Fast Adversarial Training with Smooth Convergence

Mengnan Zhao, Lihe Zhang*, Yuqiu Kong and Baocai Yin

Dalian University of Technology, China

zmnwelcome@mail.dlut.edu.cn, {zhanglihe, yqkong, ybc}@dlut.edu.cn

Abstract

Fast adversarial training (FAT) is beneficial for improving the adversarial robustness of neural networks. However, previous FAT work has encountered a significant issue known as catastrophic overfitting when dealing with large perturbation budgets, i.e. *the adversarial robustness of models declines to near zero during training. To address this, we analyze the training process of prior FAT work and observe that catastrophic overfitting is accompanied by the appearance of loss convergence outliers. Therefore, we argue a moderately smooth loss convergence process will be a stable FAT process that solves catastrophic overfitting. To obtain a smooth loss convergence process, we propose a novel oscillatory constraint (dubbed ConvergeSmooth) to limit the loss difference between adjacent epochs. The convergence stride of ConvergeSmooth is introduced to balance convergence and smoothing. Likewise, we design weight centralization without introducing additional hyperparameters other than the loss balance coefficient. Our proposed methods are attack-agnostic and thus can improve the training stability of various FAT techniques. Extensive experiments on popular datasets show that the proposed methods efficiently avoid catastrophic overfitting and outperform all previous FAT methods. Code is available at* `https://github.com/FAT-CS/ConvergeSmooth`.

1. Introduction

Recent breakthroughs in deep learning [21, 32] have aroused researchers' interest in the security of neural networks [45, 48, 39]. In particular, the advanced research proves the vulnerability of deep models to adversarial attacks [8, 11, 27]. For instance, tiny crafted perturbations can fool models in various fields into making wrong decisions [15, 50, 35, 24]. Considering the security risks brought by adversarial attacks [18, 46, 33, 44], there is a quickly growing body of work [43, 46, 18] on improving the adversarial robustness of neural networks. Among them, adversarial training is widely applied by practitioners [31, 12].

In recent years, projected gradient descent based adversarial training (PGD-AT) [25, 34] has been widely employed for its stability and effectiveness. However, this mechanism is computationally expensive. It requires multiple gradient descent steps to generate the adversarial training data [38]. An alternative of PGD-AT is the fast adversarial training (FAT) [28], which only adopts a single-step fast gradient sign method (FGSM) [11] to generate training data. Compared to PGD-AT, FAT can efficiently train models, but easily falls into catastrophic overfitting [40, 29].

A number of FAT methods have been proposed to mitigate catastrophic overfitting. For example, Wong et al. [40] use randomly initialized perturbations to enhance the diversity of adversarial perturbations. Based on it, Andriushchenko et al. [2] raise a complementary regularizer named GradAlign to maximize the gradient alignment between benign and adversarial examples explicitly. Similarly, NuAT [36] and FGSM-MEP [42] adopt nuclear norm or p-norm to regularize the adversarial training, thereby increasing the prediction alignment between benign and adversarial examples. However, the above methods can only resolve catastrophic overfitting within the limited perturbation budget ($\xi \leq 8/255$). ξ specifies the perturbation degree of adversarial training data generated by various attacks. Besides, models trained by small perturbations are vulnerable to adversarial attacks with a large ξ, *e.g.* the models trained by NuAT and FGSM-MEP at $\xi = 8/255$ perform 53% and 54% robustness against the PGD-50 attack with $\xi = 8/255$, but only 22% and 20% robustness against the same attack with $\xi = 16/255$ respectively. Therefore, we aspire to prevent catastrophic overfitting to improve the adversarial robustness of neural models at larger perturbation budgets.

By analyzing adversarial training processes of representative work, we observe that catastrophic overfitting is usually accompanied by a slight fluctuation in the classification loss for benign samples and a sharp drop in the classification

*Corresponding author

loss for adversarial examples. This motivates us to question whether a smooth loss convergence process is also a stable FAT process. Moreover, we find that an oscillating adversarial training phase may restart the FAT process after catastrophic overfitting. Fig. 1 shows the details.

According to these observations, we introduce an oscillatory constraint that limits the difference in loss between adjacent training epochs, called ConvergeSmooth. A dynamic convergence stride of ConvergeSmooth is designed considering the nonlinear decay rate of loss functions. Inspired by the smoothness of loss convergence, we further verify the effect of the proposed weight centralization on model stability. Weight centralization refers to taking the weights average of the previously trained models as the convergence center of the current model weight. Our proposed methods are attack-agnostic and thus can be combined with existing adversarial strategies in FAT, such as FGSM-RS and FGSM-MEP, to evaluate their performance.

The contributions are summarized in four aspects: **(1)** We verify that previous FAT works still suffer from catastrophic overfitting at a large ξ and then study catastrophic overfitting from the perspective of convergence instability of loss functions; **(2)** We propose a smooth convergence constraint, ConvergeSmooth, and design a dynamic convergence stride for it, to help various FAT methods avoid catastrophic overfitting on different perturbation budgets; **(3)** The weight centralization is proposed without introducing extra hyperparameters other than the loss balance coefficient to stabilize FAT; **(4)** Extensive experiments show that the proposed methods outperform the state-of-the-art FAT techniques in terms of efficiency, robustness, and stability.

2. Related Work

Adversarial attacks. Adversarial attacks are usually used to deceive deep-learning models. Goodfellow et al. [11] first discuss the adversarial attack (FGSM) within the classification task. They prove that adversarial examples x' generated by a single-step gradient backward can misclassify the model $f(\cdot;\theta)$ with high confidence. θ denotes the fixed model weights. x' is generated by

$$x' = x + \xi \cdot \text{sgn}(\nabla_x \mathcal{L}(f(x;\theta), y)), \tag{1}$$

where x is an input image, ξ represents the perturbation budget, sgn$(\cdot)$ means the sign function, $\mathcal{L}(\cdot)$ is usually the cross-entropy loss, $\nabla_x \mathcal{L}(\cdot)$ calculates the gradient of loss at x, and y denotes the ground truth labels of x. Following the FGSM [11], researchers propose a series of attack methods based on the iterative gradient backward, *e.g*. I-FGSM [22], MIM [8], and PGD [25]. Taking the PGD attack as an example, the adversarial example x'_{t+1} produced in the iteration t+1 can be formulated as

$$x'_{t+1} = \text{clip}_\xi(x'_t + \epsilon \cdot \text{sgn}(\nabla_{x'_t} \mathcal{L}(f(x'_t;\theta), y))), \tag{2}$$

where ϵ denotes the single-step stride and clip_ξ refers to projecting adversarial perturbations to a ξ-ball.

Adversarial training. Madry et al. [25] formalize the adversarial training as a min-max optimization problem,

$$\min_\theta \mathbb{E}_{(x,y)\sim D}[\max_{\delta\in[-\xi,\xi]} \mathcal{L}(f(x';\theta), y)], \tag{3}$$

where $x' = x + \delta$, δ represents the adversarial perturbations generated by various attacks such as PGD and FGSM. D is the data generator. The internal maximization maximizes the classification loss to generate adversarial perturbations with fixed model weights. The external minimization minimizes the classification loss on the generated adversarial examples when optimizing the model weights.

Actually, there is a trade-off between computational efficiency and adversarial robustness in recent adversarial training methods. Compared to PGD-AT [25], FGSM-based fast adversarial training (FAT) accelerates the training process but damages the robustness of models for the problem of catastrophic overfitting [19, 17]. To mitigate this issue, Wong et.al [40] demonstrate that the FGSM with a random start strategy (FGSM-RS) can achieve comparable performance against the PGD-AT. ZeroGrad [10] zeroes the elements of gradient that are too small to craft the perturbations. enforces the model loss to increase with the increase in perturbation size. Besides, GradAlign [2] prevents catastrophic overfitting by maximizing the alignment between gradients of benign samples and adversarial examples.

Similarly, NuAT [36] introduces a nuclear norm regularization between logits for benign and adversarial examples and uses the Bernoulli noise as the initial perturbation. ATAS [16] learns an instantiated adaptive step size that is inversely proportional to the gradient norm. It applies the adversarial perturbations from the previous epoch as the initialization of FGSM in the current training phase. In addition, Jia et al. [42] propose several prior-guided initialization methods to replace the random start strategy of FGSM-RS. Specifically, FGSM-BP adopts the adversarial perturbations from the previous batch as the attack initialization in the current batch. FGSM-MEP employs a momentum mechanism to combine all adversarial perturbations from the previous epochs to yield adversarial initialization.

Although these FAT methods have resolved catastrophic overfitting at small perturbation budgets, they still suffer from catastrophic overfitting under larger perturbation budgets (*e.g*. ξ = 16/255). Unlike these methods, we revisit the catastrophic overfitting problem from the perspective of loss convergence instability and prevent this exception by limiting the magnitude of loss fluctuations during training.

3. Proposed Method

In this section, we first study the performance of previous FAT methods when subjected to a large perturbation budget

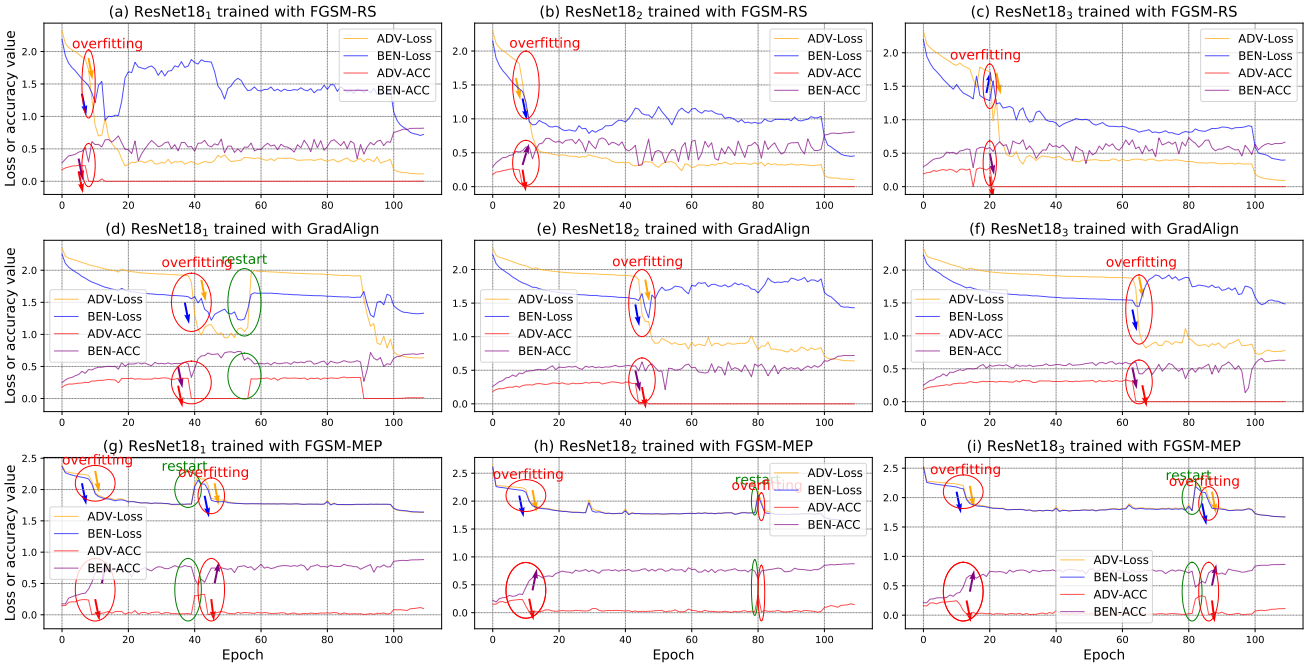

Figure 1. The training process of previous FAT methods with ResNet18 on CIFAR-10. Each FAT method is trained 3 times. ξ = 16/255. ADV-Loss and BEN-Loss denote the classification loss of models to adversarial and benign examples during training, respectively. ADV-Acc and BEN-Acc represent the classification accuracy of models to adversarial and benign examples during testing, respectively.

ξ. Then, the training processes of these methods have been analyzed for a better understanding of catastrophic overfitting. Finally, we detail the proposed methods.

3.1. Performance of FAT methods on a large ξ

Previous FAT techniques avoid catastrophic overfitting at small ξ ($\xi \leq$ 8/255). Here, we investigate their performance on a large ξ (ξ = 16/255 by default).

FGSM-RS [40]: Based on Eq. (3), this method adopts the samples with uniformly random perturbations $\delta_0 \sim \mathcal{U}(-\xi, \xi)$ as the attack initialization of FGSM,

$$\min_{\theta} \mathbb{E}_{(x,y)\sim D}[\max_{\delta_0+\delta\in[-\xi,\xi]} \mathcal{L}(f(x';\theta),y)],$$
$$x' = x + \delta_0 + \delta \quad s.t. \quad x' \in [0,1], \quad \delta_0 \sim \mathcal{U}(-\xi,\xi). \tag{4}$$

GradAlign [2]: This approach increases the gradient alignment between benign samples x and perturbed samples $x + \delta_0$, which is denoted as

$$\mathbb{E}_{(x,y)\sim D}[1 - \cos(\nabla_x \mathcal{L}(x,\theta), \nabla_{x+\delta_0}\mathcal{L}(x+\delta_0,\theta)], \tag{5}$$
$$\mathcal{L}(x,\theta) = \mathcal{L}(f(x;\theta),y).$$

$\cos(\cdot)$ computes the cosine similarity between two matrices.

FGSM-MEP [42]: Different from FGSM-RS, this method generates the initialization perturbations δ_0 based on all historical adversarial perturbations from the previous epochs and introduces a regularization expressed as

$$\mathbb{E}_{(x,y)\sim D}[||f(x';\theta) - f(x+\delta_0;\theta)||_2^2], \tag{6}$$

where $||\cdot||_2^2$ denotes the squared $\mathcal{L}_2$ distance.

Fig. 1 shows their detailed adversarial training processes of ResNet18 [13] on CIFAR-10 [20]. We find that they fall into catastrophic overfitting during the 5$\sim$20$_{\text{th}}$, 40$\sim$65$_{\text{th}}$ and 10$\sim$15$_{\text{th}}$ epochs, respectively. Graphical analysis of various models and datasets is given in the supplement.

3.2. Analysis of training process

From Fig. 1, we discover several typical phenomena of catastrophic overfitting: 1) Change from a smooth convergence state to an irregular fluctuation state; 2) A slight change (increase or decrease) in the standard classification loss $\mathcal{L}(x,\theta)$ and a rapid descent of the adversarial classification loss $\mathcal{L}(x',\theta)$; 3) Rapid decline in the classification accuracy of adversarial examples.

After catastrophic overfitting, the methods depicted in Fig. 1 are capable of restarting a stable FAT process, even though catastrophic overfitting may occur again. The observed phenomena during the FAT restart are: 1) Change from an irregular fluctuation state to a smooth convergence state; 2) Rapid increase in both $\mathcal{L}(x,\theta)$ and $\mathcal{L}(x',\theta)$ after a period of decline; 3) Rapid decrease in the classification accuracy of benign samples, while the classification accuracy of adversarial examples experiences a rapid increase.

On this basis, we can make the following conclusions. 1) $\mathcal{L}(x,\theta)$ is more stable than $\mathcal{L}(x',\theta)$ and models are prone to overfitting to adversarial perturbations; 2) Catastrophic overfitting is closely correlated to the convergence insta-

bility of adversarial training; 3) Exceptions in $\mathcal{L}(x,\theta)$ and $\mathcal{L}(x',\theta)$ occur simultaneously; 4) An oscillating adversarial training phase may trigger the FAT process to restart.

3.3. Smooth convergence for the stable FAT

Next, we describe the proposed method in detail.

Why did previous methods fail to prevent catastrophic overfitting? Despite the improvement in diversity achieved through random initialization in FGSM-RS, models are still susceptible to overfitting adversarial perturbations. GradAlign and FGSM-MEP enhance the stability of adversarial training through the constraints in Eqs. (5) and (6), respectively. However, Eq. (5) may reduce the stability of $\mathcal{L}(x,\theta)$. Meanwhile, the prediction probability $f(x+\delta_0;\theta)$ in Eq. (6) is not the sweet spot to keep the FAT stable. Unlike these methods, we ensure the convergence stability of both $\mathcal{L}(x,\theta)$ and $\mathcal{L}(x',\theta)$.

How can catastrophic overfitting be solved? Since the catastrophic overfitting is usually accompanied by a slight change in $\mathcal{L}(x,\theta)$ and a sharp decline in $\mathcal{L}(x',\theta)$, we consider a smooth loss convergence process to be a stable FAT process that resolves this issue. To this end, a complementary constraint $\mathcal{L}_{CS}$ for Eq. (3) is proposed,

$$\min_{\theta_t} \mathbb{E}_{(x,y)\sim D}[\mathcal{L}(x'_t,\theta_t)+\mathcal{L}_{CS}(t)], \tag{7}$$

which can limit the difference in losses between adjacent epochs, expressed as

$$\begin{aligned}\mathcal{L}_{CS}(t) = w_1\cdot|\mathcal{L}(x'_t,\theta_t)-\mathcal{L}(x'_{t-1},\theta_{t-1})|+\\ w_2\cdot|\mathcal{L}(x,\theta_t)-\mathcal{L}(x,\theta_{t-1})|,\quad s.t.\quad \mathcal{C}(x)=1,\end{aligned} \tag{8}$$

where θ_t represents the model weights of the t_{th} training epoch. w_1 and w_2 are hyper-parameters, $w_1, w_2 \in [0,1]$. $|\cdot|$ calculates the absolute value. $x'_t = x + \delta_{0,t} + \delta_t$. $\delta_{0,t}$ can be replaced by various attack initialization perturbations and δ_t is generated by

$$\max_{\delta_t\in[-\xi,\xi]} \mathcal{L}(x'_t,\theta_{t-1}), \tag{9}$$

where θ_{t-1} is fixed in generating δ_t.

In the practical implementation of Eq. (8), storing and computing $\mathcal{L}(x'_{t-1},\theta_{t-1})$ and $\mathcal{L}(x,\theta_{t-1})$ can consume a significant amount of memory. To overcome this challenge, we introduce u'_{t-1} and u_{t-1} to replace these terms, respectively. $u'_{t-1} = \mathbb{E}_{(x,y)\sim D}\mathcal{L}(x'_{t-1},\theta_{t-1})$ and $u_{t-1} = \mathbb{E}_{(x,y)\sim D}\mathcal{L}(x,\theta_{t-1})$ denote the mathematical expectation of loss in the t-1_{th} training epoch.

We observe that training instability is primarily caused by overfitting or underfitting of small amounts of data. Thus, in Eq. (8), the additional loss term is only applied to partial data selected by the crafted condition $\mathcal{C}(\cdot)$. Notably, exceptions in $\mathcal{L}(x,\theta)$ and $\mathcal{L}(x',\theta)$ occur simultaneously during training. Hence, $\mathcal{L}(x'_t,\theta_t)$ and $\mathcal{L}(x_t,\theta_t)$ share the same condition. $\mathcal{C}$ is constructed by the distance between the pointwise loss $\mathcal{L}(x,\theta_t)$ and the mean value u_{t-1}. This is because over-fitted or under-fitted data often yield excessively high or low classification losses.

$$\mathcal{C}(x) = (|\mathcal{L}(x,\theta_t)-u_{t-1}|\geq\gamma_t), \tag{10}$$

where γ_t ($\gamma_t \geq 0$) represents the convergence stride to select abnormal data. It is crucial to choose an appropriate value for γ_t to ensure effective training. When $\gamma_t = 0$, the adversarial training process may fail to converge as it forces the predictions of all samples to remain unchanged. On the other hand, catastrophic overfitting occurs when $\gamma_t = \infty$, leading to undesirable outcomes. Considering that the loss difference d_{t-1} ($d_{t-1} = |u_{t-1} - u_{t-2}|$) between two adjacent epochs tends to decrease non-linearly, γ_t should be a variable that varies during the training process.

$$\gamma_t = \min(\max(d_{t-1},\gamma_{min}),\gamma_{max}), \tag{11}$$

where γ_{min} and γ_{max} are hyper-parameters. γ_{max} controls the maximum convergence speed. γ_{min} ensures that the training process is not too smooth when $d_{t-1} \to 0$.

Example-based ConvergeSmooth. This approach adds the constraint individually to each sample, described as

$$\begin{aligned}\mathcal{L}^E_{CS}(t) = w_1\cdot|\mathcal{L}(x'_t,\theta_t)-u'_{t-1}|+w_2\cdot|\mathcal{L}(x,\theta_t)-u_{t-1}|,\\ s.t.\quad |\mathcal{L}(x,\theta_t)-u_{t-1}|>\gamma_t.\end{aligned} \tag{12}$$

Batch-based ConvergeSmooth. Likewise, we can apply the complementary constraint to a data batch,

$$\begin{aligned}\mathcal{L}^B_{CS}(t) = w_1\cdot|u'_B-u'_{t-1}|+w_2\cdot|u_B-u_{t-1}|,\\ s.t.\quad |u_B-u_{t-1}|>\gamma_t,\end{aligned} \tag{13}$$

where $u'_B = \mathbb{E}_{(x,y)\sim B}\mathcal{L}(x'_t,\theta_t)$, $u_B = \mathbb{E}_{(x,y)\sim B}\mathcal{L}(x,\theta_t)$.

Weight centralization. To mitigate the problem of manual parameter tuning, we introduce the weight centralization without requiring extra hyperparameters other than the coefficient w_3, $|\theta_t-\theta_{t-1}|=0 \to |\mathcal{L}(x,\theta_t)-\mathcal{L}(x,\theta_{t-1})|=0$,

$$\mathcal{L}^W_{CS}(t) = w_3\cdot||\theta_t - \frac{1}{len(\phi)}\cdot\sum_{j\in\phi}\theta_j||_p, \tag{14}$$

where $\|\cdot\|_p$ means the p-norm function (p = 2), and ϕ denotes a set of previous model weights. The model weights θ_t are restricted to the center $\frac{1}{len(\phi)}\cdot\sum_{j\in\phi}\theta_j$. The reasons behind Eq. (14) are: 1) The initial training process is stable, as indicated in Fig. 1; 2) Models at different training epochs tend to have similar weights after convergence [37].

4. Experimental Results

4.1. Experimental settings

Details. To demonstrate the effectiveness of the proposed method, we conduct comprehensive experiments on

CIFAR10	ξ	FGSM-RS	FGSM-MEP	B-RS	B-MEP
SQ↑	16/255	19.74	22.15	28.83	**29.97**
SQ+RayS↑		18.72	20.96	27.08	**28.45**

Table 1. Quantitative results of FAT methods against the Square and Ray-S attacks on CIFAR-10 with the backbone ResNet18.

CIFAR10	Clean	$\frac{8}{255}$ ♯	$\frac{10}{255}$ ♯	$\frac{12}{255}$ ♯	$\frac{16}{255}$ ♯	SQ+RayS	Time
OAAT[1]	71.30	44.29	38.56	33.68	28.48	21.06	183
AWP[41]	78.86	35.91	32.70	29.89	26.95	19.36	134
ATES[4]	74.52	43.61	38.94	35.31	30.89	22.16	-
ExAT[30]	80.78	49.21	42.15	36.98	31.26	23.08	70
E-MEP	69.84	**53.54**	**49.26**	**45.89**	**40.78**	23.78	104
B-MEP	63.84	50.31	46.77	43.97	40.13	**28.45**	101

Table 2. Quantitative results of methods against various levels of perturbation with ResNet18 as the backbone and CIFAR-10 as the dataset. ♯ denotes the PGD-10 attack.

several benchmark datasets, *i.e*. CIFAR-10 [20], CIFAR-100 [20], and Tiny ImageNet [7].

Following previous works [40, 23, 49, 42], we adopt ResNet18 [13] as the backbone on the CIFAR-10 and CIFAR-100, and choose PreActResNet18 [14] on the Tiny ImageNet. In all experiments, models are optimized using the SGD optimizer with a batch size of 128, weight decay of 5e-4, and momentum of 0.9. The initial learning rates on CIFAR-10, CIFAR-100, and Tiny ImageNet are set as 0.1, 0.1, and 0.01, respectively. Then, we optimize models with a total training epoch of 110 and decay the learning rate at the 100_{th} and 105_{th} epoch with a factor of 0.1.

We apply the proposed ConvergeSmooth in combination with two attack initialization methods on CIFAR-10 and CIFAR-100, FGSM-RS [40] and FGSM-MEP [42]. Additionally, for the Tiny ImageNet, we use FGSM-BP [42] as the initialization method. As mentioned in [42], FGSM-MEP requires consuming memory to store the previous adversarial perturbations, which limits its application on large datasets. Regarding the hyperparameters, we set γ_{max} in Eq. (11) as 0.03, 0.06, and 0.03 for CIFAR-10, CIFAR-100, and Tiny ImageNet, respectively. $\gamma_{max} = 1.5 \cdot \gamma_{min}$ and ξ = 16/255. Specific details of hyperparameter settings for $w_{1\sim3}$ are given in the supplement. All experiments are conducted on a single GeForce RTX 3090 GPU.

Baselines. We include advanced FAT methods as baselines, namely FGSM-RS [40], GradAlign [2], ZeroGrad [10], NuAT [36], ATAS [16], and FGSM-MEP [42].

Evaluation metrics. We follow [40, 16] and adopt FGSM [11], PGD-10 [25], PGD-20 [25], PGD-50 [25], C&W [3], APGD-CE [6] and Autoattack (AA) [6] to evaluate the adversarial robustness of models. PGD-n represents the PGD attack with n iterations. Autoattack combines APGD-CE and APGD-T (targeted APGD) as well as two complementary attacks (FAB [9] and Square attack [26]).

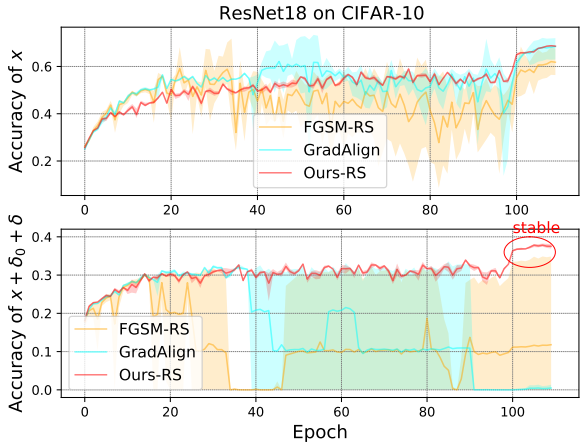

Figure 2. The training process of various FAT methods (PGD-10, ξ = 16/255) on CIFAR-10.

The training easily exhibits fluctuations, and thus individual results can not objectively reflect the performance of methods. For each method, we repeat the training process three times and report the average evaluation results of the best model across the three runs (*mbest*). Additionally, the supplement includes the best results (*best*) and the average evaluation of the final model from the three runs (*mfinal*).

In the following section, 'W-*', 'E-*', and 'B-*' mean 'weight centralization', 'example-based ConvergeSmooth', and 'batch-based ConvergeSmooth', respectively.

4.2. Significance of the Results

Before evaluating the adversarial robustness of trained models, we demonstrate the significance of the results [1]. Specifically, we show that adversarial attacks with ξ = 16/255 rarely change the true label of input. 1) We generate adversarial examples for the FGSM-RS, FGSM-MEP, and our proposed methods on CIFAR10 and CIFAR100 datasets with ξ=16/255 using the non-targeted attacks such as FGSM, PGD, C&W, APGD-CE, and the targeted attack APGD-T. The majority of the generated examples retain their true labels; 2) Tab. 1 examines the robustness of models against an ensemble of the Square (SQ) [26] and Ray-S [5] attacks, as these attacks generate strong oracle-invariant examples [1]; 3) We generate adversarial training data with ξ = 16/255 and test the robustness on various levels of perturbation. Tab. 2 provide comparative experiments between our proposed method and other AT techniques, including OAAT [1], ExAT [30], ATES [4] and AWP [41]. It is important to note that we re-implement these AT methods and apply them in the context of FAT. Overall, our methods achieve optimal performance on both oracle-invariant attacks and classical evaluation attacks. Namely, the classical evaluation attacks used in ξ = 8/255 are reliable and significant even when the value of ξ is increased to 16/255.

Notably, our methods are plugged into FGSM-RS and

Methods		Clean↑	FGSM↑	PGD-10↑	PGD-20↑	PGD-50↑	C&W↑	APGD-CE↑	AA ↑	Time (min)↓
PGD-AT [29]		65.30	46.31	40.73	35.08	33.92	30.84	33.08	26.29	370
FGSM-RS [40]		50.12	38.28	26.13	21.55	20.43	18.96	19.40	14.84	67
GradAlign [2]		58.17	39.87	33.12	26.81	24.99	22.63	23.98	17.02	135
ZeroGrad [10]		74.16	43.96	32.67	21.98	18.37	20.76	16.44	12.07	67
Ours	W-RS	70.66	45.51	36.50	27.51	24.75	23.97	23.38	17.14	71
	E-RS	62.38	42.07	36.78	**30.80**	**28.90**	23.64	**27.71**	17.55	77
	B-RS	65.42	**45.94**	**37.54**	30.01	27.85	**26.28**	26.52	**19.43**	75
NuAT [36]		74.62	44.92	35.22	25.93	23.67	24.07	22.37	18.43	101
ATAS* [16]		64.11	-	31.39	-	28.15	-	-	21.09	-
FGSM-MEP [42]		53.32	36.24	31.85	27.28	26.56	22.10	26.08	18.98	92
Ours	E-MEP	69.84	**47.18**	**40.90**	34.17	32.72	22.69	31.12	17.74	104
	B-MEP	63.84	45.48	40.13	**34.21**	**32.95**	**28.19**	**32.04**	**23.68**	101

Table 3. Quantitative results of various methods (ξ = 16/255) on the CIFAR-10 with ResNet18 as the backbone. 'ATAS*' is the result of ATAS in [16], which is superior to our reproduction. We train each method three times. The results represent the evaluation average between the best models of three training processes. Weight centralization and regularization in MEP do not work together.

Budgets (ξ)	Methods	Clean↑	FGSM↑	PGD-10↑	PGD-20↑	PGD-50↑	C&W↑	APGD-CE↑	AA ↑	Time (min)↓
12/255	GradAlign [2]	66.52	43.06	31.66	27.95	26.04	27.05	25.99	21.65	135
	NuAT [36]	72.79	51.80	41.75	38.60	37.54	35.99	36.71	32.01	101
	FGSM-MEP [42]	**74.71**	52.05	38.45	36.35	35.52	33.05	33.4	27.23	92
	B-MEP (Ours)	72.63	**54.40**	**45.23**	**42.85**	**42.14**	**36.81**	**41.62**	**33.26**	101
10/255	GradAlign [2]	83.10	55.23	36.58	30.47	28.64	31.01	26.51	23.84	135
	NuAT [36]	75.82	55.94	45.54	43.92	43.42	**41.39**	42.91	38.85	101
	FGSM-MEP [42]	**83.43**	**59.51**	42.76	39.28	37.33	37.26	35.91	32.52	92
	B-MEP (Ours)	75.96	57.26	**47.25**	**45.98**	**45.66**	41.00	**45.26**	**39.20**	101

Table 4. Quantitative results of FAT methods on various ξ with ResNet18 as the backbone and CIFAR-10 as the dataset. Models are trained and evaluated under the same ξ. The number in bold indicates the best.

ξ = 16/255	CIFAR10		CIFAR100		Time
Methods	Clean	AA	Clean	AA	hours
GradAlign [2]	55.58	13.45	35.93	7.13	8.79
NuAT [36]	74.25	13.19	20.48	7.30	7.66
FGSM-MEP [42]	65.56	15.43	20.69	6.73	6.15
B-MEP (Ours)	**69.94**	**24.27**	**48.64**	**12.05**	6.72

Table 5. The adversarial accuracy of various FAT methods with WideResNet34-10 as the backbone.

FGSM-MEP, which are fitted to the distribution of adversarial examples. Instead, methods such as OAAT are fitted to the distribution of benign samples and perform better on clean accuracy. Therefore, we use the settings as in OAAT and then apply OAAT and our B-OAAT to the FAT task. B-OAAT realizes 74.12% (+2.82% than OAAT) clean accuracy and 25.39% (OAAT) on SQ+RayS.

4.3. Results on CIFAR-10

We conduct our initial experiments on CIFAR-10 using the ResNet18 backbone. The default ξ is set to 16/255. Tab. 3 presents the main comparisons. The observations are as follows: (1) Compared with previous FAT methods, the proposed approaches achieve optimal adversarial robustness against different attacks, *e.g*. B-RS outperforms all RS-based methods and B-MEP is superior to all other methods. Meanwhile, our methods exhibit similar performance to prior work in terms of clean classification accuracy. (2) B-MEP realizes adversarial robustness approaching PGD-AT, *e.g*. B-MEP performs 32.95% robustness against the PGD-50 attack, only 0.97% lower than PGD-AT; (3) As for time consumption, B-RS takes less time (75 minutes) than GradAlign (135 minutes) and NuAT (101 minutes). This is because ConvergeSmooth only requires the additional regularization when $|u_B - u_{t-1}| > \gamma_t$ instead of adding constraints on all iterations. In addition, B-MEP (102 minutes) takes a bit more calculation cost than FGSM-MEP (92 minutes) but much less than PGD-AT (370 minutes). PGD-AT takes significantly longer than FAT works; (4) B-RS and B-MEP successfully prevent catastrophic overfitting, $e.g.$ *mfinal* (in the supplement) only shows a slight reduction than *mbest* in terms of adversarial robustness.

Fig. 2 visualizes the training process of various FAT approaches. Compared with other methods, the proposed

Methods		Clean↑	FGSM↑	PGD-10↑	PGD-20↑	PGD-50↑	C&W↑	APGD-CE↑	AA↑	Time (min)↓
PGD-AT		40.57	24.72	21.46	17.94	17.38	14.98	17.02	12.63	370
FGSM-RS [40]		30.12	15.24	12.69	10.25	9.79	8.45	9.57	6.90	67
GradAlign [2]		31.91	15.71	12.56	10.28	9.71	8.22	9.47	6.61	135
ZeroGrad [10]		47.31	23.58	17.60	12.85	11.87	11.62	11.01	7.94	67
Ours	W-RS	41.97	**26.30**	19.09	14.77	13.72	12.82	13.36	9.67	72
	E-RS	41.09	22.33	18.78	15.19	**14.43**	12.00	**13.91**	9.50	78
	B-RS	41.47	25.98	**19.44**	**15.36**	14.34	**12.91**	13.71	**9.90**	76
NuAT [36]		31.42	20.39	16.15	13.87	13.29	11.12	12.25	8.32	101
ATAS [16]		55.63	30.35	15.31	10.28	8.49	10.97	6.30	5.05	70
FGSM-MEP [42]		21.39	13.00	11.37	9.93	9.76	7.28	9.58	6.36	92
Ours	E-MEP	44.00	24.46	20.59	17.11	16.59	13.37	16.05	10.97	106
	B-MEP	41.86	**24.86**	**20.84**	**17.30**	**16.59**	**13.96**	**16.25**	**11.38**	102

Table 6. Quantitative results of various methods (ξ = 16/255) with ResNet18 as the backbone on CIFAR-100.

Methods	Clean↑	FGSM↑	PGD-10↑	PGD-20↑	PGD-50↑	C&W↑	APGD-CE↑	AA↑	Time (hour)↓
PGD-AT	32.52	16.47	13.40	10.63	10.24	7.95	10.00	6.41	67.2
FGSM-RS [40]	27.48	12.46	9.59	6.97	6.47	4.99	6.19	3.63	10.5
GradAlign [2]	28.65	13.80	10.40	7.78	7.08	5.75	6.59	3.90	20.9
ZeroGrad [10]	34.66	12.26	8.22	5.29	4.82	3.71	4.23	2.21	10.5
NuAT [36]	35.24	15.52	12.07	8.81	8.13	6.29	7.68	4.27	24.6
FGSM-BP [42]	20.02	10.18	8.38	6.98	6.77	4.56	6.57	3.62	14.3
B-BP (Ours)	33.51	**16.32**	**12.92**	**9.72**	**9.23**	**7.26**	**9.32**	**5.95**	15.4

Table 7. Quantitative results of various methods (ξ = 16/255) with PreActResNet18 as the backbone on Tiny ImageNet.

B-RS achieves optimal training stability and adversarial robustness, *e.g*. the classification accuracy of our method fluctuates slightly during three adversarial training sessions.

Various budgets. Similar experiments are performed for budgets 10/255 and 12/255. Details are given in Tab. 4. It is evident that the proposed method achieves optimal adversarial robustness across various perturbation budgets.

Various networks. We then adopt WideResNet34 with a width factor of 10 [47] as the backbone. The results are given in Tab. 5. Our proposed B-MEP also prevents the wider architectures from catastrophic overfitting.

4.4. Results on CIFAR-100

The results on CIFAR-100 with the backbone ResNet-18 are presented in Tab. 6. (1) FAT methods perform similar training consumption on CIFAR-10 and CIFAR-100 as two datasets contain the same number and size of images; (2) All proposed methods can realize stable adversarial training, as evidenced by the comparable results of *mbest* and *mfinal* (in the supplement); (3) Among the RS-based FAT works, B-RS and E-RS achieve the best and second-best adversarial robustness against different attacks; (4) B-MEP outperforms all other FAT methods; (5) Although B-MEP performs slightly worse than PGD-AT, its training process is approximately 3.5 times faster than this competitor.

4.5. Results on Tiny ImageNet

We conduct experiments on the Tiny ImageNet using the PreActResNet18 backbone to demonstrate the scalability of the proposed method to large datasets. The results are given in Tab. 7. B-BP achieves higher robustness among the FAT methods and comparable robustness to PGD-AT. As for training efficiency, B-BP (15.4 hours) requires slightly more computational cost than FGSM-BP (14.3 hours), but significantly less time than PGD-AT (67.2 hours).

4.6. Ablation studies

In the following, B-RS is selected as the backbone. Figs. 3 and 4 conduct ablation studies on γ_{max} and $\gamma_{max}/\gamma_{min}$ respectively. In Fig. 3, the classification accuracy of both benign and adversarial examples increases with γ_{max}. However, the model trained with a large γ_{max} (>0.09) suffers from catastrophic overfitting. Fig. 4 shows that the model with $\gamma_{max}/\gamma_{min}$ = 1.5 achieves optimal robustness. $\gamma_{max}/\gamma_{min}$ = 1 denotes the static convergence stride. All models can be trained stably when γ_{max} = 0.06.

Tabs. 8 and 9 provide the results of comparative experiments on w_1 and w_2 respectively. We empirically set w_2 to 1 since too small w_2 cannot prevent overfitting and too large w_2 will affect the performance. In Tab. 8, w_1 = 0 achieves optimal adversarial robustness on CIFAR100 and

Dataset	w_1	Clean↑	FGSM↑	PGD-10↑	PGD-20↑	PGD-50↑	C&W↑	APGD-CE↑	AA ↑	Stability
CIFAR100	0.0	**48.13**	**32.13**	**24.25**	**22.67**	**22.21**	**19.04**	**21.29**	**15.28**	⋆⋆⋆
	0.3	46.08	30.58	23.06	21.62	21.22	17.79	20.44	15.22	⋆⋆⋆
	0.5	44.04	29.96	22.64	21.34	20.99	17.55	19.16	13.84	⋆⋆⋆
	0.7	43.07	28.61	22.01	20.84	20.32	16.87	18.98	13.47	⋆⋆⋆
	0.9	40.45	21.05	21.62	20.54	20.15	16.33	18.91	13.53	⋆⋆⋆

Table 8. Quantitative results of the proposed method on various w_1 with ResNet18 as the backbone and the perturbation budget 12/255. 'Stability' represents the number of times the model is stable in three training repetitions. γ_{max} and w_2 are set to 0.03 and 1, respectively.

Dataset	w_2	Clean↑	FGSM↑	PGD-10↑	PGD-20↑	PGD-50↑	C&W↑	APGD-CE↑	AA ↑	Stability
CIFAR100	0.5	36.79	23.97	18.60	17.51	17.23	14.65	16.28	12.23	-
	0.7	37.31	24.36	18.56	17.37	17.26	14.46	15.96	11.54	-
	1.0	**48.13**	**32.13**	**24.25**	**22.67**	**22.21**	**19.04**	**21.29**	**15.28**	⋆⋆⋆
	1.3	44.84	30.42	23.31	21.72	21.31	18.36	19.64	14.72	⋆⋆⋆

Table 9. Quantitative results on various w_2 (γ_{max} = 0.03 and w_1 = 0) with ResNet18 as the backbone and ξ = 12/255.

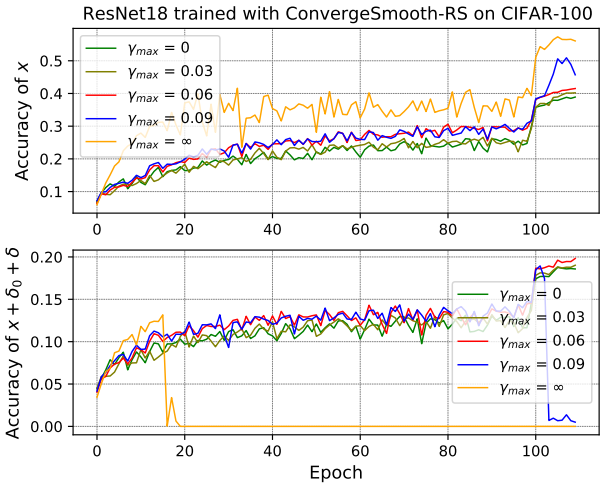

Figure 3. Ablation study of γ_{max} ($\gamma_{max}/\gamma_{min}$ = 1.5). We provide the classification accuracy of models (ResNet-18) to benign and adversarial examples (PGD-10, ξ = 16/255) on CIFAR-100.

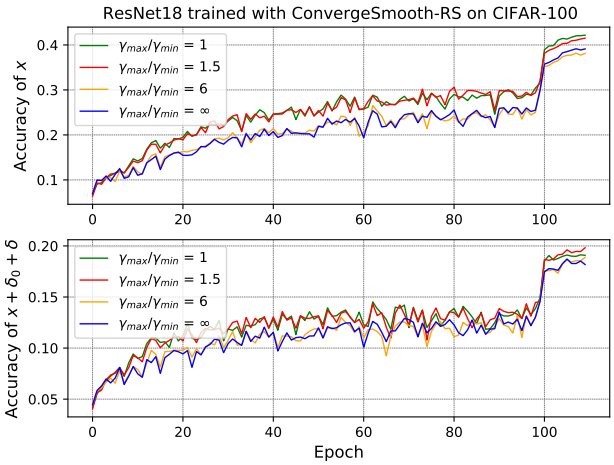

Figure 4. Ablation study of $\gamma_{max}/\gamma_{min}$ (γ_{max} = 0.06). We provide the classification accuracy of models (ResNet-18) to benign and adversarial examples (PGD-10, ξ = 16/255) on CIFAR-100.

all settings accomplish stable FAT. According to Eqs. (7) and (8), if $w_1 \neq 0$, for the data x which satisfies $\mathcal{C}(x) = 1$, $\mathcal{L}(x'_t, \theta_t)$ is assigned with higher weights (1+w_1) or lower weights (1-w_1), causing the model to excessively prioritize or neglect this portion of the data. This may weaken the performance of models. However, for Tiny ImageNet, models suffer catastrophic overfitting at $w_1 = 0$. Then, we gradually increase w_1 from 0.3 with a stride of 0.2 until the FAT process becomes stable. In addition, our weight centralization method only introduces a balance coefficient w_3. w_3 is set to 0.1 since catastrophic overfitting happens when $w_3 = 0$ and classification performance is degraded when $w_3 = 0.2$.

5. Conclusion

In this paper, we tackle the issue of catastrophic overfitting by focusing on the convergence stability of loss functions. Through experimental analysis, we find that this issue is accompanied by abnormal convergence behavior of losses. Motivated by this phenomenon, we propose complementary constraints, namely E-ConvergeSmooth and B-ConvergeSmooth, based on the adversarial training constraint. To alleviate the burden of parameter tuning, weight centralization is designed utilizing priors from previous epochs. Extensive experiments on different network architectures and datasets show that the proposed methods effectively solve catastrophic overfitting and exhibit superior robustness against various adversarial attacks.

Acknowledgments. This work was supported by the National Key R&D Program of China #2018AAA0102000 and the National Natural Science Foundation of China #62276046 and #U19B2039.

References

[1] Sravanti Addepalli, Samyak Jain, Gaurang Sriramanan, and R Venkatesh Babu. Scaling adversarial training to large perturbation bounds. In *European Conference on Computer Vision*, pages 301–316. Springer, 2022.

[2] Flammarion N Andriushchenko M. Understanding and improving fast adversarial training. In *Advances in Neural Information Processing Systems*, pages 16048–16059, 2020.

[3] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*, pages 39–57, 2017.

[4] Supriyo Chakraborty Chawin Sitawarin and David Wagner. Improving adversarial robustness through progressive hardening. *arXiv preprint arXiv:2003.09347*, 2020.

[5] Jinghui Chen and Quanquan Gu. Rays: A ray searching method for hard-label adversarial attack. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1739–1747, 2020.

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

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[8] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9185–9193, 2018.

[9] Croce Francesco and Hein Matthias. Minimally distorted adversarial examples with a fast adaptive boundary attack. In *International Conference on Machine Learning*, pages 2196–2205. PMLR, 2020.

[10] Saberi M. Eskandar M. Golgooni, Z. and M. H Rohban. Zerograd: Mitigating and explaining catastrophic overfitting in fgsm adversarial training. page arXiv preprint arXiv:2103.15476, 2021.

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

[12] Jiaxian Guo, Sidi Lu, Han Cai, Weinan Zhang, Yong Yu, and Jun Wang. Long text generation via adversarial training with leaked information. In *Proceedings of the AAAI conference on artificial intelligence*, 2018.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016.

[15] Zhezhi He, Adnan Siraj Rakin, and Deliang Fan. Parametric noise injection: Trainable randomness to improve deep neural network robustness against adversarial attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 588–597, 2019.

[16] Zhichao Huang, Yanbo Fan, Chen Liu, Weizhong Zhang, Yong Zhang, Mathieu Salzmann, Sabine Süsstrunk, and Jue Wang. Fast adversarial training with adaptive step size. *arXiv preprint arXiv:2206.02417*, 2022.

[17] Xiaojun Jia, Yong Zhang, Baoyuan Wu, Jue Wang, and Xiaochun Cao. Boosting fast adversarial training with learnable adversarial initialization. *IEEE Transactions on Image Processing*, 31:4417–4430, 2022.

[18] Yunhan Jia Jia, Yantao Lu, Junjie Shen, Qi Alfred Chen, Hao Chen, Zhenyu Zhong, and Tao Wei Wei. Fooling detection alone is not enough: Adversarial attack against multiple object tracking. In *International Conference on Learning Representations (ICLR'20)*, 2020.

[19] Hoki Kim, Woojin Lee, and Jaewook Lee. Understanding catastrophic overfitting in single-step adversarial training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8119–8127, 2021.

[20] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[22] Goodfellow I.J. Bengio S. Kurakin, A. Adversarial machine learning at scale. In *International Conference on Learning Representations (ICLR)*, 2017.

[23] Bai Li, Shiqi Wang, Suman Jana, and Lawrence Carin. Towards understanding fast adversarial training. *arXiv preprint arXiv:2006.03089*, 2020.

[24] Qizhang Li, Yiwen Guo, and Hao Chen. Yet another intermediate-level attack. In *European Conference on Computer Vision*, pages 241–257. Springer, 2020.

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

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

[27] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*, pages 372–387. IEEE, 2016.

[28] Geon Yeong Park and Sang Wan Lee. Reliably fast adversarial training via latent adversarial perturbation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7758–7767, 2021.

[29] Leslie Rice, Eric Wong, and Zico Kolter. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*, pages 8093–8104. PMLR, 2020.

[30] Amirreza Shaeiri, Rozhin Nobahari, and Mohammad Hossein Rohban. Towards deep learning models resistant to large perturbations. *arXiv preprint arXiv:2003.13370*, 2020.

[31] Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! volume 32, 2019.

[32] Vaishaal Shankar, Rebecca Roelofs, Horia Mania, Alex Fang, Benjamin Recht, and Ludwig Schmidt. Evaluating machine accuracy on imagenet. In *International Conference on Machine Learning*, pages 8634–8644. PMLR, 2020.

[33] Vasu Sharma, Ankita Kalra, Sumedha Chaudhary Vaibhav, Labhesh Patel, and Louis-Phillippe Morency. Attend and attack: Attention guided adversarial attacks on visual question answering models. In *Proc. 32nd Conf. Neural Inf. Process. Syst.(NeurIPS)*, pages 1–6, 2018.

[34] Chuanbiao Song, Kun He, Liwei Wang, and John E Hopcroft. Improving the generalization of adversarial training with domain adaptation. In *International Conference on Learning Representations (ICLR)*, 2019.

[35] Gaurang Sriramanan, Sravanti Addepalli, Arya Baburaj, et al. Guided adversarial attack for evaluating and enhancing adversarial defenses. *Advances in Neural Information Processing Systems*, 33:20297–20308, 2020.

[36] Gaurang Sriramanan, Sravanti Addepalli, Arya Baburaj, et al. Towards efficient and effective adversarial training. volume 34, pages 11821–11833, 2021.

[37] Vale Tolpegin, Stacey Truex, Mehmet Emre Gursoy, and Ling Liu. Data poisoning attacks against federated learning systems. In *Computer Security–ESORICS 2020: 25th European Symposium on Research in Computer Security, ESORICS 2020, Guildford, UK, September 14–18, 2020, Proceedings, Part I 25*, pages 480–501. Springer, 2020.

[38] Jianyu Wang and Haichao Zhang. Bilateral adversarial training: Towards fast training of more robust models against adversarial attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6629–6638, 2019.

[39] Zhipeng Wei, Jingjing Chen, Zuxuan Wu, and Yu-Gang Jiang. Cross-modal transferable adversarial attacks from images to videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15064–15073, 2022.

[40] Kolter J Z. Wong E, Rice L. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations (ICLR)*, 2020.

[41] Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. *Advances in Neural Information Processing Systems*, 33:2958–2969, 2020.

[42] Xingxing Wei Baoyuan Wu Ke Ma Jue Wang Xiaochun Cao Xiaojun Jia, Yong Zhang. Prior-guided adversarial initialization for fast adversarial training. In *Proceedings of the European conference on computer vision (ECCV)*, 2022.

[43] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. Adversarial examples for semantic segmentation and object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1369–1378, 2017.

[44] Yi Xie, Zhuohang Li, Cong Shi, Jian Liu, Yingying Chen, and Bo Yuan. Enabling fast and universal audio adversarial attack using generative model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[45] Han Xu, Yao Ma, Hao-Chen Liu, Debayan Deb, Hui Liu, Ji-Liang Tang, and Anil K Jain. Adversarial attacks and defenses in images, graphs and text: A review. *International Journal of Automation and Computing*, 17(2):151–178, 2020.

[46] Yan Xu, Baoyuan Wu, Fumin Shen, Yanbo Fan, Yong Zhang, Heng Tao Shen, and Wei Liu. Exact adversarial attack to image captioning via structured output learning with latent variables. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4135–4144, 2019.

[47] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[48] Jie Zhang, Bo Li, Jianghe Xu, Shuang Wu, Shouhong Ding, Lei Zhang, and Chao Wu. Towards efficient data free black-box adversarial attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15115–15125, 2022.

[49] Yihua Zhang, Guanhua Zhang, Prashant Khanduri, Mingyi Hong, Shiyu Chang, and Sijia Liu. Revisiting and advancing fast adversarial training through the lens of bi-level optimization. In *International Conference on Machine Learning*, pages 26693–26712. PMLR, 2022.

[50] Wenqiang Zheng and Yan-Fu Li. Mc-fgsm: Black-box adversarial attack for deep learning system. In *2021 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW)*, pages 154–159. IEEE, 2021.