Prior-Guided Adversarial Initialization for Fast Adversarial Training

Xiaojun Jia[1,2,†], Yong Zhang[3,⋆] Xingxing Wei[4], Baoyuan Wu[5], Ke Ma[6], Jue Wang[3], Xiaochun Cao[1,7,⋆]

[1]SKLOIS, Institute of Information Engineering, CAS, Beijing, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
[3] Tencent, AI Lab, Shenzhen, China
[4] Institute of Artificial Intelligence, Beihang University, Beijing, China.
[5]School of Data Science, Secure Computing Lab of Big Data, Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong, Shenzhen, China
[6] School of Computer Science and Technology, UCAS, Beijing, China
[7] School of Cyber Science and Technology, Shenzhen Campus, Sun Yat-sen University, Shenzhen 518107, China
`jiaxiaojun@iie.ac.cn; zhangyong201303@gmail.com; xxwei@buaa.edu.cn;`
`wubaoyuan@cuhk.edu.cn; make@ucas.ac.cn; arphid@gmail.com;`
`caoxiaochun@mail.sysu.edu.cn`

Abstract. Fast adversarial training (FAT) effectively improves the efficiency of standard adversarial training (SAT). However, initial FAT encounters catastrophic overfitting, *i.e.,* the robust accuracy against adversarial attacks suddenly and dramatically decreases. Though several FAT variants spare no effort to prevent overfitting, they sacrifice much calculation cost. In this paper, we explore the difference between the training processes of SAT and FAT and observe that the attack success rate of adversarial examples (AEs) of FAT gets worse gradually in the late training stage, resulting in overfitting. The AEs are generated by the fast gradient sign method (FGSM) with a zero or random initialization. Based on the observation, we propose a prior-guided FGSM initialization method to avoid overfitting after investigating several initialization strategies, improving the quality of the AEs during the whole training process. The initialization is formed by leveraging historically generated AEs without additional calculation cost. We further provide a theoretical analysis for the proposed initialization method. We also propose a simple yet effective regularizer based on the prior-guided initialization, *i.e.,* the currently generated perturbation should not deviate too much from the prior-guided initialization. The regularizer adopts both historical and current adversarial perturbations to guide the model learning. Evaluations on four datasets demonstrate that the proposed method can prevent catastrophic overfitting and outperform state-of-the-art FAT methods. The code is released at <https://github.com/jiaxiaojunQAQ/FGSM-PGI>.

**Keywords:** Fast Adversarial Training, Prior-Guided, Regularizer

---
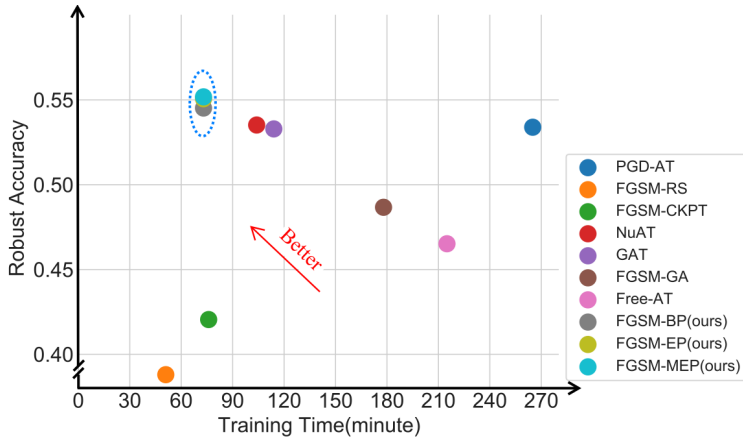
Fig. 1. PGD-10 robust accuracy and training time of different FAT methods with ResNet18 as the backbone on the CIFAR-10 dateset. x-axis denotes training time (lower is more efficient) and y-axis denotes PGD-10 accuracy (higher is more robust).

1   Introduction

Deep neural networks (DNNs) [27,5,55,16,48,44,20,29,30,28] are vulnerable to adversarial examples (AEs) [43,8,14,12,31,51,21,17,13,3,4] generated by adding imperceptible perturbations to benign images. Standard adversarial training (SAT) [32,46,22] is one of the most efficient defense methods against AEs. It adopts projected gradient descent (PGD) [32], a multi-step attack method, to generate AEs for training. But, SAT requires much time to calculate gradients of the network's input multiple times.

To improve efficiency, fast adversarial training (FAT) methods [15,35,53,23] have been proposed. Goodfellow *et al.* first [15] adopt FGSM to generate AEs for training, *i.e.,* FGSM-AT, but it encounters the catastrophic overfitting. Wong *et al.* [49] propose to combine FGSM-AT with random initialization along with early stopping to overcome the overfitting. Andriushchenko *et al.* [2] propose a method from the view of regularization to enhance the quality of AEs, *i.e.,* GradAlign. Kim *et al.* [25] propose a simple method to determine an appropriate step size of FGSM-AT to generate stronger AEs, *i.e.,* FGSM-CKPT. Sriramanan *et al.* [41] introduce a relaxation term in the training loss to improve the quality of the generated AEs, *i.e.,* GAT. Then, Sriramanan *et al.* [42] propose a Nuclear-Norm regularizer to enhance the optimization of the AE generation and the model training. These methods not only prevent catastrophic overfitting but also achieve advanced defense performance by enhancing the quality of AEs. However, compared with FGSM-RS, these advanced methods require high additional calculation costs to generate stronger AEs.

As AEs are critical for adversarial training, we study the AEs of both FAT and SAT during their training processes to explore the reason for catastrophic
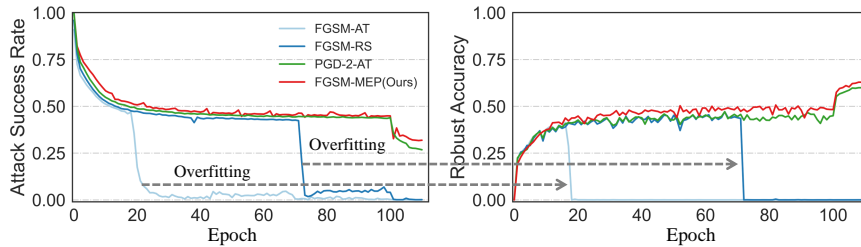
Fig. 2. The difference between the training processes of FAT and SAT methods on the CIFAR-10 dataset. Left: the attack success rate of generated AEs. Right: the PGD-10 robust accuracy of the target model. FGSM-AT and FGSM-RS encounter catastrophic overfitting after training a few epochs. But PGD-2-AT and our method can prevent the overfitting.

overfitting. Surprisingly, we observe a distinct difference between the attack success rates of their AEs in the late training stage. The attack success rate of FAT drops sharply after a few training epochs while the robust accuracy decreases to 0 (see FGSM-AT and FGSM-RS in Fig. 2). This phenomenon indicates that overfitting happens when the quality of AEs becomes worse. However, the overfitting does not exist in the two-step PGD (PGD-2-AT) that can be treated as FGSM-AT with an adversarial initialization. It means adversarial initialization could be a solution to overfitting. Based on the observation, we raise a question *"can we obtain an adversarial initialization for FGSM-AT to maintain the high quality of AEs for avoiding catastrophic overfitting without additional calculation cost?"* We investigate several initialization strategies and propose a prior-guided initialization by leveraging historically generated AEs, dubbed FGSM-PGI. Specifically, we first use the buffered perturbations of AEs from the previous batch and the previous epoch as the initialization for FGSM, respectively, called FGSM-BP and FGSM-EP. The two strategies are demonstrated to be effective. To exploit complete prior information, we then propose to leverage the buffered gradients from all previous epochs via a momentum mechanism as an additional prior, dubbed FGSM-MEP, which works the best. Furthermore, we provide a theoretical analysis for the proposed initialization method.

Moreover, we also propose a simple yet effective regularizer based on the above prior-guided initialization to guide model learning for better robustness. The current perturbation can be generated via FGSM using the prior-guided initialization. The regularizer prevents the current perturbation deviating too much from the prior-guided initialization, which is implemented by using a squared L_2 distance between the predictions of adversarial examples generated based on the current perturbation and the prior-guided initialization.

The proposed method not only prevents catastrophic overfitting but also improves the adversarial robustness against adversarial attacks. Fig. 1 illustrates the robustness and training efficiency of different FAT methods. Our main contributions are in three aspects: **1)** We propose a prior-guided adversarial initializa-

tion to prevent overfitting after investigating several initialization strategies. **2)** We also propose a regularizer to guide the model learning for better robustness by considering both the currently generated perturbation and the prior-guided initialization. **3)** Extensive experiments on four datasets demonstrate that the proposed method can outperform state-of-the-art FAT methods in terms of both efficiency and robustness.

2   Related Work

2.1   Adversarial Training

Adversarial training (AT) variants [54,37,50] are effective in defending against AEs. Madry *et al.* [32] first formulate AT as a minimax optimization problem formulated as follows:

$$\min_{\mathbf{w}} \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[\max_{\boldsymbol{\delta}\in\Omega} \mathcal{L}(f(\mathbf{x} + \boldsymbol{\delta}; \mathbf{w}), \mathbf{y})], \tag{1}$$

where $f(\cdot; \mathbf{w})$ represents a deep neural network with parameters $\mathbf{w}$, $\mathcal{D}$ represents a joint data distribution of the benign image $\mathbf{x}$ and the GT one-hot label $\mathbf{y}$. $\mathcal{L}(f(\mathbf{x}; \mathbf{w}), \mathbf{y})$ represents the classification loss function. $\boldsymbol{\delta}$ represents the adversarial perturbation, and Ω represents a threat bound which can be defined as $\Omega = \{\boldsymbol{\delta} : \|\boldsymbol{\delta}\| \leq \epsilon\}$ with the maximum perturbation strength ϵ. Multi-step AT methods generate AEs by a multi-step adversarial attack method, *i.e.*, projected gradient ascent (PGD) [32]. It can be defined as:

$$\boldsymbol{\delta}_{t+1} = \Pi_{[-\epsilon,\epsilon]} \left[\boldsymbol{\delta}_t + \alpha \cdot \text{sign} \left(\nabla_{\mathbf{x}} \mathcal{L} \left(f \left(\mathbf{x} + \boldsymbol{\delta}_t; \mathbf{w} \right), \mathbf{y} \right) \right) \right], \tag{2}$$

where $\boldsymbol{\delta}_{t+1}$ represents the perturbation at the $t+1$-th iteration, α represents the attack step size, and $\Pi_{[-\epsilon,\epsilon]}$ is the projection that maps the input to the range $[-\epsilon, \epsilon]$. This is a prime PGD-based AT framework proposed in [32]. Following this work, many advanced variants [45,47,6,24,34,38] have been proposed from different perspectives to improve model robustness. An early stopping version [37] of PGD-based AT stands out amongst them.

2.2   Fast Adversarial Training

Although multi-step AT methods can achieve good robustness, they require lots of calculation costs to generate AEs for training. Fast adversarial training variants that generate AEs by the one-step fast gradient sign method (FGSM) [15] are proposed to improve the efficiency, which can be dubbed FGSM-based AT methods. The perturbation of FGSM-AT [15] is defined as:

$$\boldsymbol{\delta} = \epsilon \cdot \text{sign} \left(\nabla_{\mathbf{x}} \mathcal{L}(f(\mathbf{x}; \mathbf{w}), \mathbf{y}) \right), \tag{3}$$

where ϵ is the maximal perturbation strength. Though FGSM-based AT accelerates the training speed of AT, it encounters catastrophic overfitting. Specifically,

after a period of training, the trained model suddenly cannot defend against the AEs generated by PGD. Wong *et al.* [49] propose to conduct FGSM-based AT with a random initialization. Its perturbation can be defined as:

$$\boldsymbol{\delta} = \Pi_{[-\epsilon,\epsilon]}\left[\boldsymbol{\eta} + \alpha \cdot \text{sign}\left(\nabla_{\mathbf{x}}\mathcal{L}(f(\mathbf{x} + \boldsymbol{\eta}; \mathbf{w}), \mathbf{y})\right)\right], \tag{4}$$

where $\boldsymbol{\eta} \in \mathbf{U}(-\epsilon, \epsilon)$ represents the random initialization, $\mathbf{U}$ is a uniform distribution, and α is set to 1.25ϵ. After that, a series of FAT methods enhances the quality of AEs to improve the defense performance. In detail, Andriushchenko *et al.* [2] demonstrate that using the random initialization cannot prevent catastrophic overfitting. They propose a regularizer to generate stronger AEs, *i.e.,* FGSM-GA. Kim *et al.* [25] reveal that FGSM-RS adopts AEs with the maximum perturbation instead of ones in the adversarial direction, resulting in the overfitting. They then propose a method to determine an appropriate step size to generate powerful AEs to improve model robustness. Sriramanan *et al.* [41] enhance the quality of the generated AEs by using a relaxation term for the classification loss, *i.e.,* GAT. Besides, Sriramanan *et al.* [42] design a Nuclear-Norm regularizer to enhance optimization of the AEs generation to improve the model robustness, *i.e.,* NuAT. Though these methods can prevent overfitting and achieve advanced defense performance, they require much additional calculation cost to conduct AT.

3   The Proposed Approach

We first present our observations of revisiting catastrophic overfitting in Sec. 3.1 and then explore several prior-guided adversarial initialization strategies for FGSM-AT to prevent the overfitting in Sec. 3.2, *i.e.,* FGSM-PGI. We propose a simple yet effective regularization method for the FGSM-PGI to further improve model robustness in Sec. 3.3. Moreover, we also provide a theoretical analysis for the proposed initialization method in Sec. 3.4.

3.1   Revisting Catastrophic Overfitting

Catastrophic overfitting is a phenomenon that the robust accuracy of FAT drops sharply to 0% in the late training stage (see the left figure of Fig. 2), which was noticed by Wong *et al.* [49]. They then found random initialization of FGSM can help avoid the overfitting with training for limited epochs. However, if the training process of FGSM-RS goes on, catastrophic overfitting still appears, as revealed in [2], which indicates that random initialization must be conducted with early stopping. Hence, FGSM-RS does not solve the overfitting fundamentally. FGSM-GA [2] proposes a method from the perspective of regularization in the loss to generate stronger AEs. FGSM-CKPT [25] proposes a method for adjusting the step size of FGSM to improve the quality of AEs in different training stages. GAT [41] adopts a relaxation term for the classification loss to generate stronger AEs during the AEs generation. NuAT [42] uses a Nuclear Norm

regularizer to improve the optimization of the AEs generation. These methods prevent catastrophic overfitting by improving the quality of AEs during the training. However, FGSM-GA requires calculating gradients multiple times for regularization, FGSM-CKPT requires multiple times of forwarding propagation for step size selection, and GAT and NuAT require much more time to conduct regularization for generating powerful AEs, which significantly reduces efficiency.

We reinvestigate catastrophic overfitting by comparing the intermediate results of the training processes of FAT and SAT on the CIFAR-10 dataset [26]. For FAT, we study two methods, *i.e.,* FGSM-AT and FGSM-RS. For SAT, we study PGD-AT [32] with the iteration time of 2, *i.e.,* PGD-2-AT. As AEs are the key that distinguishes adversarial training from conventional training, we observe the quality of adversarial examples during the whole training process, *i.e.,* whether adversarial examples can successfully attack the target model. The quality is evaluated by attack success rate (ASR). Fig. 2 illustrates the curves of the robust accuracy and the ASR of the above methods as well as ours. We also observe similar phenomenons on other benchmark datasets that are shown in the **supplementary material**.

We summarize the observations as follows. First, it can be observed that the ASRs of FGSM-AT and FGSM-RS drop sharply at the 20-th and 74-th epoch, respectively, leading to the dramatic decreases of the robust accuracy. This indicates that the model robustness would collapse against adversarial attacks if the generated adversarial examples fail in attacking against the model during training. Note that the investigation from the perspective of the quality of adversarial example (*i.e.,* ASR) is ignored by previous works. Second, FGSM-RS with random initialization delays the overfitting from the 20-th epoch to the 70-th epoch. Enhancing the initialization could alleviate the overfitting, but cannot essentially prevent it, which is also observed in [2]. Third, surprisingly, PGD-2-AT does not suffer from the overfitting, which draws our attention. PGD-2-AT can be treated as FGSM-AT with an adversarial initialization of FGSM for adversarial example generation. Even after 110 training epochs, there are still a portion of high-quality adversarial examples that can fool the model. It indicates that adversarial initialization can improve the quality of adversarial examples. The initialization requires additional gradient calculation, which is not desirable.

Table 1. Comparisons of clean and robust accuracy (%) and training time (minute) on the CIFAR-10 dataset. Number in bold indicates the best.

Method		Clean	PGD-10	PGD-20	PGD-50	C&W	AA	Time(min)
FGSM-BP	Best	**83.15**	54.59	53.55	53.2	50.24	47.47	73
	Last	**83.09**	54.52	53.5	53.33	50.12	47.17	
FGSM-EP	Best	82.75	54.8	53.62	53.27	49.86	47.94	73
	Last	81.27	55.07	54.04	53.63	50.12	46.83	
FGSM-MEP	Best	81.72	**55.18**	**54.36**	**54.17**	**50.75**	**49.00**	73
	Last	81.72	**55.18**	**54.36**	**54.17**	**50.75**	**49.00**	

3.2   Prior-Guided Adversarial Initialization

The above three observations motivate us to dive into the question *"how to obtain adversarial initialization without additional calculation cost"*. We come up with using the adversarial perturbations of historical AEs as the initialization for FGSM to conduct adversarial training, dubbed FGSM-PGI. Such perturbations serve as the prior knowledge that we can freely obtain without additional gradient calculation except for extra storage. We explore three strategies of exploiting historical perturbations, *i.e.,* taking perturbations from the previous batch, from the previous epoch, and from the momentum of all epochs, *i.e.,* FGSM-BP, FGSM-EP, and FGSM-MEP.

Prior From the Previous Batch (FGSM-BP). We store the perturbations of AEs from the previous batch and regard them as the initialization of FGSM to generate AEs in the current batch. As the batch is randomly sampled, there is no correspondence between perturbations from the previous batch and samples from the current batch. Specifically, for a data point $\mathbf{x}$, we first add the perturbation from the previous batch onto it and then conduct FGSM based on the perturbed example. The adversarial perturbation can be defined as:

$$\boldsymbol{\delta}_{B_{t+1}} = \Pi_{[-\epsilon,\epsilon]} \left[\boldsymbol{\delta}_{B_t} + \alpha \cdot \text{sign} \left(\nabla_{\mathbf{x}} \mathcal{L}(f(\mathbf{x} + \boldsymbol{\delta}_{B_t}; \mathbf{w}), \mathbf{y}) \right) \right], \tag{5}$$

where $\boldsymbol{\delta}_{B_{t+1}}$ is the adversarial perturbation in the $t+1$-th batch. Compared with FGSM-RS (see Eq. 4), we replace the random initialization with the adversarial perturbation from the previous batch.

Prior From the Previous Epoch (FGSM-EP). We store the perturbations of all adversarial samples from the previous epoch and use them as initialization of FGSM to generate adversarial perturbations for the samples in the current epoch. Note that there is a correspondence between each perturbation from the previous epoch and the sample in the current epoch. The adversarial perturbation is defined as:

$$\boldsymbol{\delta}_{E_{t+1}} = \Pi_{[-\epsilon,\epsilon]} \left[\boldsymbol{\delta}_{E_t} + \alpha \cdot \text{sign} \left(\nabla_{\mathbf{x}} \mathcal{L}(f(\mathbf{x} + \boldsymbol{\delta}_{E_t}; \mathbf{w}), \mathbf{y}) \right) \right], \tag{6}$$

where $\boldsymbol{\delta}_{E_{t+1}}$ is the adversarial perturbation in the $t+1$-th epoch. Compared with FGSM-RS (see Eq. 4), we replace the random initialization with the adversarial perturbation from the previous epoch.

Prior From the Momentum of All Previous Epochs (FGSM-MEP). To completely leverage historical adversarial perturbations during the whole training process, we propose to compute the momentum of one sample's gradients across all previous training epochs. Then, the gradient momentum is used as the initialization of FGSM for AE generation in the current epoch. There is also a correspondence between the gradient momentum and the sample in the current

epoch. The adversarial perturbation can be defined as:

$$\mathbf{g}_c = \text{sign}\left(\nabla_{\mathbf{x}}\mathcal{L}(f(\mathbf{x} + \boldsymbol{\eta}_{E_t}; \mathbf{w}), \mathbf{y})\right), \tag{7}$$

$$\mathbf{g}_{E_{t+1}} = \mu \cdot \mathbf{g}_{E_t} + \mathbf{g}_c, \tag{8}$$

$$\boldsymbol{\delta}_{E_{t+1}} = \Pi_{[-\epsilon,\epsilon]}\left[\boldsymbol{\eta}_{E_t} + \alpha \cdot \mathbf{g}_c\right], \tag{9}$$

$$\boldsymbol{\eta}_{E_{t+1}} = \Pi_{[-\epsilon,\epsilon]}\left[\boldsymbol{\eta}_{E_t} + \alpha \cdot \text{sign}(\mathbf{g}_{E_{t+1}})\right]. \tag{10}$$

Similar to FGSM-AT (Eq. 3), $\mathbf{g}_c$ is regarded as the signed gradient and $\mathbf{g}_{E_{t+1}}$ is the signed gradient momentum in the $t + 1$-th epoch. μ represents the decay factor. $\boldsymbol{\delta}_{E_{t+1}}$ is the adversarial perturbation, similar to that of FGSM-EP. $\boldsymbol{\eta}_{E_{t+1}}$ is the projected perturbation, which is used as adversarial initialization in the next epoch. Compared with FGSM-EP, FGSM-MEP exploits the gradient momentum involving information of all previous epochs as initialization, instead of using only information in the previous epoch.

3.3   Prior-guided Initialization based Regularization

We propose an effective regularization method based on the prior-guided initialization to improve model robustness. Specifically, given the prior-guided initialization, we can generate a current perturbation via FGSM with the initialization. Both the current perturbation and the initialization can be used to create adversarial examples. Forcing the two types of adversarial examples to have similar predictions could help improve the smoothness of function. The proposed regularization term can be added into the training loss to update the model parameters $\mathbf{w}_{t+1}$, as follows:

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}}[\mathcal{L}(f(\mathbf{x} + \boldsymbol{\delta}_{adv}; \mathbf{w}), \mathbf{y}) + \lambda \cdot \|f(\mathbf{x} + \boldsymbol{\delta}_{adv}; \mathbf{w}) - f(\mathbf{x} + \boldsymbol{\delta}_{pgi}; \mathbf{w})\|_2^2], \tag{11}$$

where $\boldsymbol{\delta}_{pgi}$ represents the prior-guided initialization generated by one of the above three initialization methods. $\boldsymbol{\delta}_{adv}$ represents the current adversarial perturbation generated via FGSM using $\boldsymbol{\delta}_{pgi}$ as initialization. λ is a trade-off hyperparameter. The first term is the cross-entropy loss on AEs generated using the current perturbation. The second term represents the regularization, *i.e.,* the squared L_2 distance between the predictions of the two types of adversarial examples. This term makes the learned model not only robust to currently generated AEs but also historically generated AEs. In this way, the proposed regularization term explicitly enforces the function smoothness around samples to improve model robustness.

Based on the proposed prior-guided adversarial initialization and the regularization, we can establish our FAT framework. We evaluate the three prior-guided adversarial initialization approaches on CIFAR-10 (see results in Table 1) and find that they all can prevent catastrophic overfitting and achieve advanced model robustness against adversarial attacks. FGSM-MEP works the best in terms of robust accuracy under all attack scenarios. **The FGSM-EP, FGSM-BP and FGSM-MEP algorithms are presented in the supplementary material**.

3.4   Theoretical Analysis

Proposition 1. *Let $\boldsymbol{\delta}_{pgi}$ be the prior-guided adversarial initialization in **FGSM-BP**, **FSGM-EP** or **FSGM-MEP**, $\hat{\boldsymbol{\delta}}_{adv}$ represents the current adversarial perturbation generated via FGSM using $\boldsymbol{\delta}_{pgi}$ as initialization, and α be the step size of (5), (6), (9) and (10). If $\boldsymbol{\Omega}$ is a bounded set like*

$$\boldsymbol{\Omega} = \left\{ \hat{\boldsymbol{\delta}}_{adv} \ : \ \|\hat{\boldsymbol{\delta}}_{adv} - \boldsymbol{\delta}_{pgi}\|_2^2 \leq \epsilon^2 \right\}, \tag{12}$$

and the step size α satisfies $\alpha \leq \epsilon$, it holds that

$$
\begin{aligned}
\mathbb{E}_{\hat{\boldsymbol{\delta}}_{adv} \sim \boldsymbol{\Omega}}\left[\|\hat{\boldsymbol{\delta}}_{adv}\|_2\right] &\leq \sqrt{\mathbb{E}_{\hat{\boldsymbol{\delta}}_{adv} \sim \boldsymbol{\Omega}}\left[\|\hat{\boldsymbol{\delta}}_{adv}\|_2^2\right]} \\
&\leq \sqrt{\frac{1}{d} \cdot \epsilon},
\end{aligned} \tag{13}
$$

*where $\hat{\boldsymbol{\delta}}_{adv}$ is the adversarial perturbation generated by **FGSM-BP**, **FSGM-EP** or **FSGM-MEP**, and d is the dimension of the feature space.*

The proof is deferred to the **supplementary material**. The upper bound of the proposed method is $\sqrt{\frac{1}{d}} \cdot \epsilon$ which is less than the bound $\sqrt{\frac{d}{3}} \cdot \epsilon$ of FGSM-RS provided in [2]. Due to the norm of perturbation (gradient) can be treated as the convergence criteria for the non-convex optimization problem, the smaller expectation represents that the proposed prior-guided adversarial initialization will be converged to a local optimal faster than the random initialization with the same number of iterations.

4   Experiments

4.1   Experimental Setting

Dataset Settings. To evaluate the effectiveness of the proposed FGSM-MEP, extensive experiments are conducted on four benchmark datasets that are the most widely used to evaluate adversarial robustness, *i.e.*, CIFAR-10 [26], CIFAR-100 [26], Tiny ImageNet [11], and ImageNet [11]. Following the commonly used settings in the adversarial training, we adopt ResNet18 [18] and WideResNet34-10 [52] as the backbone on CIFAR-10 and CIFAR-100, PreActResNet18 [19] on Tiny ImageNet, and ResNet50 [18] on ImageNet. We adopt the SGD optimizer [36] with a learning rate of 0.1, the weight decay of 5e-4, and the momentum of 0.9. As for CIFAR-10, CIFAR-100, and Tiny ImageNet, following the settings of [37,33], the total epoch number is set to 110. The learning rate decays with a factor of 0.1 at the 100th and 105th epoch. As for ImageNet, following the settings of [39,49], the total epoch number is set to 90. The learning rate decays with a factor of 0.1 at the 30th and 60th epoch. Experiments on ImageNet are conducted with 8 Tesla V100 and other experiments are conducted with a single Tesla V100. For our hyper-parameters, the decay factor μ is set to 0.3 and the

Table 2. Comparisons of clean and robust accuracy (%) and training time (minute) using ResNet18 on the CIFAR-10 dataset. Number in bold indicates the best.

Method		Clean	PGD-10	PGD-20	PGD-50	C&W	AA	Time(min)
PGD-AT [37]	Best	82.32	53.76	52.83	52.6	51.08	48.68	265
	Last	82.65	53.39	52.52	52.27	51.28	48.93	
FGSM-RS [49]	Best	73.81	42.31	41.55	41.26	39.84	37.07	51
	Last	83.82	00.09	00.04	00.02	0.00	0.00	
FGSM-CKPT [25]	Best	**90.29**	41.96	39.84	39.15	41.13	37.15	76
	Last	**90.29**	41.96	39.84	39.15	41.13	37.15	
NuAT [42]	Best	81.58	53.96	52.9	52.61	**51.3**	**49.09**	104
	Last	81.38	53.52	52.65	52.48	50.63	48.70	
GAT [41]	Best	79.79	54.18	53.55	53.42	49.04	47.53	114
	Last	80.41	53.29	52.06	51.76	49.07	46.56	
FGSM-GA [2]	Best	83.96	49.23	47.57	46.89	47.46	43.45	178
	Last	84.43	48.67	46.66	46.08	46.75	42.63	
Free-AT(m=8) [39]	Best	80.38	47.1	45.85	45.62	44.42	42.17	215
	Last	80.75	45.82	44.82	44.48	43.73	41.17	
FGSM-BP (ours)	Best	83.15	54.59	53.55	53.2	50.24	47.47	73
	Last	83.09	54.52	53.5	53.33	50.12	47.17	
FGSM-EP (ours)	Best	82.75	54.8	53.62	53.27	49.86	47.94	73
	Last	81.27	55.07	54.04	53.63	50.12	46.83	
FGSM-MEP (ours)	Best	81.72	**55.18**	**54.36**	**54.17**	50.75	49.00	73
	Last	81.72	**55.18**	**54.36**	**54.17**	**50.75**	**49.00**	

hyper-parameter λ is set to 8. More details are presented in the **supplementary material**.

Evaluation Metrics. For adversarial robustness evaluation, we adopt several widely used attack methods, *i.e.,* PGD [32], C&W [7], and an ensemble of diverse parameter-free attacks, AA [10] that includes APGD [10], APGD-T [10], FAB [9], and Square [1]. We set the maximum perturbation strength ϵ to 8 for all attack methods. Moreover, PGD attack is conducted with 10, 20 and, 50 iterations, *i.e.,* PGD-10, PGD-20, and PGD-50. We report the results of the checkpoint with the best accuracy under the attack of PGD-10 as well as the results of the last checkpoint. All experiments in the manuscript are performed with the multi-step learning rate strategy. We also conduct experiments by using a cyclic learning rate strategy [40] which are presented in the **supplementary material**.

Competing Methods. We compare the proposed method with a series of state-of-the-art FAT methods, *i.e.,* Free-AT[39], FGSM-RS [49], FGSM-GA[2], FGSM-CKPT[25], GAT [41] and NuAT [42]. We also compare with an advanced multi-step AT method, *i.e.,* PGD-AT[37]), which is an early stopping version of the original PGD-based AT method [32].

Table 3. Comparisons of clean and robust accuracy (%) and training time (minute) using ResNet18 on the CIFAR-100 dataset. Number in bold indicates the best.

Method		Clean	PGD-10	PGD-20	PGD-50	C&W	AA	Time(min)
PGD-AT [37]	Best	57.52	29.6	28.99	28.87	28.85	25.48	284
	Last	57.5	29.54	29.00	28.90	27.6	25.48	
FGSM-RS [49]	Best	49.85	22.47	22.01	21.82	20.55	18.29	70
	Last	60.55	00.45	00.25	00.19	00.25	0.00	
FGSM-CKPT [25]	Best	**60.93**	16.58	15.47	15.19	16.4	14.17	96
	Last	**60.93**	16.69	15.61	15.24	16.6	14.34	
NuAT [41]	Best	59.71	27.54	23.02	20.18	22.07	11.32	115
	Last	59.62	27.07	22.72	20.09	21.59	11.55	
GAT [42]	Best	57.01	24.55	23.8	23.55	22.02	19.60	119
	Last	56.07	23.92	23.18	23.0	21.93	19.51	
FGSM-GA [2]	Best	54.35	22.93	22.36	22.2	21.2	18.88	187
	Last	55.1	20.04	19.13	18.84	18.96	16.45	
Free-AT(m=8) [39]	Best	52.49	24.07	23.52	23.36	21.66	19.47	229
	Last	52.63	22.86	22.32	22.16	20.68	18.57	
FGSM-BP (ours)	Best	57.58	30.78	30.01	28.99	26.40	23.63	83
	Last	83.82	30.56	29.96	28.82	26.32	23.43	
FGSM-EP (ours)	Best	57.74	31.01	30.17	29.93	27.37	24.39	83
	Last	57.74	31.01	30.17	29.93	27.37	24.39	
FGSM-MEP (ours)	Best	58.78	**31.88**	**31.26**	**31.14**	**28.06**	**25.67**	83
	Last	58.81	**31.6**	**31.03**	**30.88**	**27.72**	**25.42**	

4.2   Results on CIFAR-10

On CIFAR-10, we adopt ResNet18 as the backbone. The results are shown in Table 2. Note that the results of using WideResNet34-10 as the backbone are shown in the **supplementary material**. The proposed FGSM-MEP prevents catastrophic overfitting and even achieves better robustness than PGD-AT in most cases. It costs much less time than PGD-AT. Specifically, under the AA attack, PGD-AT achieves an accuracy of about 48% while the proposed FGSM-MEP achieves an accuracy of about 49%.

Compared with other FAT methods, the proposed FGSM-MEP achieves comparable robustness to the previous most potent FAT method, NuAT. But FGSM-MEP costs less time than NuAT (73 min VS 104 min). Besides, for the last checkpoint, the proposed FGSM-MEP can achieve the best robustness performance under all attack scenarios. And for the best checkpoint, the proposed FGSM-MEP also achieves the best robustness performance under the PGD-10, PGD-20, and PGD-50. In terms of efficiency, our training process is about 3 times faster than Free-AT, 2.5 times faster than FGSM-GA, and 1.4 times faster than GAT and NuAT. Previous fast adversarial training variants [2,25,41,42]

Table 4. Comparisons of clean and robust accuracy (%) and training time (minute) using PreActResNet18 on Tiny ImageNet. Number in bold indicates the best.

Method		Clean	PGD-10	PGD-20	PGD-50	C&W	AA	Time(min)
PGD-AT [37]	Best	43.6	20.2	19.9	19.86	17.5	16.00	1833
	Last	45.28	16.12	15.6	15.4	14.28	12.84	
FGSM-RS [49]	Best	44.98	17.72	17.46	17.36	15.84	14.08	339
	Last	45.18	0.00	0.00	0.00	0.00	0.00	
FGSM-CKPT [25]	Best	**49.98**	9.20	9.20	8.68	9.24	8.10	464
	Last	**49.98**	9.20	9.20	8.68	9.24	8.10	
NuAT [42]	Best	42.9	15.12	14.6	14.44	12.02	10.28	660
	Last	42.42	13.78	13.34	13.2	11.32	9.56	
GAT [41]	Best	42.16	15.02	14.5	14.44	11.78	10.26	663
	Last	41.84	14.44	13.98	13.8	11.48	9.74	
FGSM-GA [2]	Best	43.44	18.86	18.44	18.36	16.2	14.28	1054
	Last	43.44	18.86	18.44	18.36	16.2	14.28	
Free-AT(m=8) [39]	Best	38.9	11.62	11.24	11.02	11.00	9.28	1375
	Last	40.06	8.84	8.32	8.2	8.08	7.34	
FGSM-BP (ours)	Best	45.01	21.67	21.47	21.43	17.89	15.36	458
	Last	47.16	20.62	20.16	20.07	15.68	14.15	
FGSM-EP (ours)	Best	45.01	21.67	21.47	21.43	17.89	15.36	458
	Last	46.00	20.77	20.39	20.28	16.65	14.93	
FGSM-MEP (ours)	Best	43.32	**23.8**	**23.4**	**23.38**	**19.28**	**17.56**	458
	Last	45.88	**22.02**	**21.7**	**21.6**	**17.44**	**15.50**	

improve the quality of adversarial examples in different ways. Though they can prevent catastrophic overfitting and improve model robustness, they all require much time for quality improvement. Differently, we improve adversarial example quality from the perspective of initialization and propose to adopt historically generated adversarial perturbation to initialize the adversarial examples without additional calculation cost.

4.3   Results on CIFAR-100

On CIFAR-100, we adopt ResNet18 as the backbone. The results are shown in Table 3. Compared with CIFAR-10, it is hard for the classification model to obtain robustness because the CIFAR-100 covers more classes. The proposed FGSM-MEP achieves comparable robustness to PGD-AT and costs less time than PGD-AT. Specifically, under the AA attack, PGD-AT achieves an accuracy of about 25%, while our FGSM-MEP also achieves an accuracy of about 25%. Note that our training process is about 3 times faster than PGD-AT. Compared with other FAT methods, the proposed FGSM-MEP also achieves the best ro-

Table 5. Comparisons of clean and robust accuracy (%) and training time (minute) using ResNet50 on the ImageNet dataset. Number in bold indicates the best.

ImageNet	Epsilon	Clean	PGD-10	PGD-50	Time (hour)
Free-AT(m=4)[39]	$\epsilon =2$	68.37	48.31	48.28	127.7
	$\epsilon =4$	63.42	33.22	33.08	
	$\epsilon =8$	52.09	19.46	12.92	
FGSM-RS [49]	$\epsilon =2$	67.65	48.78	48.67	44.5
	$\epsilon =4$	63.65	35.01	32.66	
	$\epsilon =8$	53.89	0.00	0.00	
FGSM-BP (ours)	$\epsilon =2$	**68.41**	**49.11**	**49.10**	63.7
	$\epsilon =4$	**64.32**	**36.24**	**34.93**	
	$\epsilon =8$	**53.96**	**21.76**	**14.33**	

bustness against AEs under all attack scenarios on the best and last checkpoints. In terms of training efficiency, we observe similar results on CIFAR-10.

4.4   Results on Tiny ImageNet and ImageNet

Results on Tiny ImageNet. Following the setting of [25,37], we adopt PreActResNet18 to conduct AT. The results are shown in Table 4. Compared with competing FAT methods, our FGSM-MEP achieves higher robust accuracy. Compared with PGD-AT, our FGSM-MEP achieves better robustness performance under all attack scenarios on the best and last checkpoints. In terms of training efficiency, we observe similar results on CIFAR-10 and CIFAR-100.
Results on ImageNet. Following the setting of [39,49], we adopt ResNet50 to conduct experiments. Specifically, ResNet50 is trained with the maximum perturbation strength $\epsilon = 2$, $\epsilon = 4$, and $\epsilon = 8$. The proposed FGSM-EP and FGSM-MEP require memory consumption to store the adversarial perturbation of the last epoch, which limits their application on ImageNet. Fortunately, FGSM-BP does not require memory consumption to conduct AT. Hence, on the ImageNet, we compare our FGSM-BP with FGSM-RS and Free-AT. The results are shown in Table 5. Compared with FGSM-RS and Free-AT, our FGSM-BP achieves the highest clean and robust accuracy. Our FGSM-BP requires a bit more calculation cost than FGSM-RS, but much less time than Free-AT.

4.5   Ablation Study

In this paper, we propose a regularization loss term to enforce function smoothness, resulting in improving model robustness. To validate the effectiveness of the proposed regularization, we adopt ResNet18 as the classification model to conduct ablation experiments on CIFAR-10. The results are shown in Table 6. It can be observed that combined with our regularization method, FGSM-BP, FGSM-EP, and FGSM-MEP can achieve better robustness performance under all attack scenarios.

Table 6. Ablation study of the proposed method.

CIFAR-10		Clean	PGD-50	C&W	AA	Time(min)
FGSM-RS	Best	73.81	41.26	39.84	37.07	51
	Last	83.82	00.02	0.00	0.00	
FGSM-BP w/o regularizer	Best	**86.51**	45.77	44.8	43.30	51
	Last	86.57	44.39	43.82	42.08	
FGSM-EP w/o regularizer	Best	85.97	45.97	44.6	43.39	51
	Last	86.3	44.97	43.8	42.84	
FGSM-MEP w/o regularizer	Best	86.33	46.71	45.5	43.99	51
	Last	**86.61**	45.69	44.8	43.26	
FGSM-RS with regularizer	Best	84.41	50.63	48.76	46.80	73
	Last	84.41	50.63	48.76	46.80	
FGSM-BP with regularizer	Best	83.15	53.2	50.24	47.47	73
	Last	83.09	53.33	50.12	47.17	
FGSM-EP with regularizer	Best	82.75	53.27	49.86	47.94	73
	Last	81.27	53.63	50.12	46.83	
FGSM-MEP with regularizer	Best	81.72	**54.17**	**50.75**	**49.00**	73
	Last	81.72	**54.17**	**50.75**	**49.00**	

5   Conclusion

In this paper, we investigate how to improve adversarial example quality from the perspective of initialization and propose to adopt historically generated adversarial perturbations to initialize adversarial examples. It can generate powerful adversarial examples with no additional calculation cost. Moreover, we also propose a simple yet effective regularizer to further improve model robustness, which prevents the current perturbation deviating too much from the prior-guided initialization. The regularizer adopts both historical and current adversarial perturbations to guide the model learning. Extensive experimental evaluations demonstrate that the proposed method can prevent catastrophic overfitting and outperform state-of-the-art FAT methods at a low computational cost.

References

1. Andriushchenko, M., Croce, F., Flammarion, N., Hein, M.: Square attack: a query-efficient black-box adversarial attack via random search. In: European Conference on Computer Vision. pp. 484–501. Springer (2020)
2. Andriushchenko, M., Flammarion, N.: Understanding and improving fast adversarial training. Advances in Neural Information Processing Systems **33**, 16048–16059 (2020)
3. Bai, J., Chen, B., Li, Y., Wu, D., Guo, W., Xia, S.t., Yang, E.h.: Targeted attack for deep hashing based retrieval. In: European Conference on Computer Vision. pp. 618–634. Springer (2020)
4. Bai, Y., Zeng, Y., Jiang, Y., Wang, Y., Xia, S.T., Guo, W.: Improving query efficiency of black-box adversarial attack. In: European Conference on Computer Vision. pp. 101–116. Springer (2020)
5. Bai, Y., Zeng, Y., Jiang, Y., Xia, S.T., Ma, X., Wang, Y.: Improving adversarial robustness via channel-wise activation suppressing. arXiv preprint arXiv:2103.08307 (2021)
6. Bai, Y., Zeng, Y., Jiang, Y., Xia, S., Ma, X., Wang, Y.: Improving adversarial robustness via channel-wise activation suppressing. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net (2021)
7. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 ieee symposium on security and privacy (sp). pp. 39–57. Ieee (2017)
8. Chen, S.T., Cornelius, C., Martin, J., Chau, D.H.: Robust physical adversarial attack on faster r-cnn object detector. corr abs/1804.05810 (2018). arXiv preprint arXiv:1804.05810 (2018)
9. Croce, F., Hein, M.: Minimally distorted adversarial examples with a fast adaptive boundary attack. In: International Conference on Machine Learning. pp. 2196–2205. PMLR (2020)
10. Croce, F., Hein, M.: Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In: International conference on machine learning. pp. 2206–2216. PMLR (2020)
11. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
12. Duan, R., Ma, X., Wang, Y., Bailey, J., Qin, A.K., Yang, Y.: Adversarial camouflage: Hiding physical-world attacks with natural styles. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1000–1008 (2020)
13. Duan, R., Mao, X., Qin, A.K., Chen, Y., Ye, S., He, Y., Yang, Y.: Adversarial laser beam: Effective physical-world attack to dnns in a blink. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16062–16071 (2021)
14. Finlayson, S.G., Bowers, J.D., Ito, J., Zittrain, J.L., Beam, A.L., Kohane, I.S.: Adversarial attacks on medical machine learning. Science **363**(6433), 1287–1289 (2019)
15. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples (2014)
16. Gu, J., Tresp, V., Hu, H.: Capsule network is not more robust than convolutional network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14309–14317 (2021)

17. Gu, J., Wu, B., Tresp, V.: Effective and efficient vote attack on capsule networks. arXiv preprint arXiv:2102.10055 (2021)
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
19. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: European conference on computer vision. pp. 630–645. Springer (2016)
20. Jia, X., Wei, X., Cao, X., Foroosh, H.: Comdefend: An efficient image compression model to defend adversarial examples. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6084–6092 (2019)
21. Jia, X., Wei, X., Cao, X., Han, X.: Adv-watermark: A novel watermark perturbation for adversarial examples. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 1579–1587 (2020)
22. Jia, X., Zhang, Y., Wu, B., Ma, K., Wang, J., Cao, X.: Las-at: Adversarial training with learnable attack strategy. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13398–13408 (2022)
23. Jia, X., Zhang, Y., Wu, B., Wang, J., Cao, X.: Boosting fast adversarial training with learnable adversarial initialization. IEEE Transactions on Image Processing (2022)
24. Kannan, H., Kurakin, A., Goodfellow, I.: Adversarial logit pairing. arXiv preprint arXiv:1803.06373 (2018)
25. Kim, H., Lee, W., Lee, J.: Understanding catastrophic overfitting in single-step adversarial training. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 8119–8127 (2021)
26. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
27. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. nature **521**(7553), 436–444 (2015)
28. Li, Y., Wu, B., Feng, Y., Fan, Y., Jiang, Y., Li, Z., Xia, S.T.: Semi-supervised robust training with generalized perturbed neighborhood. Pattern Recognition **124**, 108472 (2022)
29. Liang, S., Wei, X., Yao, S., Cao, X.: Efficient adversarial attacks for visual object tracking. In: European Conference on Computer Vision. pp. 34–50. Springer (2020)
30. Liang, S., Wu, B., Fan, Y., Wei, X., Cao, X.: Parallel rectangle flip attack: A query-based black-box attack against object detection. arXiv preprint arXiv:2201.08970 (2022)
31. Lin, J., Song, C., He, K., Wang, L., Hopcroft, J.E.: Nesterov accelerated gradient and scale invariance for adversarial attacks. arXiv preprint arXiv:1908.06281 (2019)
32. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017)
33. Pang, T., Yang, X., Dong, Y., Su, H., Zhu, J.: Bag of tricks for adversarial training. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net (2021)
34. Pang, T., Yang, X., Dong, Y., Xu, K., Zhu, J., Su, H.: Boosting adversarial training with hypersphere embedding. arXiv preprint arXiv:2002.08619 (2020)
35. Park, G.Y., Lee, S.W.: Reliably fast adversarial training via latent adversarial perturbation. arXiv preprint arXiv:2104.01575 (2021)
36. Qian, N.: On the momentum term in gradient descent learning algorithms. Neural networks **12**(1), 145–151 (1999)
37. Rice, L., Wong, E., Kolter, Z.: Overfitting in adversarially robust deep learning. In: International Conference on Machine Learning. pp. 8093–8104. PMLR (2020)

38. Roth, K., Kilcher, Y., Hofmann, T.: Adversarial training is a form of data-dependent operator norm regularization. arXiv preprint arXiv:1906.01527 (2019)
39. Shafahi, A., Najibi, M., Ghiasi, M.A., Xu, Z., Dickerson, J., Studer, C., Davis, L.S., Taylor, G., Goldstein, T.: Adversarial training for free! Advances in Neural Information Processing Systems **32** (2019)
40. Smith, L.N.: Cyclical learning rates for training neural networks. In: 2017 IEEE winter conference on applications of computer vision (WACV). pp. 464–472. IEEE (2017)
41. Sriramanan, G., Addepalli, S., Baburaj, A., et al.: Guided adversarial attack for evaluating and enhancing adversarial defenses. Advances in Neural Information Processing Systems **33**, 20297–20308 (2020)
42. Sriramanan, G., Addepalli, S., Baburaj, A., et al.: Towards efficient and effective adversarial training. Advances in Neural Information Processing Systems **34** (2021)
43. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
44. Wang, X., He, K.: Enhancing the transferability of adversarial attacks through variance tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1924–1933 (2021)
45. Wang, Y., Ma, X., Bailey, J., Yi, J., Zhou, B., Gu, Q.: On the convergence and robustness of adversarial training. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA. pp. 6586–6595. PMLR (2019)
46. Wang, Y., Zou, D., Yi, J., Bailey, J., Ma, X., Gu, Q.: Improving adversarial robustness requires revisiting misclassified examples. In: International Conference on Learning Representations (2019)
47. Wang, Y., Zou, D., Yi, J., Bailey, J., Ma, X., Gu, Q.: Improving adversarial robustness requires revisiting misclassified examples. In: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net (2020)
48. Wei, X., Liang, S., Chen, N., Cao, X.: Transferable adversarial attacks for image and video object detection. arXiv preprint arXiv:1811.12641 (2018)
49. Wong, E., Rice, L., Kolter, J.Z.: Fast is better than free: Revisiting adversarial training. arXiv preprint arXiv:2001.03994 (2020)
50. Wu, D., Xia, S.T., Wang, Y.: Adversarial weight perturbation helps robust generalization. vol. 33, pp. 2958–2969 (2020)
51. Xie, C., Zhang, Z., Zhou, Y., Bai, S., Wang, J., Ren, Z., Yuille, A.L.: Improving transferability of adversarial examples with input diversity. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2730–2739 (2019)
52. Zagoruyko, S., Komodakis, N.: Wide residual networks (2016)
53. Zhang, D., Zhang, T., Lu, Y., Zhu, Z., Dong, B.: You only propagate once: Accelerating adversarial training via maximal principle. arXiv preprint arXiv:1905.00877 (2019)
54. Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., Jordan, M.: Theoretically principled trade-off between robustness and accuracy. In: International Conference on Machine Learning. pp. 7472–7482. PMLR (2019)
55. Zou, W., Huang, S., Xie, J., Dai, X., Chen, J.: A reinforced generation of adversarial examples for neural machine translation. arXiv preprint arXiv:1911.03677 (2019)