Subspace Adversarial Training

Tao Li Yingwen Wu Sizhe Chen Kun Fang Xiaolin Huang
Department of Automation, Shanghai Jiao Tong University
{li.tao, yingwen_wu, sizhe.chen, fanghenshao, xiaolinhuang}@sjtu.edu.cn

Abstract

*Single-step adversarial training (AT) has received wide attention as it proved to be both efficient and robust. However, a serious problem of catastrophic overfitting exists, i.e., the robust accuracy against projected gradient descent (PGD) attack suddenly drops to 0% during the training. In this paper, we approach this problem from a novel perspective of optimization and firstly reveal the close link between the fast-growing gradient of each sample and overfitting, which can also be applied to understand robust overfitting in multi-step AT. To control the growth of the gradient, we propose a new AT method, **Sub**space **A**dversarial **T**raining (**Sub-AT**), which constrains AT in a carefully extracted subspace. It successfully resolves both kinds of overfitting and significantly boosts the robustness. In subspace, we also allow single-step AT with larger steps and larger radius, further improving the robustness performance. As a result, we achieve state-of-the-art single-step AT performance. Without any regularization term, our single-step AT can reach over $\mathbf{51}\%$ robust accuracy against strong PGD-50 attack of radius $8/255$ on CIFAR-10, reaching a competitive performance against standard multi-step PGD-10 AT with huge computational advantages. The code is released at* https://github.com/nblt/Sub-AT.

1. Introduction

Adversarial training (AT) [23], which aims to minimize the model's risk under the worst-case perturbations, is currently the most effective approach for improving the robustness of deep neural networks. For a given neural network $f(\mathbf{x}, \mathbf{w})$ with parameters $\mathbf{w}$, the optimization objective of AT can be formulated as follows:

$$\min_{\mathbf{w}} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \left[\max_{\delta\in\mathcal{B}(\mathbf{x},\epsilon)} \mathcal{L}\left(f(\mathbf{x}+\delta, \mathbf{w}), y\right) \right],$$

where $\mathcal{B}(\mathbf{x}, \epsilon)$ is the norm ball with radius ϵ and $\mathcal{L}$ is the loss function. The key issue of AT lies in solving the inner worst-case problem by generating adversarial examples.

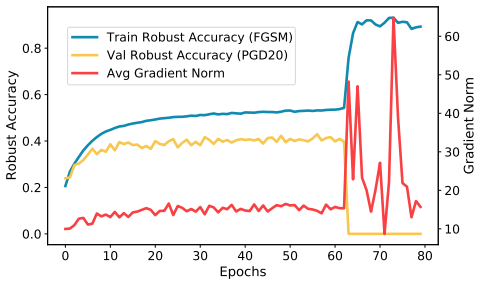


Figure 1. *Catastrophic overfitting* in single-step AT. The experiments are conducted on CIFAR-10 with PreAct ResNet18 model for adversarial robustness against ℓ_∞ perturbations of radius $8/255$. The robust accuracy of single-step Fast AT on the validation set against PGD-20 attack abruptly drops to 0 in one single epoch, characterized by a rapid explosion of the average gradient norm of each sample.

Presently the most efficient way to generate adversarial examples is the fast gradient sign method (FGSM) [9], i.e.,

$$\mathbf{x}^{\text{adv}} = \mathbf{x} + \epsilon \cdot \text{sgn}\left(\nabla_{\mathbf{x}} \mathcal{L}(f(\mathbf{x}, \mathbf{w}), y)\right).$$

Since the adversarial examples above are generated by one-step gradient propagation, the corresponding AT is called *single-step* AT. In Fig. 1, we demonstrate a standard single-step AT process where the training robust accuracy against FGSM attack keeps increasing. However, the generalization capability, i.e., the robust accuracy on the validation set under projected gradient descent (PGD) attack [23], can suddenly drop to zero, which is a typical overfitting phenomenon referred as *catastrophic overfitting* [40].

Many works [1, 16, 17, 32, 37, 40] are devoted to resolving such an intriguing overfitting problem. One approach to tackle the overfitting is to use a judiciously designed learning rate schedule as well as appropriate regularizations. For example, Wong *et al*. [40] proposed to add a random step to FGSM and introduce cyclic learning rates [30] to overcome the overfitting. Andriushchenko *et al*. [1] proposed a novel regularization term called GradAlign to further improve the quality of single-step AT solutions. However,

these methods highly rely on specifically designed learning rate schedules, which need to be tuned carefully for different tasks. Another approach is to generate more precise adversarial examples. For example, Kim *et al*. [17] suggested verifying the inner interval along the adversarial direction and searching for appropriate step size. PGD AT, a typical *multi-step* AT which generates adversarial examples using multiple iterations, can also help avoid catastrophic overfitting. However, these methods require multiple forward propagations. More seriously, overfitting can still prominently occur in multi-step AT (known as *robust overfitting*) as demonstrated by Rice *et al*. [28].

In order to understand this interesting phenomenon, let us investigate what happens at the 64-th epoch in Fig. 1 when catastrophic overfitting occurs. Before the overfitting, the training robust accuracy has already stepped into a stable stage, indicating the small norm of batch gradient $\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_{\mathbf{w}}\mathcal{L}(f(\mathbf{x}_i^{\text{adv}},\mathbf{w}),y)\right\|_2$ (n denotes the batch size). There are two possibilities for the small batch gradient: ***i***) the gradient of each sample is small; ***ii***) the gradients of samples does not converge, but they cancel each other, resulting in an overall balanced state. We then plot the average norm of each sample's gradient on one fixed training batch (i.e., $\frac{1}{n}\sum_{i=1}^{n}\left\|\nabla_{\mathbf{w}}\mathcal{L}(f(\mathbf{x}_i^{\text{adv}},\mathbf{w}),y)\right\|_2$) in red. An interesting thing is that before the overfitting, the average norm stays almost constant. However, it abruptly increases in the moment when the overfitting occurs. Intuitively, at that time, the balance of gradient is broken — the network tries to capture each sample's label with huge fluctuations, namely large gradients, a significant signal of overfitting. This phenomenon also coincides with the recent discussion on the connection between the gradient variance and generalization capability [12, 15, 25].

Inspired by the link between large gradients and overfitting, we propose to resolve the overfitting by controlling the magnitude of the gradient. A possible way is to restrict the gradient descent in a subspace instead of the whole parameter space, to prevent the excessive growth of the gradient. The key challenge lies in keeping the network's capability in such a subspace, which has been recently discussed in [20] showing that, optimizing parameters in a tiny subspace extracted from training dynamics could keep the performance. Based on this discovery, we propose a new AT method called ***Sub**space **A**dversarial **T**raining* (***Sub-AT***), which identifies such an effective subspace and conducts AT in it. From the training statistics of Sub-AT in Fig. 2a, we observe that it successfully controls the average gradient norm under a low level (the yellow dotted curve), thus resolving the catastrophic overfitting. Meanwhile, the robust accuracy is significantly improved from 0.4 to nearly 0.5 (the yellow solid curve). The sensitivity to learning rates is also fundamentally overcome as we only use a constant learning rate, and the results remain similar across a wide range of choices. As a direct extension, Sub-AT can be applied to mitigate the robust overfitting (Fig. 2b) in multi-step AT, implying the similar essence behind these two phenomena. Thus for the first time, the two overfittings, which were previously treated separately [1], are now connected and resolved in a unified approach.

Since training in subspace controls the gradient magnitude and hence fundamentally resolves the catastrophic overfitting, we now can allow larger steps and radius, which previously requires the assistance of delicate regularizations, e.g. GradAlign [1]. It brings further improvement on robustness, from which it follows that pure single-step-based AT (without regularization terms) achieves competitive robustness with standard multi-step PGD AT with great computational benefits, answering a long-existing question:

Can single-step AT achieve comparable robustness against iterative attacks than multi-step AT?

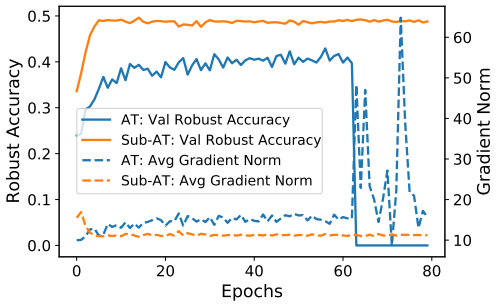
Our Sub-AT uncovers the long-neglected potential of single-step AT and can enlighten more efficient and powerful AT algorithms.

Our main contributions can be summarized as follows:

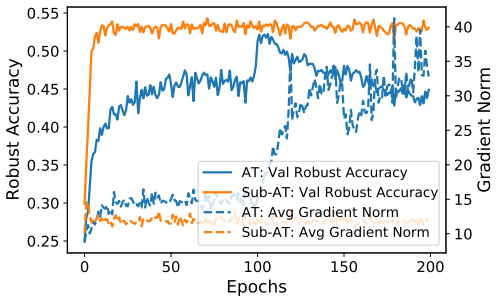
- We approach the *catastrophic overfitting* in single-step AT from a novel view of optimization and firstly reveal the close link between the fast-growing gradient of each sample and overfitting, which can also be applied to explain the *robust overfitting* in multi-step AT.
- We propose an efficient AT method, Sub-AT, which constrains AT in a carefully extracted subspace, to control the growth of gradient. It uniformly resolves both kinds of overfitting, significantly improves the robustness, and successfully overcomes the sensitivity to learning rates. It is also very easy to combine with other AT methods to bring consistent improvements.
- Our Sub-AT achieves *state-of-the-art* adversarial robustness on single-step AT and can successfully train with larger steps and larger radius, which brings further improvements. Notably, our pure single-step AT achieves over 51% robust accuracy against PGD-50 attack of $\epsilon = 8/255$ on CIFAR-10, competitive to the multi-step PGD-10 AT with great time benefits.

2. Related Work

Adversarial Training. Since deep neural networks are easily fooled by adversarial examples, many defense methods [4, 5, 8, 21–24, 27, 31, 33, 38, 39, 41, 43–45] have been proposed. Among them, AT [23], which augments the training data with adversarial perturbations, is currently the most effective way to improve the robustness of the model. According to the number of times the gradient propagation involved in generating adversarial perturbations, AT



(a) Single-step AT.



(b) Multi-step AT.

Figure 2. Resolving *catastrophic overfitting* in single-step AT and *robust overfitting* in multi-step AT. The experiments are conducted on CIFAR-10 with PreAct ResNet18 model for robustness against ℓ_∞ perturbations of radius $8/255$. In both overfittings, robust accuracy on the validation set degenerates along with an abrupt increase of the average grad norm of the sample. Our Sub-AT successfully controls the rapid increase of the average grad norm, thereby resolving both two overfittings uniformly and significantly improving the robustness.

can be mainly divided into two classes: single-step AT [1, 17, 29, 32, 40] and multi-step AT [23, 38, 44]. Single-step AT has proven to be both efficient and robust [29, 40] and thus receives much attention. For example, Free AT [29] achieves remarkable robustness performance using a single-step gradient with redundant batches and accumulative perturbations. Multi-step AT, such as PGD AT [23] and TRADES [44], generally provides better robustness guarantees than single-step AT as it generates strong adversarial perturbations. However, its computational cost is relatively high as multiple forward and back propagations are required during batch training.

Overfitting in Adversarial Training. Both single-step AT and multi-step AT suffer from overfitting problems (known as catastrophic overfitting [40] and robust overfitting [28], respectively) where the robust test accuracy suddenly begins to decrease as the training proceeds. The problem can be more severe in single-step AT as the robust test accuracy against PGD attack can abruptly drop to 0% only in one epoch. Many works are devoted to resolving such an intriguing overfitting problem. Among them, Wong *et al*. [40] first suggested adding a random step to FGSM and adopting cyclic learning rates, which provides competitive robustness against PGD AT with significant time advantages. Andriushchenko *et al*. [1] designed a novel regularization term called GradAlign to improve the gradient alignment inside the perturbation set and provide better robustness. Sriramanan *et al*. [32] introduced a relaxation term to find more suitable gradient-directions for attack. However, these methods rely on a judiciously selected learning rate schedule, or a proper regularization coefficient [17]. Towards understanding the overfitting, Vivek *et al*. [36] discovered that models trained via single-step AT learn to prevent the model from generating effective adversarial examples and introduced dropout scheduling to mitigate it. Kim *et al*. [17] observed the distortion of the sample-wise decision boundary during the overfitting and suggested verifying the adversarial examples along the adversarial direction. However, their explanations are limited to single-step AT, and the robustness performance is still inferior. In this work, we understand overfitting from a general perspective of optimization and explain both kinds of overfitting in AT uniformly, bringing a huge improvement in robustness.

Training in Subspace. Many works focus on the low-dimensionality essence of neural network training [11, 34, 35]. The pioneering work [19] first proposed to train neural networks in a reduced subspace via random projection and discovered that, the required dimension to obtain 90% performance of regular training is far less than the original parameters' dimension. The following work [10] improved the random bases training by considering different layers and re-drawing the random bases at each step. Different from random projection, Li *et al*. [20] proposed to train neural networks in low-dimensional subspaces extracted from training dynamics, and obtained comparable performance as regular training. We also take advantage of the subspace extracted from the training dynamics of AT and constrain the training in it, thereby successfully controlling the magnitude of the gradient and keeping the training performance.

3. Methodology

3.1. Investigating Catastrophic Overfitting

First, we focus on an interesting phenomenon in Fig. 3a: when catastrophic overfitting occurs, the natural accuracy on training data goes through a collapse. Recall that at the same time, the robust training accuracy goes through a sudden increase, as illustrated in Fig. 1. These two phenomena suggest that before the overfitting, the network learns robust features that benefit both robust and natural accuracy. However, when overfitting occurs, the network turns to capture the adversarial information in training data (with adversar-

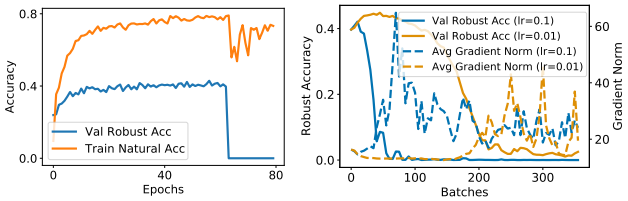
ial perturbations), which harms natural accuracy, i.e., generalization capability to natural examples. Further, it loses generalization capability to new adversarial examples generated by PGD attacks, as they may lie very close to natural examples. The adversarial information is so "hard" to learn that, the network has to go through a huge fluctuation — and eventually overfit it — resulting in nearly zero robust accuracy on test data.

Since adversarial examples are relatively "hard" to learn, we pay attention to the evolution of each sample's gradient and consider the average norm of the gradient to analyze the training status. Specifically, we record the robust accuracy of Fast AT [40] against PGD-20 attack and the average norm of the gradient in one fixed training batch with size $n = 256$ during the training (the estimations for batch normalization [14] are frozen for sample-wise gradient estimation). Fig. 1 illustrates the statistics when the catastrophic overfitting occurs, where we observe that the abrupt decrease of the robust accuracy highly coincides with the sudden increase of the average gradient norm. This phenomenon implies that during the catastrophic overfitting, the gradient of each sample suddenly increases, resulting in a huge fluctuation in training and, eventually, significant degeneration on robust generalization.

To further investigate the link between increasing gradients and overfitting, we examine the detailed statistics in the 64th epoch, when catastrophic overfitting happens. We record the statistics after each iteration. For comparison, we also consider decaying the learning rate from 0.1 to 0.01 before the 64th epoch training, as increasing gradient indicates an excessive learning rate. The results are illustrated in Fig. 3b, where we observe that although for both learning rates catastrophic overfitting eventually occurs, a smaller learning rate could help. It achieves better robustness accuracy and remarkably postpones the overfitting with a better-controlled average gradient norm (the yellow dotted curve). We conclude the findings as follows: ***i***) the overfitting is closely related to the fast-growing of the average gradient norm, and a delicately chosen learning rate could help suppress the growth of the gradient, which results in recent advances on adopting heuristic learning rates [40]; ***ii***) to control the catastrophic overfitting, we have to control the growth of the average gradient norm.

3.2. Controlling the Gradient Magnitude

Let us focus on how to control the gradient magnitude of each sample during the training. Our main idea is to constrain the gradient descent of AT in a low-dimensional subspace instead of the whole parameter space, thus implicitly suppressing the fast growth of the gradient. First, we consider how to obtain such a subspace that is effective for AT. Recently, Li *et al*. [20] proposed an algorithm called DLDR, which effectively extracts a low-dimensional subspace for



(a) (b)

Figure 3. (a) natural accuracy on training data also collapses at the overfitting; (b) the variation of the statistics in the 64th epoch: switching to a smaller learning rate could alleviate the catastrophic overfitting with a better-controlled gradient norm.

optimization from the training trajectory. It generally contains two steps:

- **Step 1:** sample model checkpoints $\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_t\}$ during the regular training where we align the model parameters as a vector $\mathbf{w}_i$ with length of the parameters' number N;
- **Step 2:** perform singular vector decomposition (SVD) on the aligned parameter matrix $[\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_t]$ and obtain mutually orthogonal bases of the subspace $[\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_d]$ with dimensionality d.

In this work, we apply DLDR algorithm [20] to extract the effective subspace for AT. Note that we sample the model parameters *before* the overfitting occurs since in this period the network learns robust features beneficial for both robust and natural accuracy, as demonstrated in Sec. 3.1. We expect that optimizing the network in such an extracted subspace could overcome the overfitting and obtain good robustness. After extracting the subspace, we rewind the model parameters to initialization and constrain AT optimization in subspace by projecting the gradient onto it. The detailed algorithm is summarized in Algorithm 1.

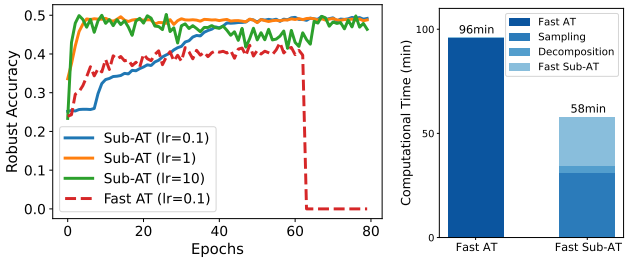
Sampling Strategy. We adopt a simple sampling strategy for DLDR: sampling twice uniformly in each epoch training. A more delicate strategy is promising to improve the performance. For sampling epochs, we expect that the best performance will be obtained with sampling right before the overfitting. Sampling in the start of the training is not good, as the subspace cannot be well estimated. We conduct DLDR with a bit more epochs in a safe region when the overfitting certainly has not happened. The dimensionality of the subspace d is set to 80 on CIFAR-10 for single-step AT by default. We provide a detailed sampling strategy in Appendix A.

Training Performance. In Fig. 2, we demonstrate that Sub-AT successfully controls the average gradient norm under a constant low level, thereby resolving both catastrophic

Algorithm 1 Subspace Adversarial Training (Sub-AT)

Require: The dimensionality of the subspace d, the number of sampling times t for DLDR, learning rate α, batch size n and training data $\{(\mathbf{x}_i, y_i)\}$;

Phase 1: obtaining the orthonormal bases of subspace $[\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_d]$;
Sample a parameter trajectory $\{\mathbf{w}_1^s, \mathbf{w}_2^s, \dots, \mathbf{w}_t^s\}$ along AT training with a certain strategy;
$\overline{\mathbf{w}} = \frac{1}{t}\sum_{i=1}^{t} \mathbf{w}_i^s$;
$W = [\mathbf{w}_1^s - \overline{\mathbf{w}}, \mathbf{w}_2^s - \overline{\mathbf{w}}, \dots, \mathbf{w}_t^s - \overline{\mathbf{w}}]$;
Perform spectral decomposition on $W^\top W$ and obtain the largest d eigenvalues $[\sigma_1^2, \sigma_2^2, \dots, \sigma_d^2]$ with corresponding eigenvectors $[\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_d]$;
$\mathbf{u}_i = \frac{1}{\sigma_i} W \mathbf{v}_i$;
Phase 2: conducting AT in extracted subspaces;
$k \leftarrow 0$;
$P = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_d]$;
$\mathbf{w}_0 = \mathbf{w}_1^s$;
while not converging **do**
    Sample mini-batch data $\{(\mathbf{x}_i, y)\}_{i=1}^n$;
    Generate adversarial examples $\{\mathbf{x}_i^{\text{adv}}\}_{i=1}^n$;
    $\mathbf{g}_k^{\text{adv}} \leftarrow \frac{1}{n}\sum_{i=1}^{n} \nabla_{\mathbf{w}} \mathcal{L}(f(\mathbf{x}_i^{\text{adv}}, \mathbf{w}_k), y)$;
    $\mathbf{w}_{k+1} \leftarrow \mathbf{w}_k - \alpha P \left(P^\top \mathbf{g}_k^{\text{adv}}\right)$;
    $k \leftarrow k + 1$;
end while
Return $\mathbf{w}_k$;



(a) (b)

Figure 4. (a) Sub-AT is extremely robust to learning rate; (b) detailed time consumption analysis on CIFAR-10.

and robust overfitting meanwhile significantly improving the robustness performance. Then in Fig. 4a, we demonstrate that Sub-AT is highly robust to a wide range of learning rates and can converge only with a constant learning rate. Notably, with a large learning rate, Sub-AT can converge quickly in a few epochs and keep the performance without overfitting. Thus Sub-AT fundamentally overcomes the sensitivity to learning rates and obtains true robustness to overfitting. In this work, we set the constant learning rate as 1 by default for both stable and efficient training.

Computational Analysis. The computational overhead of Sub-AT consists of two parts: DLDR and subspace training. The DLDR part contains two steps: sampling and decomposition. The time consumption on the decomposition is negligible compared to the sampling as it only involves a spectral decomposition of a $t \times t$ matrix and two matrix productions. The training of Sub-AT has almost the same computational complexity as the standard AT with little additional cost on the gradient projection. Detailed computational analysis on CIFAR-10 is illustrated in Fig. 4b. We observe that Sub-AT reduces total training time overhead compared to standard AT training as it only samples a piece of the trajectory with efficient training in the subspace.

4. Experiments

In this section, we conduct comprehensive experiments to verify the effectiveness of Sub-AT in resolving the overfitting of both single-step and multi-step AT. We first demonstrate that Sub-AT successfully resolves catastrophic overfitting and achieves state-of-the-art robustness performance in single-step AT. Then we show that after obtaining the subspace, Sub-AT allows larger steps and larger radius, which further improves the robustness. Finally, we apply Sub-AT to mitigate robust overfitting in multi-step AT and show that leveraging the subspace, weak single-step AT is able to achieve even better robustness than multi-step AT, revealing the great potential of single-step AT.

4.1. Experiment Setup

Datasets. Three datasets are considered in our experiments: CIFAR-10/100 [18] and Tiny-ImageNet [7]. We randomly split the original training set as training and validation set according to a ratio of 9:1 [3]. Due to the limited space, we place the Tiny-ImageNet results in Appendix C.

Attack. We consider two typical types of adversarial perturbations: ℓ_∞ norm with radius $\epsilon = 8/255$ and ℓ_2 norm with radius $\epsilon = 128/255$. For single-step AT, we focus on ℓ_∞ norm attack and use the recommended step size of $\alpha = 1.25\epsilon$ described by Wong *et al.* [40]. For multi-step AT, we generate adversarial perturbations with 10 steps attacks of step size $\alpha = 2/255$ for ℓ_∞ norm and $\alpha = 15/255$ for ℓ_2 norm, the standard PGD AT following the setting of Rice *et al.* [28]. We consider PGD-20 [28], PGD-50 (with 50 iterations and 10 restarts) [40] and also Auto-Attack, a strong and reliable attack recently proposed by [6], for a rigorous evaluation on robustness.

Training. For all experiments, we use PreAct ResNet-18 [13] model as a default choice. Experiments with Wide-ResNet [42] can be found in Appendix D. Three learning rate schedules are considered: ***i***) *cyclic* schedule [1, 40] which can help overcome the overfitting; ***ii***) *piecewise* schedule, i.e., training the model for 200 epochs with an initial learning rate 0.1 and decaying by ten at 100 and 150

Table 1. Performance comparisons of single-step AT on CIFAR-10/100. The robustness is evaluated under **PGD-50** attack.

Schedule	Method	Subspace	CIFAR-10				CIFAR-100			
			Robust Accuracy		Natural Accuracy		Robust Accuracy		Natural Accuracy	
			Best	Final	Best	Final	Best	Final	Best	Final
cyclic	Fast AT	–	45.82	45.69	82.36	83.26	16.72	0.00	34.51	47.23
cyclic	GradAlign	–	47.02	46.73	80.43	81.34	24.57	24.22	50.82	51.92
piecewise	Fast AT	–	39.95	0.00	73.13	89.93	17.84	0.00	41.54	46.46
piecewise	Single AT	–	35.48	32.43	83.68	86.86	16.18	0.91	55.88	59.15
piecewise	Free AT	–	47.30	47.00	79.37	79.98	23.50	22.93	50.91	51.64
piecewise	GradAlign	–	42.16	0.02	71.64	88.77	23.80	15.60	49.30	53.40
piecewise	GAT	–	50.03	41.37	82.38	84.45	23.12	20.40	58.33	57.09
constant	Fast Sub-AT	Fast AT	48.22	47.88	82.36	82.74	24.97	24.55	52.74	53.09
constant	GradAlign Sub-AT	GradAlign	48.88	48.40	79.82	80.84	**25.69**	**25.46**	52.65	52.92
constant	GAT Sub-AT	GAT	**51.15**	**50.80**	81.76	81.61	23.40	22.96	57.71	58.45

Table 2. Results of single-step Sub-AT with a larger training ϵ against ℓ_∞ perturbations of radius $8/255$ on CIFAR-10 ($\alpha = 1.25\epsilon$).

Method	Subspace	Best			Final			Time
		Natural	PGD-50	AA	Natural	PGD-50	AA	
Fast AT	–	73.13	39.95	37.55	89.93	0.00	0.00	1.6h
Fast Sub-AT ($\epsilon = 8/255$)	Fast AT	82.36	48.22	44.20	82.74	47.88	43.89	1.0h
Fast Sub-AT ($\epsilon = 12/255$)	Fast AT	80.74	50.38	45.84	80.91	49.64	45.40	1.0h
Fast Sub-AT ($\epsilon = 16/255$)	Fast AT	78.64	51.46	46.11	79.13	51.22	46.03	1.0h
Fast Sub-AT ($\epsilon = 12/255$)	GAT	80.77	52.41	46.80	80.72	52.30	46.80	2.1h
Fast Sub-AT ($\epsilon = 14/255$)	GAT	79.96	**53.35**	**47.25**	80.14	**53.02**	**46.92**	2.1h
PGD-10 AT	–	80.50	50.79	47.29	82.92	42.51	41.08	7.0h

epochs (as our default setting for base AT), which is commonly used and produces best robustness performance as suggested by Rice *et al*. [28]; ***iii***) *constant* schedule, which is adopted for our Sub-AT to show its insensitivity to learning rates. We set the learning rate as 1 *without* a schedule and train the model for 40 epochs in subspace with a sufficient convergence. We use a batch size of 128 and SGD optimizer with momentum 0.9 and weight decay 10^{-4}. Data augmentations, such as 4-pixel padding, random cropping, and horizontal flipping, are applied.

Evaluation. We consider both the best and final robustness performance during the training and use the difference between them to evaluate the degree of overfitting. The model checkpoint that achieves the best robust accuracy on the validation set is selected as the best model, while the final is an average of the last five epochs [28] (except that cyclic learning rates use the last epoch). The time consumption is measured on an Nvidia Geforce GTX 2080 TI. For Sub-AT, we repeat over five independent runs. The standard deviations in tables are omitted as they are very small ($\leq 0.45\%$), which hardly affects the results.

4.2. Resolving Catastrophic Overfitting

First, we consider resolving the catastrophic overfitting in single-step AT. We set the Fast AT [40], i.e., FGSM AT with a random initialization, as the baseline, and also consider other recently proposed methods for preventing the overfitting, including GradAlign (with a recommended coefficient $\lambda = 0.2$) [1], Stable Single-AT (with $c = 3$ check points) [17], Free AT ($m = 8$) [29] and GAT [32] (with a default coefficient $\lambda = 10$). We evaluate robustness using PGD-50 attack [1,40] and set the training epochs to 200 [17] to closely examine whether the overfitting will occur.

The results of different methods on both CIFAR-10 and CIFAR-100 datasets are presented in Tab. 1, where we apply Sub-AT to different base methods with the subspaces extracted from the corresponding training trajectory. The base methods use a piecewise schedule, where the overfitting mostly occurs. We observe that under piecewise learning rates of 200 epochs, Fast AT and FGSM AT with GradAlign still meet serious catastrophic overfitting. A carefully designed cyclic learning rate schedule [1, 40] helps overcome the overfitting, but it leads to an inferior robust performance (*e.g.*, $\mathbf{-1.86}\%$ on CIFAR-10) compared with GradAlign Sub-AT. Without any other regularization technique, our naive Fast Sub-AT is already able to obtain better robustness than FGSM AT with GradAlign regularization [1]. GAT [32] indeed overcomes the catastrophic overfitting and achieves impressive robustness among base methods, but it still potentially suffers from overfitting problems, as the dif-

ference between the best and final is large. Combined with Sub-AT, we are able to mitigate the overfitting and consistently improve the robustness.

To ensure the robustness improvements, we conduct additional evaluations via Auto-Attack. On CIFAR-10, our GAT Sub-AT achieves $46.33 \pm 0.37\%$ robust accuracy (best) while base GAT achieves $45.29 \pm 0.53\%$ (best), and on CIFAR-100, our GradAlign Sub-AT achieves $21.64 \pm 0.22\%$ (best) while cyclic GradAlign achieves $20.30 \pm 0.11\%$ (best). Thus via Sub-AT, we obtain state-of-the-art robustness performance on single-step AT, further reducing the robustness gap to multi-step AT. Note that our good performance does not rely on the results obtained during the DLDR sampling, as both vanilla Fast AT and GradAlign meet serious overfitting during the training and only obtain a poor result far from satisfactory.

4.3. Towards Larger Steps and Larger Radius

After resolving the catastrophic overfitting, we demonstrate that Sub-AT overcomes the overfittings in training with a large step and radius, further improving the robustness performance.

Single-step AT with a larger step. Although Wong *et al*. [40] discovered that adding a random initialization to FGSM AT could help avoid catastrophic overfitting, it only holds when the step size α is not too large. In fact, applying α larger than $12/255$ could still meet serious catastrophic overfitting (as illustrated in Fig. 3 of [40] for $\epsilon = 8/255$). Since Sub-AT resolves the overfitting, we expect that it can allow training with a large α. To this end, we consider CIFAR-10 and repeat Fast Sub-AT experiments in Sec. 4.2 with α ranging from $1/255$ to $16/255$ and record the final robustness performance (note that we are in the same subspace). From the results in Fig. 5, we observe that Sub-AT successfully resolves the overfitting for large step sizes and further improves the robustness, showing that using a large step indeed benefits robustness.

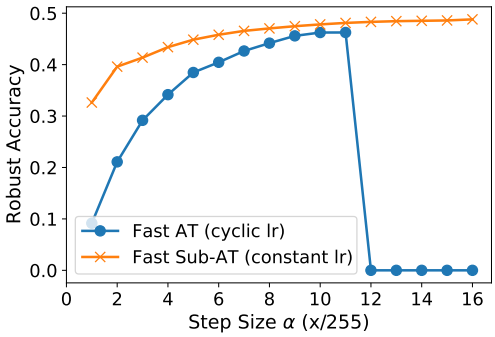


Figure 5. **Final** robust test accuracy of different single-step methods under PGD-20 attack over different step sizes for $\epsilon = 8/255$.

Single-step AT with a larger radius. AT with a larger radius generally could provide better robustness guarantees against potential adversarial attacks. However, it is hard to train a model that is robust to a large ϵ, especially for single-step AT [1]. By constraining the training in subspace, we can conduct single-step AT for a large ϵ with ease. Similar to last section, we repeat the Fast Sub-AT experiments in Sec. 4.2 with training radius $\epsilon \in [8/255, 20/255]$ (by default $\alpha = 1.25\epsilon$). We select the best checkpoints (against PGD-20 attack of radius $8/255$) of different settings and plot their robust accuracy curves with respect to different strengths of attack. In Fig. 6, we observe that within a certain range, training with a larger radius consistently improves robustness against attacks of different radii (especially for the large one), showing that the model's robustness is genuinely improved. However, there also exist limitations as expected, as excessive perturbations will harm the valuable information of training data, resulting in a degenerated performance. For example, the best robustness against $8/255$ attack is achieved with training radius $16/255$.

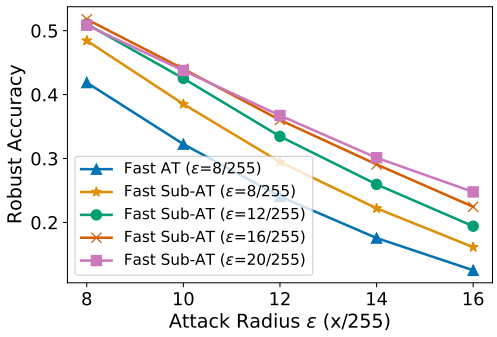


Figure 6. Test robust accuracy under PGD-20 attack with respect to different perturbation radius. We evaluate the checkpoint with the best performance on the validation set.

We summarize the results of training with larger steps and larger radius in Tab. 2, where we also report total time consumption (including DLDR phase for Sub-AT) as well as evaluations on PGD-50 and Auto-Attack. Main findings include: ***i***) simply by increasing the training step size and radius, Sub-AT significantly improves model robustness against both strong attacks, without suffering catastrophic overfitting nor additional time cost; ***ii***) notably, pure single-step-based Sub-AT achieves a very competitive robust accuracy under Auto-Attack than multi-step PGD-10 AT and even better robust accuracy under PGD-50 attack. We also get rid of the serious problem of robust overfitting, which PGD-10 AT suffers. ***iii***) more promisingly, Sub-AT only takes a seventh of total training time compared with PGD-10 AT, which is a considerable superiority.

Table 3. Results of multi-step AT and Sub-AT on CIFAR-10/100 against **PGD-20** attack with ℓ_2 and ℓ_∞ norm perturbations.

Dataset	Norm	Radius	Settings	Robust Accuracy			Natural Accuracy		
				Best	Final	Diff.	Best	Final	Diff.
CIFAR-10	ℓ_∞	$\epsilon = \frac{8}{255}$	AT	51.09	42.92	8.17	80.50	82.92	-2.42
			Sub-AT	**52.79**	**52.31**	**0.48**	80.46	80.47	-0.01
	ℓ_2	$\epsilon = \frac{128}{255}$	AT	67.73	65.21	2.52	88.17	88.82	-2.42
			Sub-AT	**69.14**	**69.01**	**0.13**	88.87	88.84	-0.01
CIFAR-100	ℓ_∞	$\epsilon = \frac{8}{255}$	AT	26.80	19.38	7.42	52.29	53.27	-0.98
			Sub-AT	**27.50**	**27.02**	**0.48**	52.41	52.18	0.23
	ℓ_2	$\epsilon = \frac{128}{255}$	AT	40.21	34.98	5.23	61.98	60.28	1.70
			Sub-AT	**41.48**	**41.00**	**0.48**	62.62	63.13	-0.51

Table 4. Results of applying Fast AT with a larger training radius ϵ to multi-step AT (ℓ_∞ attack of radius 8/255, CIFAR-10).

Method	Subspace	Best			Final			Time
		Natural	PGD-50	AA	Natural	PGD-50	AA	
PGD-10 AT	–	80.50	50.79	47.29	82.92	42.51	41.08	7.0h
PGD-10 Sub-AT	PGD-10 AT	80.46	52.48	48.37	80.47	52.01	47.88	4.9h
Fast Sub-AT ($\epsilon = 12/255$)	PGD-10 AT	81.02	53.32	48.58	81.25	52.85	48.21	3.9h
Fast Sub-AT ($\epsilon = 16/255$)	PGD-10 AT	79.63	**54.17**	**49.14**	79.94	**54.11**	**48.86**	3.9h

4.4. Extensions to Multi-step AT

We then apply Sub-AT to mitigate robust overfitting in multi-step AT, where we set standard PGD-10 AT [23] as the baseline. To numerically show the degree of overfitting, we report the difference between the best and final robust accuracy. The difference is crucial as, generally, we can only attain the final performance without using validation set. From the results in Tab. 3, we observe that robust overfitting occurs in every setting of AT as expected, and the gap between best and final can be as large as 8.17% (on CIFAR-10 with ℓ_∞ norm). By restricting AT in subspace, we reduce the gap to less than 0.5% meanwhile significantly improving the robust accuracy (*e.g.*, $+\mathbf{1.7}\%$ on CIFAR-10). In the good subspace extracted from DLDR [20], generalization on clean data is also kept compared with standard AT. More examinations via Auto-Attack are presented in Appendix B, where we observe that Sub-AT can indeed mitigate robust overfitting and consistently improve the robustness, rather than as a result of gradient masking [2, 26]. We note that the improvements in robustness naturally come from the low-dimensional optimization, and the results are promising to be further improved by combining other enhancement techniques, such as modifications on loss function [3].

Finally, we apply Fast Sub-AT to the subspace extracted from multi-step PGD-10 AT. From the results in Tab. 4, we observe that, surprisingly, Fast Sub-AT with a larger training radius remarkably outperforms the PGD-10 Sub-AT. It implies that we can achieve strong robustness only with the guidance of weak adversarial examples in the subspace, which also brings computational benefits. This encouraging finding reveals the previously underestimated potential of single-step AT and can provide a new scheme for designing more robust and efficient AT methods.

5. Conclusion

In this paper, we focus on the serious catastrophic overfitting in single-step AT. From a novel perspective of optimization, we reveal the close link between the fast-growing gradient of each sample and overfitting, which can also explain the robust overfitting in multi-step AT. To control the growth of gradient, Sub-AT is proposed to constrain AT in a carefully extracted subspace. It successfully resolves both kinds of overfitting and hence significantly improves the robustness. Leveraging the subspace, we allow single-step AT with larger steps and radius, further improving the robustness. Weak adversarial examples generated from single-step AT can be trained to obtain even better robustness than those from multi-step PGD AT in subspace, revealing the great potential of single-step AT. As a result, pure single-step-based AT achieves comparable robustness to standard PGD-10 AT with only one-seventh of the training time, which is a solid step towards more efficient AT methods.

Acknowledgements

We are very grateful for Hongkai Zheng, Qinghua Tao and anonymous reviewers for the useful feedback on the paper. The research leading to these results has received funding from National Key Research and Development Project (No.2018AAA0100702), National Natural Science Foundation of China 61977046, and Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102). X. Huang is the corresponding author.

References

[1] Maksym Andriushchenko and Nicolas Flammarion. Understanding and improving fast adversarial training. In *Proceedings of the Advances In Neural Information Processing Systems 33 (NeurIPS)*, volume 33, pages 16048–16059, 2020. 1, 2, 3, 5, 6, 7

[2] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning (ICML)*, pages 274–283. PMLR, 2018. 8, 11

[3] Tianlong Chen, Zhenyu Zhang, Sijia Liu, Shiyu Chang, and Zhangyang Wang. Robust overfitting may be mitigated by properly learned smoothening. In *International Conference on Learning Representations (ICLR)*, 2020. 5, 8, 11

[4] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning (ICML)*, pages 1310–1320. PMLR, 2019. 2

[5] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track (Round 2)*, 2020. 2

[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 (ICML)*, pages 2206–2216. PMLR, 2020. 5, 11

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009. 5, 11

[8] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9185–9193, 2018. 2

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

[10] Frithjof Gressmann, Zach Eaton-Rosen, and Carlo Luschi. Improving neural network training in low dimensional random bases. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 12140–12150, 2020. 3

[11] Guy Gur-Ari, Daniel A Roberts, and Ethan Dyer. Gradient descent happens in a tiny subspace. *arXiv preprint arXiv:1812.04754*, 2018. 3

[12] Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning (ICML)*, pages 1225–1234. PMLR, 2016. 2

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 5, 11

[14] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning (ICML)*, pages 448–456. PMLR, 2015. 4

[15] Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *International Conference on Learning Representations (ICLR)*, 2020. 2

[16] Peilin Kang and Seyed-Mohsen Moosavi-Dezfooli. Understanding catastrophic overfitting in adversarial training. *arXiv preprint arXiv:2105.02942*, 2021. 1

[17] Hoki Kim, Woojin Lee, and Jaewook Lee. Understanding catastrophic overfitting in single-step adversarial training. In *AAAI*, pages 8119–8127, 2020. 1, 2, 3, 6

[18] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Technical Report*, 2009. 5

[19] Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. In *International Conference on Learning Representations (ICLR)*, 2018. 3

[20] Tao Li, Lei Tan, Qinghua Tao, Yipeng Liu, and Xiaolin Huang. Low dimensional landscape hypothesis is true: DNNs can be trained in tiny subspaces. *arXiv preprint arXiv:2103.11154*, 2021. 2, 3, 4, 8, 11

[21] Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, and Jun Zhu. Defense against adversarial attacks using high-level representation guided denoiser. In *the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[22] Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, and Jun Zhu. Defense against adversarial attacks using high-level representation guided denoiser. In *the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1778–1787, 2018. 2

[23] 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. 1, 2, 3, 8

[24] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Jonathan Uesato, and Pascal Frossard. Robustness via curvature regularization, and vice versa. In *the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9078–9086, 2019. 2

[25] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5947–5956, 2017. 2

[26] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, pages 506–519, 2017. 8, 11

[27] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *the IEEE Symposium on Security and Privacy (SP)*, pages 582–597, 2016. 2

[28] Leslie Rice, Eric Wong, and Zico Kolter. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning (ICML)*, pages 8093–8104. PMLR, 2020. 2, 3, 5, 6

[29] 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. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pages 3353–3364, 2019. 3, 6

[30] Leslie N Smith. Cyclical learning rates for training neural networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 464–472. IEEE, 2017. 1

[31] Yang Song, Taesup Kim, Sebastian Nowozin, Stefano Ermon, and Nate Kushman. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2018. 2

[32] Gaurang Sriramanan, Sravanti Addepalli, Arya Baburaj, et al. Guided adversarial attack for evaluating and enhancing adversarial defenses. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 20297–20308, 2020. 1, 3, 6

[33] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations (ICLR)*, 2018. 2

[34] Mark Tuddenham, Adam Prügel-Bennett, and Jonathan Hare. Quasi-newton's method in the class gradient defined high-curvature subspace. *arXiv preprint arXiv:2012.01938*, 2020. 3

[35] Oriol Vinyals and Daniel Povey. Krylov subspace descent for deep learning. In *Artificial Intelligence and Statistics (AISTATS)*, pages 1261–1268. PMLR, 2012. 3

[36] BS Vivek and R Venkatesh Babu. Single-step adversarial training with dropout scheduling. In *the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 947–956, 2020. 3

[37] BS Vivek, Arya Baburaj, and R Venkatesh Babu. Regularizer to mitigate gradient masking effect during single-step adversarial training. In *CVPR Workshops*, pages 66–73, 2019. 1

[38] Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *International Conference on Learning Representations (ICLR)*, 2019. 2, 3

[39] Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning (ICML)*, pages 5286–5295. PMLR, 2018. 2

[40] Eric Wong, Leslie Rice, and J. Zico Kolter. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations (ICLR)*, 2020. 1, 3, 4, 5, 6, 7

[41] Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 2958–2969, 2020. 2

[42] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *British Machine Vision Conference (BMVC)*, 2016. 5, 12

[43] Dinghuai Zhang, Tianyuan Zhang, Yiping Lu, Zhanxing Zhu, and Bin Dong. You only propagate once: Painless adversarial training using maximal principle. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 227–238, 2019. 2

[44] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning (ICML)*, pages 7472–7482. PMLR, 2019. 2, 3

[45] Jingfeng Zhang, Xilie Xu, Bo Han, Gang Niu, Lizhen Cui, Masashi Sugiyama, and Mohan Kankanhalli. Attacks which do not kill training make adversarial learning stronger. In *International Conference on Machine Learning (ICML)*, pages 11278–11287. PMLR, 2020. 2