{\Delta\in B(0;\delta)} \frac{1}{|\mathcal{D}|} \sum_{(x,y)\in\mathcal{D}} \ell(f(x), y; \theta + \Delta) \ . \tag{6}$$

For any positive value δ and any $\mu \in \Theta$, define $B(\mu; \delta)$ as a L_2 norm ball with a radius of δ centered on μ in Θ space. Its parameter δ is a small positive value as a hyperparameter, which can be expressed as in equation (7).

$$B(\mu; \delta) = \{\theta \in \Theta : \|\theta - \mu\|_2 \le \delta\} \ . \tag{7}$$

In actual use of AMP, the factors of computation cost and training speed are considered, and then the method of minimizing mini-batch is adopted to minimize AMP loss $\mathcal{L}_{AMP}$. In more detail, a mini-batch $\mathcal{B}$ from a random small batch is adopted to approximate AMP loss in equation (6), as shown in equation (8).

$$\mathcal{L}_{AMP}(\theta) \approx \max_{\Delta_{\mathcal{B}}\in B(0;\delta)} \frac{1}{|\mathcal{B}|} \sum_{(x,y)\in\mathcal{B}} \ell(f(x), y; \theta + \Delta_{\mathcal{B}}) \ . \tag{8}$$

Then, introducing the AMP method into the adversarial training framework to minimize the adversarial loss $\rho(\theta)$ in equation (1) can be approximately formalized as an optimization problem, which is called the AT-AMP method, as shown in equation (9).

$$\min_{\theta_{AT-AMP}^{*}} \rho(\theta) \ , \ \rho(\theta) = \max_{\Delta_{\mathcal{B}} \in B(0;\delta)} \frac{1}{|\mathcal{B}|} \sum_{(x,y)\in\mathcal{B}} \max_{\|x'-x\|_p \le \varepsilon} \ell(f(x'), y; \theta + \Delta_{\mathcal{B}}) \ . \tag{9}$$

Algorithm 1. AT-AMP algorithm process

Symbol description: Training dataset $\mathcal{D} = \{(x, y)\}$, Batch scale m , Loss function ℓ , Initial DNN model parameter θ_0 , Outer learning rate η , Inner learning rate ν and ξ Inner iteration quantity N , L_2 norm δ , Adversarial examples norm constraint ε

while θ_k not converged **do**
  Update epoch: $k \leftarrow k+1$
  Sample mini-batch $\mathcal{B} = \{(x_i, y_i)\}_{i=1}^{m}$ from training dataset $\mathcal{D}$
  Initialize perturbation vector: $\Delta_{\mathcal{B}} \leftarrow \vec{0}$
  for $n \leftarrow 1$ **to** N **do**
    while($\|x_i' - x_i\|_p \le \varepsilon$ and the preset number of iterations is completed**)**
      Compute gradient:

$$a \leftarrow \nabla_x \ (f(x_i), y_i; \theta_k + \Delta_{\mathcal{B}})$$

      Update AE: $x_i' \leftarrow x_i + \upsilon\vec{a}$
    end while
    Compute gradient:

$$\vec{b} \leftarrow \frac{1}{m}\sum_{i=1}^{m} \nabla_\theta \ell(f(x_i'), y_i; \theta_k + \Delta_{\mathcal{B}})$$

    Update perturbation vector: $\Delta_{\mathcal{B}} \leftarrow \Delta_{\mathcal{B}} + \xi\vec{b}$
    if $\|\Delta_{\mathcal{B}}\|_2 > \delta$ **then**
      Normalize perturbation: $\Delta_{\mathcal{B}} \leftarrow \delta \frac{\Delta_{\mathcal{B}}}{\|\Delta_{\mathcal{B}}\|_2}$
    end if
  end for
  Compute gradient:

$$\vec{c} \leftarrow \frac{1}{m}\sum_{i=1}^{m} \nabla_\theta \ell(f(x_i'), y_i; \theta_k + \Delta_{\mathcal{B}})$$

  Update parameter: $\theta_k \leftarrow \theta_k - \eta\vec{c}$
end while

Formalizing the description based on the above formula, each batch $\mathcal{B}$ corresponds to a perturbation vector $\Delta_{\mathcal{B}}$ on the parameter θ. This training involves the maximization of two inner layers and the minimization of an outer layer: The first maximization of the inner layer is used to perturb normal input examples searching for adversarial examples; the second maximization of the inner layer is used to update $\Delta_{\mathcal{B}}$ in the direction of augmenting ERM loss as a way to perturb weight parameter; The minimization of the outer layer loops on random batches and uses mini-batch SGD to minimize the adversarial loss $\rho(\theta)$. The detailed AT-AMP algorithm process is shown in Algorithm 1. It should be noted that the learned θ^*_{AT-AMP} is used to predict as a parameter of the deep neural network model after training, but it is used without perturbation in the testing stage.

The flat minima of neural networks can generalize better than the sharp ones. A convincing reason is that a flat minimum of the training function loss curve can acquire lower generalization loss when the test function loss curve is shifted from the training function loss for random perturbations, as shown in Fig. 2. For the improved adversarial robustness, why is the flatter minimum more work? Our insight is that the adversarial training optimizes the weight loss landscape over the adversarial examples, which are then generated by adding small perturbations to each example as a way to obtain a worst-case, which means that the AT method considers a local worst-case on a sample-by-sample basis but does not cover the overall situation over multiple samples. The reason for the AT-AMP method's working is that it looks for the flatter minimum of the weight loss landscape, which affects the loss of all examples to the extent that a global worst-case at the model level can be obtained.

4 Results and Discussion

In this section, we conduct comprehensive experiments to evaluate the effectiveness of incorporating AMP approach into AT framework establishing double perturbation mechanism including its vanilla AT robustness, benchmarking the state-of-the-art robustness and comparisons to other classical regularization techniques and modern data augmentation techniques. Moreover, we discuss AMP approach in improving the robust generalization gap and the robust overfitting effect.

4.1 Experimental Settings

Datasets. Our experiments were conducted primarily using two threat models (L_∞ norm constraint and L_2 norm constraint) and three different datasets of SVHN [43], CIFAR-10 [15] and CIFAR-100 [15].

Training and Evaluation Details. We default to using PreAct ResNet-18 for most experiments, with the exception of the experiments in Table 2 and Table 3 with the large-capacity network structure WideResNet34-10 [44]. In all training, an SGD optimizer with momentum of 0.9 was used for 200 epochs, and a piece-wise learning rate schedule (in the 100th and 150th epoch, the learning rate is reduced by a factor of 10 and the initial learning rate is 0.1) was used. Simple image data augmentation methods such as 32-size random crop with four pixel size padding and random image horizontal flip tricks are applied. We use two common metrics that are widely used: natural test accuracy and robust test accuracy, which are the classification accuracies on the original and the attacked testsets, respectively.

Attack Methods. For the L_2 threat model, perturbation constraint $\varepsilon = 128/255$, step size 15/255 for all datasets. For L_∞ threat model, perturbation constraint $\varepsilon = 8/255$, step size 2/255 for CIFAR-10 and CIFAR-100 datasets, and step size 1/255 for SVHN datasets. Unless otherwise specified, our default training and test attacks use PGD, which is recognized as one of the strongest algorithms for first-order attacks. Training attack uses PGD-10 and test attack uses PGD-20 or other attacks.

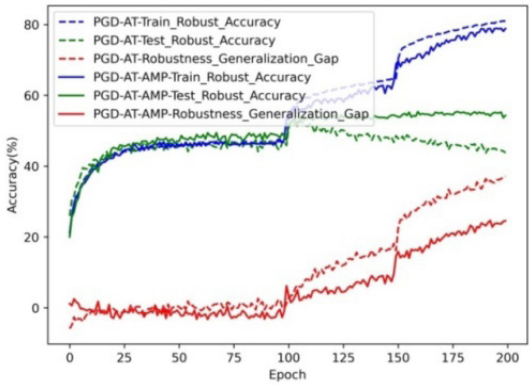


Fig. 3. The curve of the accuracy of PGD-AT and PGD-AT-AMP on CIFAR-10 dataset under the L_∞ norm constraint ($\varepsilon = 8/255$) applying PreAct ResNet-18 model for 200 epochs
(Dash lines show the PGD-AT; solid lines represent the PGD-AT-AMP. Blue, Green and Red curves represent the accuracy of train robust, test robust and robustness generalization gap respectively.)

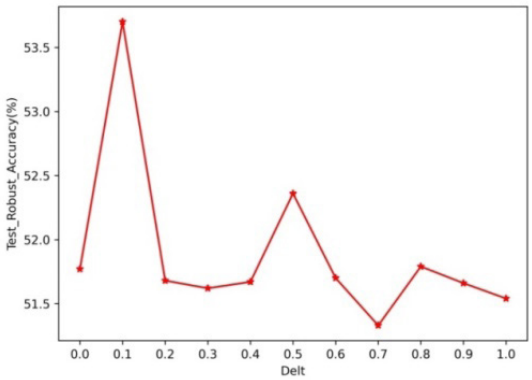


Fig. 4. The curve of the test robust accuracy of PGD-AT-AMP varies with hyperparameters δ on CIFAR-10 dataset applying PreAct ResNet-18 model under the L_∞ norm constraint ($\varepsilon = 8/255$)
(The experimental results show that $\delta = 0.1$ is the best hyperparameters setting.)

4.2 A Case Study on Vanilla AT and AT-AMP

At first, we use PreAct ResNet-18 to carry out PGD-AT-AMP experiments under the L_∞ norm constraint of the CIFAR-10 dataset, in which the hyperparameter of the AMP method follows the best hyperparameter $\delta = 0.5$ in the original paper [23]. For other detailed settings, please refer to Section 4.1 Experimental Settings. It is observed that the model trained by the PGD-AT-AMP method does improve the robust generalization gap (the red solid line is lower than the red dotted line) and the robust overfitting (accuracy curves no longer get worse after more than 100 epochs), as shown in Fig. 3. Therefore, we further explored the influence of the AMP hyperparameter setting on the robust test accuracy and provided some guidance for further experimental exploration. We refer to the original paper [23] and explore several cases of test robust accuracy in the PGD-AT-AMP method with hyperparameters $\delta \in \{0,\ 0.1,\ 0.2,\ 0.3,\ 0.4,\ 0.5,\ 0.6,\ 0.7,\ 0.8,\ 0.9,\ 1\}$, as shown in Fig. 4. The experimental research results show that $\delta = 0.1$ is the best hyperparameter setting, and we will adopt this best hyperparameter setting in the subsequent experiments.

In view of the fact that under the constraint of L_∞ norm of the CIFAR-10 dataset, the PreAct ResNet-18 model trained by the PGD-AT-AMP method further improves the test robustness, so we further experimented under two threat models (i.e. L_∞ and L_2 norm) of two datasets (SVHN and CIFAR-100 dataset) and the CIFAR-10 dataset L_2 norm constraint, and also achieved more advanced test robustness than PGD-AT. The experimental results are shown in Table 1. For detailed analysis and discussion, see sections 4.4 and 4.5.

As the previous research work [12] shows that the network model with a larger capacity is needed to achieve the model's adversarial robustness, we have conducted experiments on the larger network structure WideResNet34-10, and the results show that the PGD-AT-AMP method has been proved to promote the adversarial robustness of the model again. The detailed results are shown in Table 2. For detailed analysis and discussion, see sections 4.4 and 4.5.

Table 1. Accuracy (%) of PGD-AT and PGD-AT-AMP applying PreAct ResNet-18 model on different datasets and threat models (L_∞ norm: $\varepsilon = 8/255$; L_2 norm: $\varepsilon = 128/255$) over 5 random runs

(The best signifies the highest accuracy in the whole epoch while last signifies the accuracy at the end of the 200 epochs. The best results are marked in bold.)

Dataset	Norm	Method	Robustness Accuracy		Natural Accuracy	
			Best	Last	Best	Last
SVHN	L_∞	PGD-AT	53.34 ± 0.07	44.46 ± 0.25	92.14 ± 0.13	89.56 ± 0.37
		PGD-AT-AMP	**58.47 ± 0.18**	**57.38 ± 0.21**	**93.25 ± 0.09**	**92.19 ± 0.34**
	L_2	PGD-AT	66.63 ± 0.26	65.05 ± 0.21	93.34 ± 0.09	93.12 ± 0.24
		PGD-AT-AMP	**72.36 ± 0.32**	**67.86 ± 0.26**	**95.25 ± 0.16**	**94.76 ± 0.11**
CIFAR-10	L_∞	PGD-AT	51.77 ± 0.19	44.36 ± 0.34	**81.68 ± 0.19**	**81.57 ± 0.21**
		PGD-AT-AMP	**53.85 ± 0.32**	**53.26 ± 0.18**	80.36 ± 0.14	80.44 ± 0.11
	L_2	PGD-AT	68.13 ± 0.11	65.90 ± 0.31	89.54 ± 0.06	88.93 ± 0.15
		PGD-AT-AMP	**71.25 ± 0.09**	**71.13 ± 0.08**	**90.03 ± 0.26**	**89.13 ± 0.21**
CIFAR-100	L_∞	PGD-AT	27.26 ± 0.14	20.34 ± 0.20	**56.36 ± 0.16**	**54.85 ± 0.36**
		PGD-AT-AMP	**30.52 ± 0.21**	**29.69 ± 0.18**	53.69 ± 0.26	53.36 ± 0.28
	L_2	PGD-AT	41.39 ± 0.18	35.54 ± 0.27	62.83 ± 0.09	60.34 ± 0.23
		PGD-AT-AMP	**44.93 ± 0.16**	**44.63 ± 0.23**	**63.85 ± 0.22**	**62.51 ± 0.31**

Table 2. Accuracy (%) of PGD-AT and PGD-AT-AMP on WideResNet34-10 across different datasets on L_∞ norm constraint threat models ($\varepsilon = 8/255$) over 5 random runs

(The best signifies the highest accuracy in the whole epoch while last signifies the accuracy at the end of the 200 epochs. The best results are marked in bold.)

Dataset	Norm	Method	Robustness Accuracy		Natural Accuracy	
			Best	Last	Best	Last
CIFAR-10	L_∞	PGD-AT	54.16 ± 0.16	43.52 ± 0.22	**84.16 ± 0.13**	**84.57 ± 0.21**
		PGD-AT-AMP	**55.54 ± 0.19**	**53.89 ± 0.26**	83.89 ± 0.06	83.59 ± 0.12
CIFAR-100	L_∞	PGD-AT	29.95 ± 0.06	24.02 ± 0.09	**56.69 ± 0.16**	**56.23 ± 0.24**
		PGD-AT-AMP	**31.93 ± 0.15**	**31.54 ± 0.11**	55.88 ± 0.13	55.25 ± 0.08

4.3 Benchmarking the State-of-the-art Robustness

In this section, we benchmark the state-of-the-art robustness of our proposed AT-AMP double-perturbation mechanism method against white-box and black-box attacks on CIFAR-10 dataset. Two types of adversarial training framework approaches are discussed: One is only dependent on original data: 1) AT; 2) TRADES; and 3) MART. The other makes advantage of extra data: 1) Pre-training; and 2) RST.

For CIFAR-10 under L_∞ attack with $\varepsilon = 8/255$, we train WideResNet34-10 for AT, TRADES, MART, Pre-training and RST, as described in their original works. For pre-training, we fine-tune 50 epochs using a learning rate of 0.001 as described in [45].

In the white-box attack scenario, we use FGSM, PGD-20, PGD-100 and CW_∞ [46] attacks, where PGD-x denotes the number of iterations using PGD attack as well as CW_∞ attack implemented in the form of PGD-100 using CW loss.

In order to eliminate the suspicion of obfuscated gradients in the proposed method, the black-box attack method of SPSA attack is implemented. In the black-box attack scenario, we use the query-based attack SPSA [47], where for gradient estimation we use a perturbation size of 0.001, step size of 0.01 and a batch size of 256 samples for each gradient estimation.

Finally we also tested a more powerful parameter-free attack method Auto Attack (AA) [48], which contains three white-box attack methods: APGD-CE [48], APGD-DLR [48], FAB [49] and a black-box attack method Square Attack [50]. The great advantage of the AA attack is the use of an ensemble of multiple parameter-free attacks to verify the robustness of the model. Other base hyperparameters of the baselines are configured as per their original paper. The experimental results are shown in Table 3.

The experimental results show that the proposed double perturbation mechanism consistently improves the robustness of improved methods (AT with pre-training, TRADES, MART, and RST) based on the adversarial training framework under various attacks, and also show that the Auto Attack is the strongest attack among several attacks, which also demonstrates the power of the integrated attack.

The results in Table 3 show that the proposed double perturbation mechanism still has robustness improvement under the white-box attack and the black-box attack, thus indicating that the proposed double perturbation mechanism improves the model robustness not due to obfuscated gradients or masking, improper tuning of hyper-parameters of attacks.

Table 3. Robustness accuracy (%) of AT and improved method based on AT with WideResNet34-10 model across different attacks (white-box and black-box attacks) on CIFAR-10 datasets under L_∞ norm constraint threat models ($\varepsilon = 8/255$) over 5 random runs
(Limited to space and all standard deviations are less than 0.4% so they are ignored. The best results are marked in bold.)

Defense	Natural	FGSM	PGD-20	PGD-100	CW_∞	SPSA	AA
PGD-AT	**84.45**	61.55	54.18	53.43	52.32	61.13	51.30
PGD-AT-AMP	83.74	**62.83**	**56.26**	**55.83**	**53.46**	**62.53**	**53.16**
TRADES	82.27	61.26	54.56	54.03	52.36	61.25	52.22
TRADES-AMP	**83.65**	**63.52**	**57.37**	**57.21**	**55.11**	**63.56**	**55.86**
MART	82.19	61.65	56.39	55.86	52.69	59.02	50.23
MART-AMP	**83.24**	**63.83**	**58.69**	**57.43**	**54.42**	**61.96**	**53.49**
Pre-training	85.56	63.15	55.43	54.32	53.67	62.19	53.85
Pre-training-AMP	**86.36**	**65.63**	**59.56**	**58.61**	**57.35**	**64.43**	**56.45**
RST	**87.53**	**69.31**	60.19	60.03	58.63	67.15	58.32
RST-AMP	86.34	67.45	**61.86**	**61.53**	**59.35**	**68.85**	**59.63**

4.4 Effect on Robust Generalization Gap

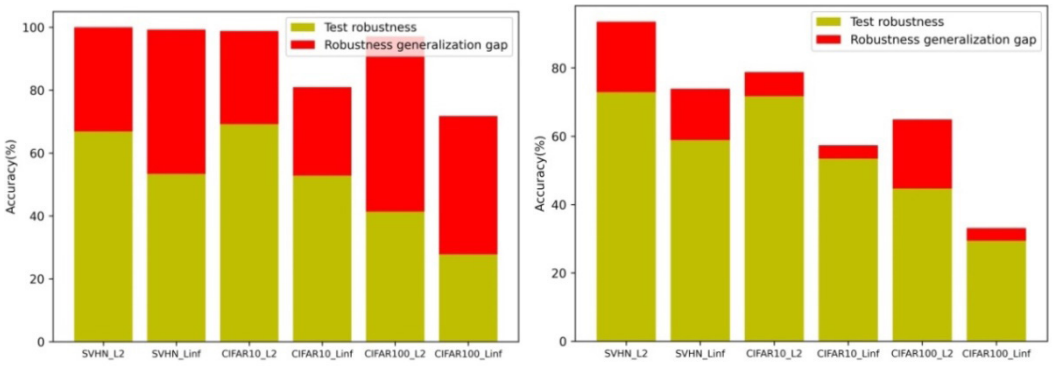
As shown in Fig. 1 for the robust generalization gap in the CIFAR-100 dataset, we used PreAct ResNet-18 for adversarial training under different threat models on the SVHN and CIFAR-10 datasets (with the same parameter settings as Section 4.1), and similar robust generalization gap phenomena were observed. Among them, the CIFAR-10 and CIFAR-100 datasets have similar phenomena where the robust generalization gap becomes larger after the first learning rate drop. However, for the SVHN dataset, unlike the CIFAR-10 and CIFAR-100 datasets, the robust generalization gap becomes significantly larger much earlier, which appears in about 10 epochs. In conclusion, the robust generalization gap issue is across different threat models and different datasets.

Therefore, we conduct experiments to explore whether finding a flatter minimum of the weight loss landscape in adversarial training framework is beneficial to narrow the gap of robust generalization. We use the PGD-AT-AMP method to conduct experiments under three different datasets and two threat models, and the histogram of experimental research results is shown in Fig. 5. The experimental results in Fig. 5 show that finding a flatter minimum of the weight loss landscape in the adversarial training framework does directly result in a smaller robust generalization gap in the training process.

4.5 Effect on Robust Overfitting

The PGD-AT obtains the best test robustness accuracy after the first learning rate drop and begins overfitting on CIFAR-10 and CIFAR-100 datasets. However, overfitting of the SVHN dataset occurs much earlier at around 10 epochs. In comparison, PGD-AT-AMP stays stable and continues to improve test robust accuracy along with more training epochs (In addition to the case on SVHN dataset under the L_2 norm constraint), as shown in Fig. 6. The PGD-AT-AMP method improves the robustness of all datasets (L_2 and L_∞ threat models) but sacrifices the natural accuracy on CIFAR-10 and CIFAR-100 datasets (L_∞ threat model), as shown in Table 1 and Table 2.

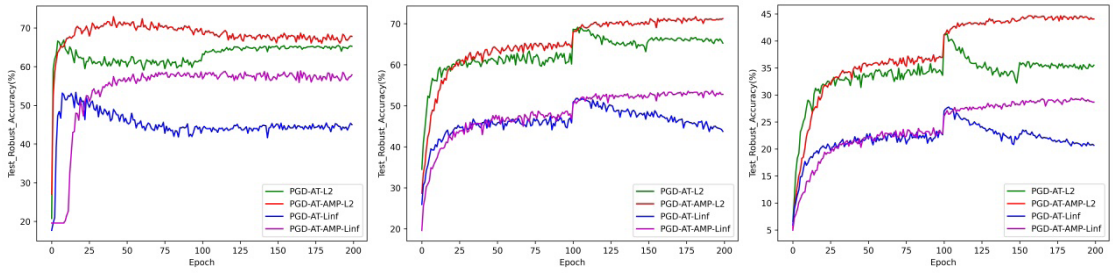
This is consistent with the results of Preetum Nakkiran [35], and the adversarial robustness accuracy may be inherently at odds with natural classification accuracy. We perspective is that the pictures in the CIFAR dataset are more complex than color digits in the SVHN dataset, so they are more challenging. However, experiments show that PGD-AT-AMP is a general method to consistently improve the best and final robustness across diverse datasets and different threat models by finding the flatter minimum of the weight loss landscape.



(a) PGD-AT (b) PGD-AT-AMP

Fig. 5. Comparison of best test robust accuracy and test robust generalization gap (the difference between best train robust accuracy and best test robust accuracy) of PGD-AT and PGD-AT-AMP methods on different three datasets and different two threat models (L_∞ norm $\varepsilon = 8/255$ and L_2 norm $\varepsilon = 128/255$)

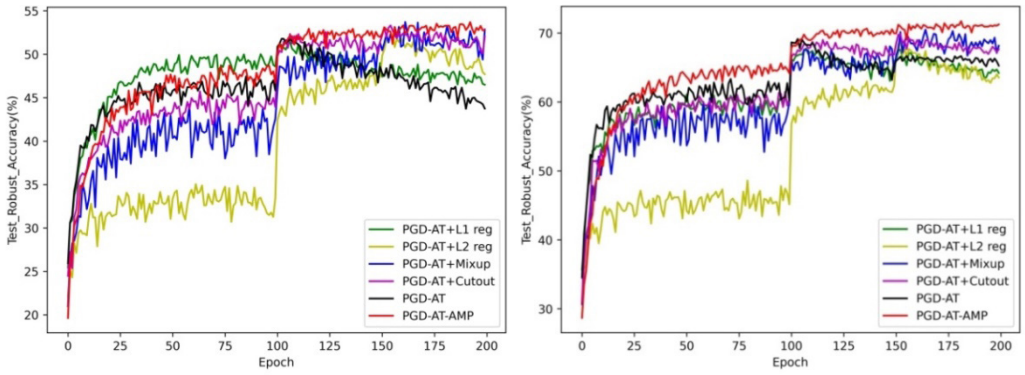
(Yellow and Red curves represent the accuracy of test robust and robustness generalization gap respectively.)



(a) SVHN (b) CIFAR-10 (c) CIFAR-100

Fig. 6. The curve of the test robust accuracy of PGD-AT and PGD-AT-AMP on three different benchmark datasets under the L_∞ norm constraint ($\varepsilon = 8/255$) and the L_2 norm constraint ($\varepsilon = 128/255$) applying PreAct ResNet-18 model for 200 epochs

(Red, Green, purple and Blue curves represent the test robust accuracy of PGD-AT-AMP method under the L_2 norm constraint, PGD-AT method under the L_2 norm constraint, PGD-AT-AMP method under the L_∞ norm constraint and PGD-AT method under the L_∞ norm constraint respectively.)



(a) CIFAR-10 L_∞ norm (b) CIFAR-10 L_2 norm

Fig. 7. The curve of the test robust accuracy of PGD-AT + L1 regularization, PGD-AT + L2 regularization, PGD-AT + Mixup, PGD-AT + Cutout, PGD-AT and PGD-AT-AMP on CIFAR-10 dataset under the L_∞ norm constraint ($\varepsilon = 8/255$) and the L_2 norm constraint ($\varepsilon = 128/255$) using PreAct ResNet-18 for 200 epochs

(Green, yellow, Blue, purple, black and Red curves represent the test robust accuracy of PGD-AT + L1 regularization method, PGD-AT + L2 regularization method, PGD-AT + Mixup method, PGD-AT + Cutout method, PGD-AT and PGD-AT-AMP method respectively.)

4.6 Comparisons to Regularization Techniques and Data Augmentation

Due to the existence of robust generalization gap and robust overfitting problem, we consider using classical regularization technique and classical dataset augmentation technique to conduct adversarial training under different threat models (i.e. L_∞ and L_2 norm) of CIFAR-10 dataset, and further compare them with PGD-AT-AMP method. Detailed experimental research results are shown in Table 4 and Table 5. These experimental results manifest the generalizability of our approach across different datasets, different threat models, different network structures and diverse attack.

In this part, we compare the AMP with L_1 and L_2 regularization techniques and mixup [51] and cutout [52] data augmentation techniques. For the two threat models L_2 and L_∞ norm, we refer to the best hyperparameter settings found in Leslie Rice et al. [13]. The hyperparamete λ for L_1 and L_2 regularization is, 5×10^{-6} / 5×10^{-3} respectively, the hyperparameter $\alpha = 1.4$ for mixup, and the hyperparameter patch length for cutout is 14. For AMP method, we set best hyperparameter $\delta = 0.1$, and we follow Section 4.1 for other training settings. We display the test robustness accuracy and natural accuracy in Table 4 (L_∞ norm threat model) and Table 5 (L_2 norm threat model). Our experimental results are consistent with the previous section in that AMP approach does improve the test robustness in the best case and last case scenarios by some margin. In addition, we visualized the learning curve in Fig. 7. The learning curve shows that PGD-AT-AMP shows better performance than other regularization techniques and data augmentation techniques. However, the Table 4 and Table 5 experimental results show PGD-AT-AMP method compared to other methods (except for the PGD-AT+Mixup method) impairs its natural accuracy on CIFAR-10 dataset under L_∞ threat model, which shows that PGD-AT-AMP is also affected by the trade-off between robustness accuracy and natural accuracy [21]. Under the L_2 threat model on CIFAR-10 dataset, PGD-AT-AMP method has superior natural classification accuracy compared to all other methods and many methods are observed to have similar natural classification accuracy. The above experimental results suggest that the L_∞ threat model may be more difficult and more challenging compared to the L_2 threat model.

Table 4. Accuracy (%) of PGD-AT and PGD-AT with other techniques on CIFAR-10 dataset applying PreAct ResNet-18 model under L_∞ threat model ($\varepsilon = 8/255$) over 5 random runs

(The best signifies the highest robustness accuracy in the whole epoch while last signifies the robustness accuracy at the end of the 200 epochs. The best results are marked in bold.)

Method	Robustness Accuracy		Natural Accuracy	
	Best	Last	Best	Last
PGD-AT + L1 regularization	51.92 ± 0.32	48.74 ± 0.36	**82.72 ± 0.32**	83.42 ± 0.26
PGD-AT + L2 regularization	51.71 ± 0.21	47.38 ± 0.42	81.09 ± 0.41	81.95 ± 0.45
PGD-AT + Mixup	52.83 ± 0.24	49.76 ± 0.81	78.76 ± 0.61	78.50 ± 1.21
PGD-AT + Cutout	52.79 ± 0.14	50.35 ± 0.38	80.99 ± 0.26	**83.65 ± 0.24**
PGD-AT	51.77 ± 0.19	44.36 ± 0.34	81.68 ± 0.19	81.57 ± 0.21
PGD-AT-AMP	**53.85 ± 0.32**	**53.26 ± 0.18**	80.36 ± 0.14	80.44 ± 0.11

Table 5. Accuracy (%) of PGD-AT and PGD-AT with other techniques on CIFAR-10 dataset applying PreAct ResNet-18 model under L_2 threat model ($\varepsilon = 128/255$) over 5 random runs

(The best signifies the highest robustness accuracy in the whole epoch while last signifies the robustness accuracy at the end of the 200 epochs. The best results are marked in bold.)

Method	Robustness Accuracy		Natural Accuracy	
	Best	Last	Best	Last
PGD-AT + L1 regularization	67.96 ± 0.26	63.73 ± 0.42	88.24 ± 0.13	88.34 ± 0.25
PGD-AT + L2 regularization	67.87 ± 0.35	63.69 ± 0.38	88.75 ± 0.16	87.57 ± 0.21
PGD-AT + Mixup	70.19 ± 0.39	68.27 ± 0.26	87.92 ± 0.23	86.96 ± 0.34
PGD-AT + Cutout	69.56 ± 0.24	67.59 ± 0.32	88.45 ± 0.29	88.03 ± 0.15
PGD-AT	69.15 ± 0.13	65.93 ± 0.35	89.57 ± 0.09	88.96 ± 0.18
PGD-AT-AMP	**71.25 ± 0.09**	**71.13 ± 0.08**	**90.03 ± 0.26**	**89.13 ± 0.21**

5 Conclusions

This paper is a further step towards solving the troubles of adversarial training framework robust overfitting and robust generalization gap found in recent studies. Inspired by the idea of smoothness, we seek a solution to explicitly flatten the weight loss landscape in the adversarial training framework, establishing a double perturbation mechanism that infuses input and weight parameter perturbations that directly bound the flatness of the weight loss landscape to facilitate the DNN model to find a more flat minimum, called AT-AMP method. Although the idea is plain, improving robust overfitting and robust generalization gap in adversarial training framework is surprisingly effective. A large number of experiments have shown that our idea achieves advanced performance in adversarial training framework. A substantial body of experimental results has also proved the fact that a flatter weight loss landscape to facilitate the DNNs model to find a more flat minimum often leads to smaller robust generalization gap and robust overfitting in adversarial training framework utilizing on-the-fly produced adversarial examples.

Unfortunately, although promising progress have been made, the underlying causes of the trouble with robust overfitting and robust generalization gap are still left outstanding. Despite the fact that the adversarial robustness accuracy of the DNNs model has been improved, the natural accuracy of the normal examples will be damaged in some scenarios, and the introduction of the double perturbation mechanism proposed in this paper will increase some of the computational overhead in the adversarial training framework. Our future research work will be conducted in this way because we believe that these are very intriguing and momentous questions.

Acknowledgement

This work has been supported by the Special Funds of Central Government of China for Guiding Local Science and Technology Development under Grant No. [2018] 4008, the Science and Technology Planned Project of Guizhou Province, China under Grant No. [2020] 2Y013, the Postgraduate Research Fund of Guizhou Province (Qianjiaohe YJSKYJJ [2021] 102).

References

[1] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
[2] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019.
[3] A. Graves, N. Jaitly, Towards end-to-end speech recognition with recurrent neural networks, in: Proc. 2014 International Conference on Machine Learning (ICML), 2014.
[4] C. Chen, A. Seff, A. Kornhauser, J. Xiao, Deepdriving: Learning affordance for direct perception in autonomous driving, in: Proc. 2015 IEEE International Conference on Computer Vision (ICCV), 2015.
[5] Y. Liao, A. Vakanski, M. Xian, A deep learning framework for assessing physical rehabilitation exercises, IEEE Transactions on Neural Systems and Rehabilitation Engineering 28(2)(2020) 468-477.
[6] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, Intriguing properties of neural networks, in: Proc. 2014 International Conference on Learning Representations (ICLR), 2014.
[7] I. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, in: Proc. 2015 International Conference on Learning Representations (ICLR), 2015.
[8] N. Papernot, P. McDaniel, X. Wu, S. Jha, A. Swami, Distillation as a defense to adversarial perturbations against deep neural networks, in: Proc. 2016 IEEE Symposium on Security and Privacy (SP), 2016.
[9] D. Wu, S. Xia, Y. Wang, Adversarial weight perturbation helps robust generalization, in: Proc. 2020 Advances in Neural Information Processing Systems (NIPS), 2020.
[10] Y. Wang, X. Ma, J. Bailey, J. Yi, B. Zhou, Q. Gu, On the convergence and robustness of adversarial training, in: Proc. 2019 International Conference on Machine Learning (ICML), 2019.
[11] Y. Wang, D. Zou, J. Yi, J. Bailey, X. Ma, Q. Gu, Improving adversarial robustness requires revisiting misclassified examples, in: Proc. 2020 International Conference on Learning Representations (ICLR), 2020.
[12] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards deep learning models resistant to adversarial attacks, in: Proc. 2018 International Conference on Learning Representations (ICLR), 2018.
[13] L. Rice, E. Wong, Z. Kolter, Overfitting in adversarially robust deep learning, in: Proc. 2020 International Conference on Machine Learning (ICML), 2020.
[14] L. Schmidt, S. Santurkar, D. Tsipras, K. Talwar, A. Madry, Adversarially robust generalization requires more data, in:

Proc. 2018 Advances in Neural Information Processing Systems (NIPS), 2018.
[15] A. Krizhevsky, G. Hinton, Learning multiple layers of features from tiny images, Technical Report, University of Toronto, 2009.
[16] K. He, X. Zhang, S. Ren, J. Sun, Identity mappings in deep residual networks, in: Proc. 2016 European Conference on Computer Vision (ECCV), 2016.
[17] H. Li, Z. Xu, G. Taylor, C. Studer, T. Goldstein, Visualizing the loss landscape of neural nets, in: Proc. 2018 Advances in Neural Information Processing Systems (NIPS), 2018.
[18] B. Neyshabur, S. Bhojanapalli, D. McAllester, N. Srebro, Exploring generalization in deep learning, in: Proc. 2017 Advances in Neural Information Processing Systems (NIPS), 2017.
[19] V.U. Prabhu, D.A. Yap, J. Xu, J. Whaley, Understanding adversarial robustness through loss landscape geometries. <https://arxiv.org/abs/1907.09061>, 2019 (accessed 03.12.20).
[20] F. Yu, C. Liu, Y. Wang, L. Zhao, X. Chen, Interpreting adversarial robustness: A view from decision surface in input space. <https://arxiv.org/abs/1810.00144>, 2018 (accessed 26.09.20).
[21] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. Ghaoui, M. Jordan, Theoretically principled trade-off between robustness and accuracy, in: Proc. 2019 International Conference on Machine Learning (ICML), 2019.
[22] Y. Carmon, A. Raghunathan, L. Schmidt, J.C. Duchi, P.S. Liang, Unlabeled data improves adversarial robustness, in: Proc. 2019 Advances in Neural Information Processing Systems (NIPS), 2019.
[23] Y. Zheng, R. Zhang, Y. Mao, Regularizing neural networks via adversarial model perturbation, in: Proc. 2021 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021.
[24] Z. He, A.S. Rakin, D. Fan, Parametric noise injection: Trainable randomness to improve deep neural network robustness against adversarial attack, in: Proc. 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
[25] C. Xie, Y. Wu, L. Maaten, A. Yuille, K. He, Feature denoising for improving adversarial robustness, in: Proc. 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
[26] C. Guo, M. Rana, M. Cissé, L. Maaten, Countering adversarial images using input transformations, in: Proc. 2018 International Conference on Learning Representations (ICLR), 2018.
[27] X. Ma, B. Li, Y. Wang, S. Erfani, S. Wijewickrema, G. Schoenebeck, D. Song, M. Houle, J. Bailey, Characterizing adversarial subspaces using local intrinsic dimensionality, in: Proc. 2018 International Conference on Learning Representations (ICLR), 2018.
[28] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, P. McDaniel, Ensemble adversarial training: Attacks and defenses, in: Proc. 2018 International Conference on Learning Representations (ICLR), 2018.
[29] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. Celik, A. Swami, Practical black-box attacks against machine learning, in: Proc. 2017 ACM on Asia Conference on Computer and Communications Security, 2017.
[30] Q. Liu, T. Liu, Z. Liu, Y. Wang, Y. Jin, W. Wen, Security analysis and enhancement of model compressed deep learning systems under adversarial attacks, in: Proc. 2018 Asia and South Pacific Design Automation Conference, 2018.
[31] G. Dhillon, K. Azizzadenesheli, Z. Lipton, J. Bernstein, J. Kossaifi, A. Khanna, A. Anandkumar, Stochastic activation pruning for robust adversarial defense, in: Proc. 2018 International Conference on Learning Representations (ICLR), 2018.
[32] A. Athalye, N. Carlini, D. Wagner, Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples, in: Proc. 2018 International Conference on Machine Learning (ICML), 2018.
[33] J. Uesato, J. Alayrac, P. Huang, R. Stanforth, A. Fawzi, P. Kohli, Are labels required for improving adversarial robustness? in: Proc. 2019 Advances in Neural Information Processing Systems (NIPS), 2019.
[34] D. Yin, K. Ramchandran, P. Bartlett, Rademacher complexity for adversarially robust generalization, in: Proc. 2019 International Conference on Machine Learning (ICML), 2019.
[35] P. Nakkiran, Adversarial robustness may be at odds with simplicity. <https://arxiv.org/abs/1901.00532>, 2019 (accessed 20.03.21).
[36] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, A. Madry, Robustness may be at odds with accuracy, in: Proc. 2019 International Conference on Learning Representations (ICLR), 2019.
[37] P. Chaudhari, A. Choromanska, S. Soatto, Y. LeCun, C. Baldassi, C. Borgs, J. Chayes, L. Sagun, R. Zecchina, Entropy-sgd: Biasing gradient descent into wide valleys, in: Proc. 2017 International Conference on Learning Representations (ICLR), 2017.
[38] N. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, P. Tang, On large-batch training for deep learning: generalization gap and sharp minima, in: Proc. 2017 International Conference on Learning Representations (ICLR), 2017.
[39] C. Xie, M. Tan, B. Gong, A. Yuille, Q. Le, Smooth adversarial training. <https://arxiv.org/abs/2006.14536> 2020, (accessed 03.07.21).
[40] T. Ishida, I. Yamane, T. Sakai, G. Niu, M. Sugiyama, Do we need zero training loss after achieving zero training error? in: Proc. 2020 International Conference on Machine Learning (ICML), 2020.
[41] P. Foret, A. Kleiner, H. Mobahi, B. Neyshabur, Sharpness-aware minimization for efficiently improving generalization, in: Proc. 2021 International Conference on Learning Representations (ICLR), 2021.
[42] V.N. Vapnik, Statistical learning theory, Wiley, 1998.
[43] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, A. Ng, Reading digits in natural images with unsupervised feature learning, in: Proc. 2011 Advances in Neural Information Processing Systems (NIPS) Workshop on Deep Learning and

Unsupervised Feature Learning, 2011.
[44] S. Zagoruyko, N. Komodakis, Wide residual networks. <https://arxiv.org/abs/1605.07146>, 2016 (accessed 06.09.20).
[45] D. Hendrycks, K. Lee, M. Mazeika, Using pre-training can improve model robustness and uncertainty, in: Proc. 2019 International Conference on Machine Learning (ICML), 2019.
[46] N. Carlini, D. Wagner, Towards evaluating the robustness of neural networks, in: Proc. 2017 IEEE Symposium on Security and Privacy (SP), 2017.
[47] J. Uesato, B. O'Donoghue, P. Kohli, A. Oord, Adversarial risk and the dangers of evaluating against weak attacks, in: Proc. 2018 International Conference on Machine Learning (ICML), 2018.
[48] F. Croce, M. Hein, Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks, in: Proc. 2020 International Conference on Machine Learning (ICML), 2020.
[49] F. Croce, M. Hein, Minimally distorted adversarial examples with a fast adaptive boundary attack, <https://arxiv.org/abs/1907.02044>, 2019 (accessed 03.04.21).
[50] M. Andriushchenko, F. Croce, N. Flammarion, M. Hein, Square attack: a query-efficient black-box adversarial attack via random search, in: Proc. 2020 European Conference on Computer Vision (ECCV), 2020.
[51] H. Zhang, M. Cisse, Y. Dauphin, D. Lopez-Paz, Mixup: Beyond empirical risk minimization, in: Proc. 2018 International Conference on Learning Representations (ICLR), 2018.
[52] T. DeVries, G. Taylor, Improved regularization of convolutional neural networks with cutout. < https://arxiv.org/abs/1708.04552>, 2017 (accessed 03.11.20).