Taxonomy Driven Fast Adversarial Training

Kun Tong[1], **Chengze Jiang**[1], **Jie Gui**[*1,2,3], **Yuan Cao**[4]
[1] Southeast University, Nanjing, China
[2] Engineering Research Center of Blockchain Application, Supervision And Management (Southeast University), Ministry of Education, China
[3] Purple Mountain Laboratories, China
[4] Ocean University of China, China
tongkun@seu.edu.cn, czjiang@ieee.org, guijie@seu.edu.cn, cy8661@ouc.edu.cn

Abstract

Adversarial training (AT) is an effective defense method against gradient-based attacks to enhance the robustness of neural networks. Among them, single-step AT has emerged as a hotspot topic due to its simplicity and efficiency, requiring only one gradient propagation in generating adversarial examples. Nonetheless, the problem of catastrophic overfitting (CO) that causes training collapse remains poorly understood, and there exists a gap between the robust accuracy achieved through single- and multi-step AT. In this paper, we present a surprising finding that the taxonomy of adversarial examples reveals the truth of CO. Based on this conclusion, we propose taxonomy driven fast adversarial training (TDAT) which jointly optimizes learning objective, loss function, and initialization method, thereby can be regarded as a new paradigm of single-step AT. Compared with other fast AT methods, TDAT can boost the robustness of neural networks, alleviate the influence of misclassified examples, and prevent CO during the training process while requiring almost no additional computational and memory resources. Our method achieves robust accuracy improvement of 1.59%, 1.62%, 0.71%, and 1.26% on CIFAR-10, CIFAR-100, Tiny ImageNet, and ImageNet-100 datasets, when against projected gradient descent PGD10 attack with perturbation budget 8/255. Furthermore, our proposed method also achieves state-of-the-art robust accuracy against other attacks. Code is available at https://github.com/bookman233/TDAT.

Introduction

Deep neural networks have made significant progress in many fields and achieved remarkable performance. However, they are vulnerable to adversarial examples (AEs) with carefully designed perturbations (Ding et al. 2023; Li et al. 2023; He et al. 2023; Chen et al. 2023). Adversarial training (AT) (Mao et al. 2022; Tsiligkaridis and Roberts 2022) has emerged as the most effective approach to enhance the robustness of neural networks against the threat from worst-case perturbations. Furthermore, AT methods are classified into two categories based on the method used to generate AEs for training: multi-step and single-step AT (Gao et al. 2022; Zhang et al. 2019a; Ding et al. 2022). While multi-step AT is found to be more advantageous in improving the robustness of neural networks, it requires more time to compute gradients of the neural network's input multiple times (Park and Lee 2021; Yu and Sun 2022; Izmailov et al. 2018; Zhang et al. 2020). As a result, single-step AT methods gain significant attention as a research hotspot due to their effectiveness and efficiency (Phan et al. 2023; Chiang, Chan, and Wu 2021; Qin et al. 2023).

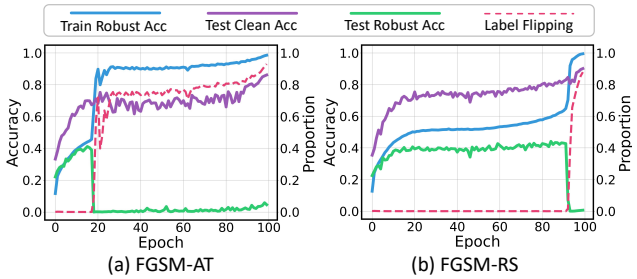


Figure 1: Catastrophic Overfitting (CO) and Label Flipping. The red line indicates the proportion of Case 4 (as shown in Fig. 3) to all misclassified examples, which infers the proportion of label flipping examples explodes once CO occurs. The robust accuracy is evaluated against PGD-10 attack.

Despite the single-step AT achieving impressive performance on both efficiency and improving the robustness of neural networks, it still suffers from a serious and puzzling problem: catastrophic overfitting (CO) (Madry et al. 2018). This problem leads trained models vulnerable to multi-step attacks on test dataset despite remaining robust to single-step attacks on training dataset after a few batches, which is shown in Fig.1. It has been reported that CO is due to the sharp decrease in the generalization ability of the neural network during AT (Zhang et al. 2022). Recently, there are new perspectives and solution schemes presented to prevent CO, containing prior-guided initialization method (Jia et al. 2022a), subspace adversarial training (Li et al. 2022), noise fast adversarial training (de Jorge Aranda et al. 2022), etc.

Although considerable efforts have been made to explain the CO problem and enhance the robustness of neural networks (Chen and Ji 2022; Athalye, Carlini, and Wagner 2018; Mei et al. 2023; Pang et al. 2022; Rice, Wong, and Kolter 2020; Chen et al. 2021), its underlying mechanisms remain inadequately identified. From the perspective of fail-

*Corresponding author.

ure phenomenon of standard single-step AT, the robust accuracy against fast gradient sign method (FGSM) (Goodfellow, Shlens, and Szegedy 2015) on the training dataset is excessively high and the robust accuracy against projected gradient descent (PGD) (Madry et al. 2018) on the test dataset almost approaches zero. Consequently, CO is attributed to the overfitting of the neural network to AEs during AT, thereby causing the collapse of generalization to AEs from the test dataset. However, the following two questions are still unidentified.

- Why does robust accuracy fluctuate drastically within such a small number of batches?
- Why does the CO problem have almost no effect on the clean accuracy of the test dataset?

These two problems prompt us to find a new perspective to explain and resolve the CO in single-step AT.

To this end, this paper presents a taxonomy of the AEs and analyzes the impact of different categories of training examples during single-step AT, which shows that the changing pattern of accuracy in CO is directly mapped to changes in the number of specific examples. Moreover, the label flipping phenomenon (as presented in the red line of Fig.1) is revealed and investigated, which leads to that the adversarial noise generated by FGSM contradicts the original goal of maximizing loss. On this basis, we propose **T**axonomy **D**riven fast **A**dversarial **T**raining (TDAT) which incorporates dynamic label relaxation, batch momentum initialization, and taxonomy driven loss function. Finally, comprehensive experiments demonstrate that our method helps neural networks achieve better robust and clean accuracy than other state-of-the-art defense methods. The main contributions are summarized as follows.

- To understand and explain the problem of CO, we present a taxonomy to investigate the impact of different AEs to single-step AT and identify the changes in the quantity of which examples lead to the occurrence of CO.
- Based on the above analysis, we propose the TDAT which systematically improves AT across initialization, label relaxation, and loss function.
- The systemic experiments are designed and performed on standard datasets to evaluate our TDAT with other state-of-the-art AT methods. Results show that the TDAT achieves better robust accuracy than state-of-the-art single-step and even multi-step AT methods.

Related Work

Adversarial Training

AT is a technique that aims to improve the robustness of neural networks against adversarial attacks by introducing adversarial perturbations during the training process. Formally, let $D = \{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^n$ be a dataset consisting of n examples and L classes with $\boldsymbol{x}_i \in \mathbb{R}^d$ as the benign example in the d-dimensional space and $\boldsymbol{y}_i$ is the corresponding one-hot label. The objective of AT is to train a neural network $f(\cdot)$ with parameters ϕ to be robust against adversarial attacks. Mathematically, AT is formulated as a min-max problem:

$$\min_{\phi} \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{D}}\Big[\max_{\boldsymbol{\delta}\in\mathcal{N}(\boldsymbol{x},\epsilon)} \mathcal{L}(f(\boldsymbol{x}+\boldsymbol{\delta},\phi),\boldsymbol{y})\Big], \quad (1)$$

where $\mathcal{D}$ represents the data distribution, $\mathcal{N}(\cdot)$ is the norm constraint with maximum perturbation budget ϵ, $\boldsymbol{\delta}$ denotes the adversarial perturbation, and $\mathcal{L}(\cdot)$ denotes the loss function. Specifically, the objective of the inner maximization is generating the worst adversarial perturbations for the neural network, while the outer minimization updates the model to improve robustness (Bai et al. 2021; Pang et al. 2021; Li and Liu 2023; Kim, Lee, and Lee 2021). There are two categories of AT based on the number that adversarial perturbations are optimized: single-step and multi-step AT.

Single-step AT The FGSM is one of the most famous methods for generating AEs and is used in the single-step AT. The principle of the FGSM is formulated as

$$\boldsymbol{x}_{\text{adv}} = \boldsymbol{x} + \epsilon \cdot \text{sign}\Big(\nabla_{\boldsymbol{x}}\mathcal{L}(f(\boldsymbol{x},\phi),\boldsymbol{y})\Big), \quad (2)$$

where $\boldsymbol{x}_{\text{adv}}$ is adversarial example. Since the FGSM only requires one-step gradient propagation for generating AEs, the associated AT is classified as the single-step version. However, this scheme is vulnerable to CO when faced with multi-step adversarial attacks, as reported in (Wong, Rice, and Kolter 2020; Müller, Kornblith, and Hinton 2019; Sriramanan, Gor, and Feizi 2022).

Multi-step AT On the other aspect, Madry $et\ al.$ utilize the multi-step adversarial attack to achieve the inner maximization problem of AT (as defined in equation (1)), which is implemented as the PGD method (Madry et al. 2018):

$$\boldsymbol{x}_{\text{adv}}^{t+1} = \Pi_{\epsilon}\Big(\boldsymbol{x}_{\text{adv}}^{t} + \alpha \cdot \text{sign}\big(\nabla_{\boldsymbol{x}_{\text{adv}}^{t}}\mathcal{L}(f(\boldsymbol{x}_{\text{adv}}^{t},\phi),\boldsymbol{y})\big)\Big), \quad (3)$$

where t and α represent the t-th iteration step and step size, respectively. The projection operation $\Pi_{\epsilon}(\cdot)$ is utilized to ensure that the perturbation is within a certain bound. It has been reported that multi-step AT (Cui et al. 2021; Zhang et al. 2019b) can help neural networks achieve stronger robustness than single-step AT, due to its better attack strength. Nonetheless, the implementation process of multi-step AT requires more computational and time resources, as it involves performing multiple back and forward propagations.

Catastrophic Overfitting

AT suffers from CO, leading to training instability and even failure in the training process. In (Wong, Rice, and Kolter 2020), Wong $et\ al.$ identify the mode that causes CO in single-step AT and present FGSM-RS to prevent CO. After that, Andriushchenko $et\ al.$ present a new regularization method, GradAlign, that improves the quality of AEs generation to prevent CO via maximizing the gradient alignment (Andriushchenko and Flammarion 2020). As a result, the problem of CO attracts much attention, and solutions are presented (Miyato et al. 2018; Jia et al. 2022a). One of the effective approaches is to develop regularization methods (Sankaranarayanan et al. 2018; Herrmann et al. 2022). For instance, Sriramanan $et\ al.$ present a relaxation method to

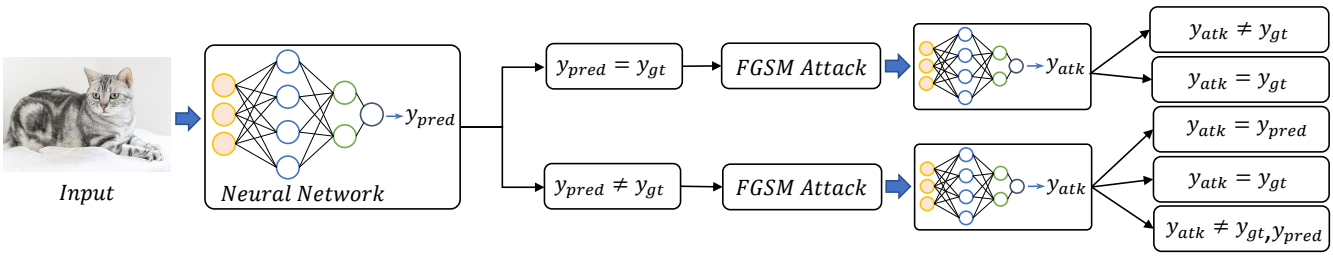


Figure 2: Taxonomy of adversarial examples during single-step AT. The five cases in this figure correspond to Case 1 to Case 5, from top to bottom. y_{gt} represents ground truth.

standard loss, that finds more appropriate gradient directions to increase attack efficiency (Sriramanan et al. 2020). NuAT (Sriramanan et al. 2021) adopts a nuclear norm regularizer to enforce function smoothing in the vicinity of data examples. Nonetheless, these methods require additional computational resources for worse AEs, limiting their expandability. In addition, it has been reported that the initialization method of AEs is significant (Wong, Rice, and Kolter 2020; Tramèr et al. 2018). Jia *et al.* explore the difference between the training process of vanilla AT and fast AT to explain the reason for CO and propose a prior-guided FGSM initialization method to avoid CO (Jia et al. 2022a).

Problem Description and Analysis

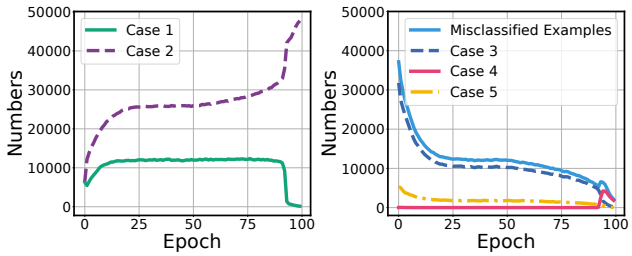
In this section, a taxonomy is defined that divides all AEs used for AT into five categories. The connection between CO and the proposed taxonomy is then analyzed in detail.

Taxonomy and Catastrophic Overfitting

CO problem that occurs during single-step AT leads to the robust accuracy of the neural network against PGD rapidly decreasing and training accuracy rapidly increasing within a few epochs. Currently, CO is explained as the overfitting of the neural network on the AEs generated by FGSM when performing training. However, this explanation fails to identify the reason for the drastic degradation of robust accuracy and the negligible impact of clean accuracy on test dataset. Therefore, these problems motivate us to explore a new perspective to explain CO in single-step AT.

Subsequently, we divide examples into five categories from the perspective of whether the clean examples are correctly classified by the neural network before being attacked and whether these examples are successfully attacked. On this basis, we observe the quantity change of each category examples when performing single-step AT. First, the five cases are defined as

- Case 1: The clean example is correctly classified by the network, and the corresponding adversarial example is misclassified by the network.
- Case 2: The clean example is correctly classified by the network, and the corresponding adversarial example is also correctly classified by the network.
- Case 3: The i-class clean example is misclassified as j-class example by the network, and the corresponding adversarial example is classified by the network as j-class example.
- Case 4: The i-class clean example is misclassified as j-class example by the network, and the corresponding adversarial example is classified by the network as i-class example.
- Case 5: The i-class clean example is misclassified as j-class example by the network, and the corresponding adversarial example is classified by the network as k-class example where $k \neq i, j$.



(a) (b)

Figure 3: Numbers of the five cases for each epoch during single-step AT. For each epoch, the sum of the number of five cases equals the size of the training dataset.

Fig. 2 is provided to better illustrate the taxonomy of examples in the single-step AT procedure. After that, we perform the single-step AT using ResNet-18 on the CIFAR-10 dataset against the FGSM attack with perturbation budget of 8/255, step size of 8/255, and random initialization (Wong, Rice, and Kolter 2020). The quantity change of the five cases with epoch is presented in Fig. 3. Specifically, the green and purple lines in Fig. 3(a) represent the number of examples in cases 1 and 2, respectively. The dark blue, red, and orange lines in Fig. 3(b) denote the number of examples in cases 3, 4, and 5, respectively.

Concretely, as observed from Fig. 3(a), when CO occurs in the 91th epoch, FGSM possesses an extremely low attack success rate (as indicated by the green line), thereby the number of AEs is too small to maintain single-step AT. Meanwhile, according to Fig. 3(b) and Fig.1(b), we observe a surprising phenomenon of **label flipping**, where most misclassified clean examples are attacked to ground truth. In this situation, the inner maximization problem as presented in

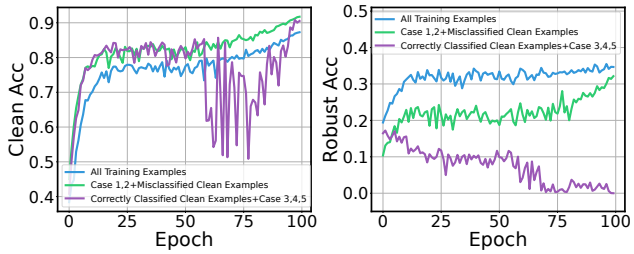


Figure 4: Clean (left) and robust (right) accuracy on test dataset when performing FGSM-RS.

(1) is ineffective. AEs are generated to satisfy the objective of the outer minimization problem instead of its initial objective, leading to the sudden collapse of single-step AT and counterintuitive phenomena. Specifically, a substantial number of unsuccessfully attacked examples (Case 2) and almost offenseless AEs (Case 4) ensure that the model continues to only learn the distribution around clean examples, thereby leading to minor impacts on the clean accuracy of test dataset and degradation of robustness. Under the min-max optimization paradigm of AT (1), we recognize an inherent imbalance between inner and outer optimization, whereby the intensity of minimization optimization surpasses that of maximization optimization. Consequently, it is significant to strengthen the maximization problem by exploiting the diversity of AEs and improve the minimization problem to better learn AEs through label adjustment. We then introduce dynamic label relaxation and batch momentum initialization, which respectively enhance the minimization and maximization optimization.

Misclassified Examples and Training Instability

In order to further investigate the impact of different kinds of AEs on training stability, the correctly classified and misclassified clean examples are respectively used to generate AEs for performing single-step AT. To this end, the ResNet-18 is trained by the FGSM-RS (Wong, Rice, and Kolter 2020) method with perturbation budget 8/255 and step size 4/255 on CIFAR-10 dataset. The clean and robust accuracy of the neural network is presented in Fig. 4. The blue line represents the accuracy of standard single-step AT, while the green line denotes the accuracy of neural networks trained with AEs from cases 1, 2, and misclassified clean examples. The purple line represents the accuracy of neural networks trained with clean examples correctly classified by the neural network and AEs from cases 3, 4, and 5. Compared with green and blue line in Fig. 4, the violent oscillations of the purple line indicate AEs from misclassified examples cause the instability of single-step AT. Meanwhile, we realize the decrease in the number of AEs impairs the robust accuracy. Additionally, as the solid blue line in Fig. 3, the number of clean examples misclassified by the neural network occupies a considerable proportion of total training examples. Therefore, it is valuable to completely utilize misclassified examples rather than not using them to generate AEs, which motivates us to introduce the taxonomy driven loss.

Methodology

In this section, we present the TDAT from the perspective of initialization method for AEs, dynamic label relaxation, and loss function. The implementation of our proposed method is outlined in Algorithm 1 and presented in detail as follows.

Batch Momentum Initialization

Considerable works have demonstrated the significance of example initialization methods in single-step AT. Appropriate initialization methods effectively mitigate the problem of CO, where the radius and distribution of initialization methods respectively determine the strength and diversity of AEs. Recently, it has been reported that the approach for addressing CO by exploiting prior perturbation information improves initialization and enhances robust accuracy (Jia et al. 2022a). Nonetheless, this method requires extensive memory resources to load the entire dataset, making it infeasible for larger scale datasets.

In light of the diversity of AEs, it is critical when achieving the maximization of the single-step AT as defined in (1). Generative adversarial perturbations can well satisfy the diversified distribution of AEs and reduce the memory requirement simultaneously. Following this, we leverage batch historical perturbation information and momentum method to further improve AEs diversity. The adversarial perturbation is defined as

$$\boldsymbol{\delta} = \arg\max_{\boldsymbol{\delta}\in\mathcal{N}(\boldsymbol{x},\epsilon)} \mathcal{L}(f(\boldsymbol{x}+\boldsymbol{\delta}_{\text{m-1}}+\boldsymbol{\delta},\phi),\boldsymbol{y}), \tag{4}$$

$$\boldsymbol{\delta}_{\text{m}} = \alpha\cdot\boldsymbol{\delta}_{\text{m-1}} + (1-\alpha)\cdot\boldsymbol{\delta}, \tag{5}$$

where $\boldsymbol{\delta}$ denotes adversarial perturbations generated by clean examples initialized with momentum perturbation $\boldsymbol{\delta}_{\text{m-1}}$ from last batch. The momentum perturbation $\boldsymbol{\delta}_{\text{m}}$ is updated with factor α and employed as the initialization perturbation for next batch. This initialization approach stabilizes AT and prevents CO via enhancing diversity and improving the directional changes of adversarial perturbations.

Dynamic Label Relaxation

Owing to the inherent unbalance under AT and inner maximization performance limited by the attack step of FGSM, we propose a dynamic label relaxation method to find a better gradient updating scheme for the neural network by lowering the classification expectation against the AEs. The proposed dynamic label relaxation method is formulated as

$$\hat{\boldsymbol{y}} = \boldsymbol{y}\cdot\gamma + (\boldsymbol{y}-\mathbf{1})\cdot\frac{\gamma-1}{L-1}, \tag{6}$$

where $\hat{\boldsymbol{y}}$ denotes the relaxation label, $\boldsymbol{y}$ signifies one-hot label, L represents the number of classes for the training, and γ represents the label relaxation factor, which is formulated as

$$\gamma = \begin{cases} \beta\cdot\tanh(1-\frac{epoch}{EPOCHs}), & \text{if } \beta\cdot\tanh(1-\frac{epoch}{EPOCHs}) \geq \gamma_{\min}, \\ \gamma_{\min}, & \text{if } \beta\cdot\tanh(1-\frac{epoch}{EPOCHs}) < \gamma_{\min}, \end{cases} \tag{7}$$

where $epoch$ denotes the current epoch number, $EPOCHs$ denotes the total number of epochs for training, parameter β

Algorithm 1: Taxonomy Driven Fast Adversarial Training

Parameters: clean example $\boldsymbol{x}$ and one-hot label $\boldsymbol{y}$ with length L; learning rate μ; number of total epochs $EPOCHs$; previous momentum perturbation from last batch $\boldsymbol{\delta}_{\text{m-1}}$; current momentum perturbation $\boldsymbol{\delta}_{\text{m}}$ for next batch; relaxation factor γ; dynamic relaxation label $\hat{\boldsymbol{y}}$; momentum factor α; scale factor β;
Return: parameter ϕ of the neural network $f(\cdot)$;

1: Initialize momentum perturbation $\boldsymbol{\delta}_0$ with uniform distribution $(-\epsilon, \epsilon)$.
2: **for** $epoch$ in $EPOCHs$ **do**
3: $\quad\gamma \leftarrow \beta \cdot \tanh(1 - \frac{epoch}{EPOCHs})$;
4: $\quad$**if** $\gamma < \gamma_{\text{min}}$ **then**
5: $\quad\quad\gamma \leftarrow \gamma_{\text{min}}$;
6: $\quad$**end if**
7: $\quad m \leftarrow 1$;
8: $\quad$**for** $batch$ in $BATCHes$ **do**
9: $\quad\quad p \leftarrow \text{Softmax}(f(\boldsymbol{x}))\,[i]$, where $\boldsymbol{y}\,[i] = 1$;
10: $\quad\quad\hat{\boldsymbol{y}} \leftarrow \boldsymbol{y} \cdot \gamma + (\boldsymbol{y} - \boldsymbol{1}) \cdot \frac{\gamma - 1}{L - 1}$;
11: $\quad\quad\boldsymbol{g}_x \leftarrow \text{sign}\Big(\nabla_{\boldsymbol{x}} \mathcal{L}(f(\boldsymbol{x} + \boldsymbol{\delta}_{\text{m-1}}, \phi), \hat{\boldsymbol{y}})\Big)$;
12: $\quad\quad\boldsymbol{\delta} \leftarrow \text{Clip}\Big(\boldsymbol{\delta}_{\text{m-1}} + \epsilon \cdot \boldsymbol{g}_x, -\epsilon, \epsilon\Big)$;
13: $\quad\quad\mathcal{L} \leftarrow \mathcal{L}_{\text{CE}} + \lambda \cdot \|f(\boldsymbol{x} + \boldsymbol{\delta}) - f(\boldsymbol{x})\|_2 \cdot \tanh(1 - p)$;
14: $\quad\quad g_\phi \leftarrow \nabla_\phi \mathcal{L}$;
15: $\quad\quad\phi \leftarrow \phi - \mu g_\phi$;
16: $\quad\quad\boldsymbol{\delta}_{\text{m}} \leftarrow \alpha \cdot \boldsymbol{\delta}_{\text{m-1}} + (1 - \alpha) \cdot \boldsymbol{\delta}$;
17: $\quad\quad m \leftarrow m + 1$.
18: $\quad$**end for**
19: **end for**
20: Return parameter ϕ.

is utilized to control the relaxation extent of label, parameter γ is dynamically changed with AT process and its minimum value γ_{min} ensures the relaxed label correctly guide the neural network to update. Hence, the value of γ_{min} can decrease with increasing the class number of the training dataset for lower expectations and boosting robustness. The function $\tanh(\cdot)$ is adopted to accelerate the decrease of γ.

Specifically, the parameter of our dynamic label relaxation method decreases as training progresses and converges to γ_{min}. In the early stage of training, dynamic label relaxation helps the neural network obtain better clean accuracy, which is essentially achieved by reducing the impact of noisy examples on the neural network. As the training process proceeds, the parameters of dynamic label relaxation become smaller (label smoothness increases), which prompts the neural network to correctly classify the adversarial examples without pursuing high confidence.

Taxonomy Driven Loss

Motivated by analysis of misclassified examples and training instability, we develop regularization terms with relaxation labels to mitigate the training instability caused by misclassified examples and prevent CO. First, we review the mathematical representation of the standard cross entropy loss $\mathcal{L}_{\text{CE}} = -\sum_{i=1}^{L} \hat{\boldsymbol{y}}_i \log f(\boldsymbol{x} + \boldsymbol{\delta})$, where the parameter $\hat{y}_i$ denotes the ith elements of relaxed label $\hat{\boldsymbol{y}}$. After that, the taxonomy driven loss is established to relieve the influence of AEs generated by the misclassified clean examples as

$$\mathcal{L}_{\text{TD}} = \mathcal{L}_{\text{CE}} + \lambda \cdot \|f(\boldsymbol{x} + \boldsymbol{\delta}) - f(\boldsymbol{x})\|_2 \cdot \tanh(1 - p), \quad (8)$$

where $p \in [0, 1]$ (line 9 in Algorithm 1) denotes the desired class confidence of example $\boldsymbol{x}$, with $\mathcal{L}_{\text{TD}}$ degrades to standard cross entropy loss for $p = 1$. λ is a trade-off hyperparameter to balance value of regularization term. Introducing the faster descending term $\tanh(1 - p)$ is aimed at separating and penalizing examples with lower value of p more severely. This loss focuses more on examples that may be misclassified to reduce the negative influence of misclassified examples to single-step AT and enhance stability.

Experiments and Analysis

Experimental Setup

Datasets and Training Analysis and comparison experiments are performed and evaluated on the **CIFAR-10** (Krizhevsky, Hinton et al. 2009), **CIFAR-100** (Krizhevsky, Hinton et al. 2009), **Tiny ImageNet** (Le and Yang 2015), and **ImageNet-100** (Deng et al. 2009) datasets, which are standard datasets for AT. We adopt ResNet-18 as the backbone to perform all experiments. Furthermore, we evaluate the best and last epoch of different methods to validate training stability. The Last and Best are obtained from two different model checkpoints during training stage. Specifically, the Best denotes that the model achieves the best robustness against PGD attack and the Last denotes the model's robustness against PGD attack in the last epoch. Bold numbers indicate the best results of single-step AT. For hyperparameters of CIFAR-10, CIFAR-100, Tiny ImageNet, and ImageNet-100, relaxation factor γ_{min} is set to 0.15, 0.05, 0.025, and 0.05, respectively. They are dependent on the number of classes in the corresponding dataset. Momentum factor is set to 0.75 for all experiments. More training details are shown in the **Supplementary Materials**.

Attacks for Evaluation Representative and state-of-the-art adversarial attack methods are considered to evaluate the performance of the TDAT method, $i.e.$, FGSM (Goodfellow, Shlens, and Szegedy 2015), MIFGSM (Dong et al. 2018), PGD with 10-step, 20-step, and 50-step versions (Madry et al. 2018), BIM (Kurakin, Goodfellow, and Bengio 2017), AutoAttack (Croce and Hein 2020), C&W (Carlini and Wagner 2017), APGD (Croce and Hein 2020), and Square (Andriushchenko et al. 2020) attacks. For each attack method, the perturbation budget is constrained by ℓ_∞ norm with budget $\epsilon = 8/255$ to evaluate the robust accuracy of the neural network and compare it with other existing AT methods. For the single-step attack methods, the step size is set to $8/255$, while it is set to $2/255$ against the multi-step attack methods.

Baselines The systemic experiments are provided for performing comparisons among our TDAT with other typical or state-of-the-art single-step AT methods, covering FGSM-RS (Wong, Rice, and Kolter 2020), Free (Shafahi et al. 2019),

Methods	Steps		Clean Acc	FGSM	MIFGSM	BIM	PGD 10	PGD 20	PGD 50	AA	C&W	APGD	Square
FGSM-RS	1	Best	51.67	31.02	23.05	22.42	22.61	22.04	21.75	18.72	20.92	21.87	23.00
		Last	51.67	31.02	23.05	22.42	22.61	22.04	21.75	18.72	20.92	21.87	23.00
Free(m=8)	-	Best	52.06	32.13	25.06	24.48	24.74	24.09	24.04	20.23	22.43	23.99	24.86
		Last	52.06	32.13	25.06	24.48	24.74	24.09	24.04	20.23	22.43	23.99	24.86
GAT	1	Best	57.49	36.77	29.63	28.91	29.14	28.60	28.30	23.11	25.14	28.42	27.69
		Last	57.58	36.85	29.55	28.87	29.06	28.43	28.30	23.02	24.97	28.43	27.64
FGSM-SDI	1	Best	58.64	37.23	29.19	28.60	28.78	27.99	27.67	23.27	25.85	27.83	29.00
		Last	58.54	37.19	29.17	28.53	28.71	28.00	27.72	23.18	25.55	27.89	28.94
FGSM-PGI	1	Best	**58.78**	40.02	31.84	31.43	31.94	31.30	31.19	25.65	28.23	31.21	30.35
		Last	**58.82**	39.83	31.56	31.22	31.65	31.18	30.89	25.43	27.75	30.93	30.27
GradAlign	1	Best	54.90	35.28	27.50	26.77	27.13	26.52	26.22	22.30	25.01	26.39	27.20
		Last	55.22	35.51	27.40	26.82	27.12	26.42	26.24	22.19	24.94	26.52	27.62
N-FGSM	1	Best	54.41	35.00	27.59	26.99	27.01	26.55	26.34	22.81	25.08	26.31	27.42
		Last	54.41	35.00	27.59	26.99	27.01	26.55	26.34	22.81	25.08	26.31	27.42
Ours	1	Best	57.32	**40.29**	**33.73**	**33.33**	**33.56**	**33.17**	**33.06**	**26.61**	**28.47**	**33.15**	**31.06**
		Last	57.32	**40.29**	**33.73**	**33.33**	**33.56**	**33.17**	**33.06**	**26.61**	**28.47**	**33.15**	**31.06**
MART	10	Best	54.51	38.62	32.37	32.00	32.18	31.68	31.59	26.07	28.01	31.55	29.92
		Last	54.75	38.52	32.18	31.75	31.85	31.37	31.21	25.71	27.81	31.22	29.87
LAS-AWP	10	Best	58.75	40.66	32.98	32.42	32.58	31.91	31.74	27.23	29.59	31.74	32.30
		Last	58.75	40.66	32.98	32.42	32.58	31.91	31.74	27.23	29.59	31.74	32.30

Table 1: Accuracy Comparisons of Different AT Methods on CIFAR-100. Bold numbers in this table indicate the best results of single-step AT. Best indicates the best accuracy during training stage and Last indicates the accuracy from last epoch.

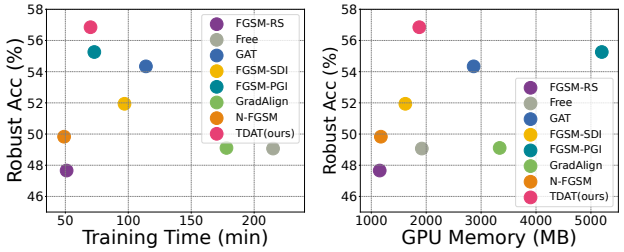


Figure 5: Comparison of robust accuracy (PGD-10), training time, and GPU memory of different single-step AT methods on CIFAR-10 dataset. The method in upper left corner of each subfigure denotes the better robustness and efficiency.

GAT (Sriramanan et al. 2020), FGSM-SDI (Jia et al. 2022c), FGSM-PGI (Jia et al. 2022a), GradAlign (Andriushchenko and Flammarion 2020), and N-FGSM (de Jorge Aranda et al. 2022). Furthermore, the multi-step AT is also considered for demonstrating the extension of TDAT, which contains MART (Wang et al. 2020) and LAS-AWP (Jia et al. 2022b). Specially, FGSM-RS (Wong, Rice, and Kolter 2020) are implemented with larger step size $10/255$ to avoid CO.

Comparison Experiments and Analysis

Due to page limitations, the performance of TDAT compared with other AT methods on CIFAR-10, Tiny ImageNet, and ImageNet-100 is presented in the Tables 1, 2, and 3 of **Supplementary Materials**, respectively.

Results on CIFAR-10 Our method achieves the best robust accuracy in the best checkpoint against FGSM ($+1.13\%$), MIFGSM ($+1.92\%$), BIM ($+1.70\%$), PGD-10 ($+1.59\%$), PGD-20 ($+1.51\%$), PGD-50 ($+1.52\%$), and APGD ($+1.32\%$) attacks compared to other single-step AT methods. Meanwhile, our method exhibits competitive robust accuracy against AA, C&W, and Square attacks, with only a slight decrease of 0.27%, 0.7%, and 0.29% compared to the best method. Furthermore, the results also verify that our proposed method yields better robust accuracy compared to multi-step AT. Moreover, as shown in Fig. 5, TDAT can obtain the state-of-the-art robustness while requiring less time and memory resources.

Results on CIFAR-100 The experiment results of TDAT compared with other AT methods on the CIFAR-100 dataset are presented in Table 1. Our methods demonstrate competitive performance compared to single- and multi-step AT methods. Specifically, our method achieves the best robust accuracy when defending against FGSM (0.27%), MIFGSM ($+1.89\%$), BIM ($+1.90\%$), PGD-10 ($+1.62\%$), PGD-20 ($+1.87\%$), PGD-50 ($+1.87\%$), AA ($+0.96\%$), C&W ($+0.24\%$), APGD ($+1.94\%$), and Square ($+0.71\%$) attacks compared with other existing single-step AT methods. Moreover, its best performance is generated by the checkpoint obtained from the last epoch, which avoids early stopping method and eliminates comparing the performance of each checkpoint.

Results on Tiny ImageNet Our method achieves a competitive robust accuracy in the best checkpoint against

FGSM (+1.83%), MIFGSM (+1.05%), BIM (+0.77%), PGD-10 (+0.71%), PGD-20 (+0.6%), PGD-50 (+0.58%), APGD (+0.64%), and Square (+1.17%) attacks compared to the other single-step AT methods. On the other aspect, there is only a small gap between the performance obtained at the last training epoch and the best epoch. Specifically, our method achieves state-of-the-art results against FGSM (+2.2%), MIFGSM (+2.48%), BIM (+2.43%), PGD-10 (+2.35%), PGD-20 (+2.35%), PGD-50 (+2.33%), AA (+1.19%), C&W (+1.45%), APGD (+2.29%), and Square (+1.94%) attacks from the last epoch.

Results on ImageNet-100 TDAT achieves a competitive robust accuracy in the best checkpoint against FGSM (+0.14%), MIFGSM (+1%), BIM (+1.24%), PGD-10 (+1.26%), PGD-20 (+1.22%), PGD-50 (+1.4%), APGD (+1.3%), and Square (+1.06%) attacks compared to the other single-step AT methods. Furthermore, our proposed method obtains the best robust accuracy against all considered adversarial attacks compared to other AT methods in term of the last training epoch.

Ablation Study

Effects of Each Component First, when only batch momentum initialization is involved, the training stability is enhanced with CO eliminated, and the robustness of the neural network is improved. However, the robust accuracy is still unsatisfactory. Second, only utilizing the proposed loss or dynamic label relaxation is able to improve the clean accuracy and accelerate convergence, yet cannot void CO, leading to a breakdown in robust precision. The results demonstrate a noticeable improvement in the clean accuracy of the neural network upon employing the proposed loss but at the expense of robust accuracy. When dynamic label relaxation and taxonomy driven loss are involved simultaneously, robust accuracy can be improved with a little loss of clean accuracy (as presented in the second and last row of Table 2).

Effects of Components Cooperation The best performance of TDAT is obtained by adopting all of the components, achieving an accuracy of 82.25% for clean accuracy and 56.85% for robust accuracy. Generally, the robust accuracy of neural networks is significantly improved with only a small reduction in clean accuracy when all components are involved. Note that the stability of the training procedure in AT is essential and eliminates the requirements for performance comparison between each epoch or the early stopping strategy. The best checkpoint of our method can be obtained directly from last epoch.

Init†	Label‡	Loss*	Clean Acc Best/Last	PGD-10 Best/Last
✗	✗	✗	64.29/91.33	41.70/14.60
✓	✗	✗	82.50/82.91	53.82/53.70
✓	✓	✗	86.73/89.04	47.66/45.17
✓	✗	✓	83.83/84.10	51.00/50.55
✓	✓	✓	82.25/82.25	**56.85/56.85**

Table 2: Ablation Study Results. Init†, Label‡ and Loss* represent batch momentum initialization, dynamic label relaxation and taxonomy driven loss function. Bold numbers in this table indicate the best robust accuracy of single-step AT.

Inner†	Outer*	Clean Acc Best/Last	PGD-10 Best/Last
0.15	0.9	85.49/85.56	40.69/40.64
0.15	0.6	85.76/85.99	44.49/44.19
0.15	0.15	82.25/82.25	**56.85/56.85**
0.6	0.15	81.92/81.99	54.93/54.61
0.9	0.15	81.22/81.43	54.60/54.24
0.6	0.6	83.41/83.42	54.01/53.49
0.9	0.9	83.88/84.01	52.47/52.19

Table 3: Analyses on Label Relaxation Factor. Bold numbers in this table indicate the best robust accuracy of single-step AT. Inner† and Outer* represent relaxation factor of inner maximization and outer minimization, respectively.

Relaxation Factor Dynamic label relaxation is adopted on the labels of both inner and outer optimization. The corresponding analyses are presented in Table 3 and Fig. 6. Specifically, when the factor of the inner maximization problem remains at 0.15, increasing the factor of the outer minimization problem can significantly reduce the robust accuracy. Meanwhile, when the factor of the outer minimization problem remains at 0.15, only increasing the factor of the maximization problem leads to a mild decrease in both clean and robust accuracy. These findings validate our previous thought that improving the outer minimization problem is more crucial than the inner maximization problem. Moreover, we discover that using same factor in both inner and outer optimization is significant from Table 3. Therefore, as shown in Fig. 6, we further select 35 relaxation factors ranging from 0.11 to 1 in order to demonstrate the impact of same factors used in both optimization. As the factor increases, the robust accuracy gradually decreases and the clean accuracy gradually increases. More analyses on training epochs are provided in **Supplementary Materials**.

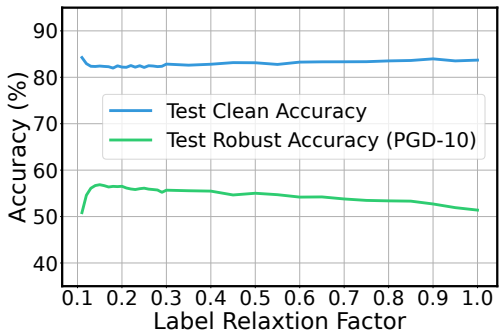


Figure 6: Clean and Robust (PGD-10) Accuracy of our method on CIFAR-10 with different relaxation factors.

Conclusion

In this paper, we find that catastrophic overfitting (CO) occurs in single-step adversarial training (AT) which is caused by a specific kind of adversarial examples (AEs). Specifically, the aim of generating AEs in single-step AT is distorted, resulting in a sudden decrease in generalization. To address this issue, we proposed a taxonomy of AEs to investigate the relationship between kinds of training examples and CO. On this basis, taxonomy driven fast AT (TDAT) is proposed, which involves batch momentum initialization, dynamic label relaxation, and taxonomy driven loss. Our TDAT is a systematic improvement of the single-step AT, thereby can be regarded as a new training paradigm. Comprehensive experimental results demonstrate the proposed method successfully alleviates CO and achieves significantly improved robust accuracy compared with other state-of-the-art single- and even multi-step AT methods.

Acknowledgments

This work was supported in part by the grant of the National Science Foundation of China under Grant 62172090, 62202438; Start-up Research Fund of Southeast University under Grant RF1028623097; CAAI-Huawei MindSpore Open Fund. We thank the Big Data Computing Center of Southeast University for providing the facility support on the numerical calculations in this paper.

References

Andriushchenko, M.; Croce, F.; Flammarion, N.; and Hein, M. 2020. Square Attack: A Query-Efficient Black-Box Adversarial Attack via Random Search. In *Proceedings of the European Conference on Computer Vision(ECCV)*, 484–501.

Andriushchenko, M.; and Flammarion, N. 2020. Understanding and improving fast adversarial training. *Advances in Neural Information Processing Systems*, 33: 16048–16059.

Athalye, A.; Carlini, N.; and Wagner, D. 2018. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning*, 274–283. PMLR.

Bai, T.; Luo, J.; Zhao, J.; Wen, B.; and Wang, Q. 2021. Recent Advances in Adversarial Training for Adversarial Robustness. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence(IJCAI)*, 4312–4321. Survey Track.

Carlini, N.; and Wagner, D. 2017. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*, 39–57.

Chen, H.; and Ji, Y. 2022. Adversarial Training for Improving Model Robustness? Look at Both Prediction and Interpretation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10): 10463–10472.

Chen, T.; Zhang, Z.; Liu, S.; Chang, S.; and Wang, Z. 2021. Robust overfitting may be mitigated by properly learned smoothening. In *International Conference on Learning Representations*.

Chen, Z.; Wang, Z.; Huang, J.-J.; Zhao, W.; Liu, X.; and Guan, D. 2023. Imperceptible Adversarial Attack via Invertible Neural Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(1): 414–424.

Chiang, P.-H.; Chan, C.-S.; and Wu, S.-H. 2021. Adversarial pixel masking: A defense against physical attacks for pretrained object detectors. In *Proceedings of the 29th ACM International Conference on Multimedia*, 1856–1865.

Croce, F.; and Hein, M. 2020. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning*, 2206–2216. PMLR.

Cui, J.; Liu, S.; Wang, L.; and Jia, J. 2021. Learnable boundary guided adversarial training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15721–15730.

de Jorge Aranda, P.; Bibi, A.; Volpi, R.; Sanyal, A.; Torr, P.; Rogez, G.; and Dokania, P. 2022. Make some noise: Reliable and efficient single-step adversarial training. *Advances in Neural Information Processing Systems*, 35: 12881–12893.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Ding, D.; Zhang, M.; Feng, F.; Huang, Y.; Jiang, E.; and Yang, M. 2023. Black-Box Adversarial Attack on Time Series Classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6): 7358–7368.

Ding, J.; Bu, T.; Yu, Z.; Huang, T.; and Liu, J. 2022. SNN-RAT: Robustness-enhanced spiking neural network through regularized adversarial training. *Advances in Neural Information Processing Systems*, 35: 24780–24793.

Dong, Y.; Liao, F.; Pang, T.; Su, H.; Zhu, J.; Hu, X.; and Li, J. 2018. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9185–9193.

Gao, R.; Wang, J.; Zhou, K.; Liu, F.; Xie, B.; Niu, G.; Han, B.; and Cheng, J. 2022. Fast and reliable evaluation of adversarial robustness with minimum-margin attack. In *International Conference on Machine Learning*, 7144–7163. PMLR.

Goodfellow, I.; Shlens, J.; and Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations*.

He, B.; Liu, J.; Li, Y.; Liang, S.; Li, J.; Jia, X.; and Cao, X. 2023. Generating Transferable 3D Adversarial Point Cloud via Random Perturbation Factorization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(1): 764–772.

Herrmann, C.; Sargent, K.; Jiang, L.; Zabih, R.; Chang, H.; Liu, C.; Krishnan, D.; and Sun, D. 2022. Pyramid adversarial training improves vit performance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13419–13429.

Izmailov, P.; Podoprikhin, D.; Garipov, T.; Vetrov, D.; and Wilson, A. G. 2018. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*.

Jia, X.; Zhang, Y.; Wei, X.; Wu, B.; Ma, K.; Wang, J.; and Cao, X. 2022a. Prior-Guided Adversarial Initialization for Fast Adversarial Training. In *Proceedings of the European Conference on Computer Vision(ECCV)*, 567–584.

Jia, X.; Zhang, Y.; Wu, B.; Ma, K.; Wang, J.; and Cao, X. 2022b. LAS-AT: adversarial training with learnable attack strategy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13398–13408.

Jia, X.; Zhang, Y.; Wu, B.; Wang, J.; and Cao, X. 2022c. Boosting fast adversarial training with learnable adversarial initialization. *IEEE Transactions on Image Processing*, 31: 4417–4430.

Kim, H.; Lee, W.; and Lee, J. 2021. Understanding catastrophic overfitting in single-step adversarial training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 8119–8127.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Kurakin, A.; Goodfellow, I. J.; and Bengio, S. 2017. Adversarial Machine Learning at Scale. In *International Conference on Learning Representations*.

Le, Y.; and Yang, X. S. 2015. Tiny ImageNet Visual Recognition Challenge.

Li, B.; and Liu, W. 2023. WAT: Improve the Worst-Class Robustness in Adversarial Training. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12): 14982–14990.

Li, C.; Qiu, Q.; Zhang, Z.; Guo, J.; and Cheng, X. 2023. Learning Adversarially Robust Sparse Networks via Weight Reparameterization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(7): 8527–8535.

Li, T.; Wu, Y.; Chen, S.; Fang, K.; and Huang, X. 2022. Subspace adversarial training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13409–13418.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations*.

Mao, X.; Chen, Y.; Duan, R.; Zhu, Y.; Qi, G.; Li, X.; Zhang, R.; Xue, H.; et al. 2022. Enhance the visual representation via discrete adversarial training. *Advances in Neural Information Processing Systems*, 35: 7520–7533.

Mei, S.; Zhao, C.; Ni, B.; and Yuan, S. 2023. Towards Interpreting and Utilizing Symmetry Property in Adversarial Examples. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(8): 9126–9133.

Miyato, T.; Maeda, S.-i.; Koyama, M.; and Ishii, S. 2018. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8): 1979–1993.

Müller, R.; Kornblith, S.; and Hinton, G. E. 2019. When does label smoothing help? *Advances in Neural Information Processing Systems*, 32.

Pang, T.; Lin, M.; Yang, X.; Zhu, J.; and Yan, S. 2022. Robustness and accuracy could be reconcilable by (proper) definition. In *International Conference on Machine Learning*, 17258–17277. PMLR.

Pang, T.; Yang, X.; Dong, Y.; Su, H.; and Zhu, J. 2021. Bag of tricks for adversarial training. In *International Conference on Learning Representations*.

Park, G. Y.; and Lee, S. W. 2021. Reliably fast adversarial training via latent adversarial perturbation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7758–7767.

Phan, H.; Yin, M.; Sui, Y.; Yuan, B.; and Zonouz, S. 2023. CSTAR: Towards Compact and Structured Deep Neural Networks with Adversarial Robustness. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(2): 2065–2073.

Qin, Y.; Xiong, Y.; Yi, J.; and Hsieh, C.-J. 2023. Training Meta-Surrogate Model for Transferable Adversarial Attack. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(8): 9516–9524.

Rice, L.; Wong, E.; and Kolter, Z. 2020. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*, 8093–8104. PMLR.

Sankaranarayanan, S.; Jain, A.; Chellappa, R.; and Lim, S. N. 2018. Regularizing deep networks using efficient layerwise adversarial training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Shafahi, A.; Najibi, M.; Ghiasi, M. A.; Xu, Z.; Dickerson, J.; Studer, C.; Davis, L. S.; Taylor, G.; and Goldstein, T. 2019. Adversarial training for free! *Advances in Neural Information Processing Systems*, 32.

Sriramanan, G.; Addepalli, S.; Baburaj, A.; et al. 2020. Guided adversarial attack for evaluating and enhancing adversarial defenses. *Advances in Neural Information Processing Systems*, 33: 20297–20308.

Sriramanan, G.; Addepalli, S.; Baburaj, A.; et al. 2021. Towards efficient and effective adversarial training. *Advances in Neural Information Processing Systems*, 34: 11821–11833.

Sriramanan, G.; Gor, M.; and Feizi, S. 2022. Toward Efficient Robust Training against Union of L_p Threat Models. *Advances in Neural Information Processing Systems*, 35: 25870–25882.

Tramèr, F.; Kurakin, A.; Papernot, N.; Goodfellow, I.; Boneh, D.; and McDaniel, P. 2018. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations*.

Tsiligkaridis, T.; and Roberts, J. 2022. Understanding and Increasing Efficiency of Frank-Wolfe Adversarial Training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 50–59.

Wang, Y.; Zou, D.; Yi, J.; Bailey, J.; Ma, X.; and Gu, Q. 2020. Improving adversarial robustness requires revisiting misclassified examples. In *International Conference on Learning Representations*.

Wong, E.; Rice, L.; and Kolter, J. Z. 2020. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations*.

Yu, M.; and Sun, S. 2022. FE-DaST: Fast and effective data-free substitute training for black-box adversarial attacks. *Computers & Security*, 113: 102555.

Zhang, D.; Zhang, T.; Lu, Y.; Zhu, Z.; and Dong, B. 2019a. You only propagate once: Accelerating adversarial training via maximal principle. *Advances in Neural Information Processing Systems*, 32.

Zhang, H.; Yu, Y.; Jiao, J.; Xing, E.; El Ghaoui, L.; and Jordan, M. 2019b. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, 7472–7482. PMLR.

Zhang, J.; Xu, X.; Han, B.; Niu, G.; Cui, L.; Sugiyama, M.; and Kankanhalli, M. 2020. Attacks which do not kill training make adversarial learning stronger. In *International conference on machine learning*, 11278–11287. PMLR.

Zhang, Y.; Zhang, G.; Khanduri, P.; Hong, M.; Chang, S.; and Liu, S. 2022. Revisiting and advancing fast adversarial training through the lens of bi-level optimization. In *International Conference on Machine Learning*, 26693–26712. PMLR.