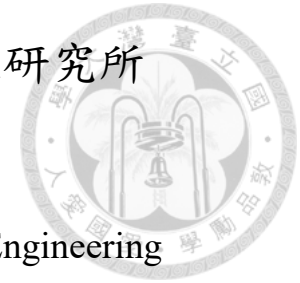


國立臺灣大學電機資訊學院資訊工程研究所

碩士論文

Department of Computer Science and Information Engineering

College of Electrical Engineering and Computer Science

National Taiwan University

Master Thesis

利用漸進式自蒸餾方法改進對抗式訓練

Annealing Self-Distillation Rectification Improves Adversarial Training

吳由由

Yu-Yu Wu

指導教授: 陳尚澤 博士

Advisor: Shang-Tse Chen Ph.D.

中華民國 112 年 6 月

June, 2023

doi:10.6342/NTU202301129

國立臺灣大學碩士學位論文
口試委員會審定書
MASTER'S THESIS ACCEPTANCE CERTIFICATE
NATIONAL TAIWAN UNIVERSITY

利用漸進式自蒸餾方法改進對抗式訓練

Annealing Self-Distillation Rectification Improves Adversarial Training

本論文係吳由由君（學號 R10922018）在國立臺灣大學資訊工程學系完成之碩士學位論文，於民國 112 年 6 月 12 日承下列考試委員審查通過及口試及格，特此證明。

The undersigned, appointed by the Department of Computer Science and Information Engineering on 12 June 2023 have examined a Master's thesis entitled above presented by YU-YU WU (student ID: R10922018) candidate and hereby certify that it is worthy of acceptance.

口試委員 Oral examination committee:

陳向澤　　曹昱

（指導教授 Advisor）

陳祝嵩

系主任/所長 Director: 洪士灝

doi:10.6342/NTU202301129

doi:10.6342/NTU202301129

Acknowledgements

在碩士期間我學習了做研究的方法，從找一個好的題目，到尋找問題的解法，執行實驗，再把自己的論述條理分明的寫成文字，每一步都讓人感受到做研究的困難與艱辛。從一開始 survey 了很多方向的論文，但技術上無從下手，到後來感受到自己工程技術不斷進步，相關的問題都能以工程方法解決，唯獨研究上仍做不出突破。無論在哪個階段，都讓人感到煎熬萬分。直到最後，有幸能完成本作，內心充滿著欣慰，也對幫助過我的人感到無限感恩。

感謝我的指導老師 陳尚澤教授，一直以來花了很多時間和我開會、討論，並給予了我時間、資源與自由，讓我能不受限制的往自己感興趣的方向研究，並在我需要時，幫我申請更多的運算資源，讓我能很奢侈、不顧及成本的從一次次的失敗中學習。也謝謝老師給了我許多寫作上的建議與指導，因為有老師的細心指正，提供諸多寶貴意見，本篇論文才能更臻完備，在此獻上我最深刻的謝意。

修業期間，感謝 王竑睿同學，與我互相砥礪，一起討論研究與課業上的問題。並在我遇到瓶頸時，給予我鼓勵讓我能堅持下去。感謝 王俊傑同學，作為實驗室的守護神，裡裡外外幫實驗室打理機器、環境，因為有你負責維護實驗室的運作，其他人才能無後顧之憂的做研究。

最後，我要感謝我的家人一路上不離不棄的陪伴與支持，沒有你們，我大概沒辦法在今天完成碩士學位。

doi:10.6342/NTU202301129

謹將此篇作品獻給每一個曾幫助過我的人。

doi:10.6342/NTU202301129

摘要

在標準的對抗訓練中，模型以獨熱標籤 (one-hot label) 作為優化目標，當對抗式攻擊 (Adversarial Attack) 對資料造成的變異在事先定義可接受的範圍之內，模型都將以相同的標籤作為目標學習。也就是，一定變異內的資料都標上相同的標籤。然而，這種給予標籤的方式使模型忽略了對抗式攻擊產生的變異對資料所帶來的潛在分佈飄移 (distribution shift) 現象，因而導致過度擬合 (overfitting) 的問題出現在對抗式訓練中。為了解決這個問題並強化模型防禦對抗式攻擊，首先，我們先分析了強健模型的特徵，並發現了強健模型更傾向產生平滑且校正良好的輸出。基於這項觀察，我們提出了一個簡單但有效的方法「漸進式自蒸餾校正」Annealing Self-Distillation Rectification，ADR），生成對資料分佈改變更精確描述的軟標籤，對模型訓練提供更好的引導，以準確反應攻擊下的資料分佈偏移狀況。透過漸進式自蒸餾校正得到的校正分佈軟標籤 (rectified labels)，我們得以在沒有預先訓練模型 (pre-trained models) 和額外大量計算的情況下顯著提昇模型的強健性。此外，透過以校正後的軟標籤替換損失函數中的硬標籤，我們的方法可以很方便的與其他對抗式訓練的演算法結合。我們的實驗在廣泛的資料集與模型架構上都得到了強勁的表現，驗證了漸進式自蒸餾校正是個有效改善對抗式訓練，並防止過度擬合的方法。

關鍵字：對抗式訓練、過度擬合、知識蒸餾

doi:10.6342/NTU202301129

doi:10.6342/NTU202301129

Abstract

In standard adversarial training, models are optimized to fit one-hot labels within allowable adversarial perturbation budgets. However, the ignorance of underlying distribution shifts brought by perturbations causes the problem of robust overfitting. To address this issue and enhance adversarial robustness, we analyze the characteristics of robust models and identify that robust models tend to produce smoother and well-calibrated outputs. Based on the observation, we propose a simple yet effective method, **A**nnealing Self-**D**istillation **R**ectification (ADR), which generates soft labels as a better guidance mechanism that accurately reflects the distribution shift under attack during adversarial training. By utilizing ADR, we can obtain rectified distributions that significantly improve model robustness without the need for pre-trained models or extensive extra computation. Moreover, our method facilitates seamless plug-and-play integration with other adversarial training techniques by replacing the hard labels in their objectives. We demonstrate the efficacy of ADR through extensive experiments and strong performances across datasets.

doi:10.6342/NTU202301129

Keywords: Adversarial Training, Robust Overfitting, Knowledge Distillation

doi:10.6342/NTU202301129

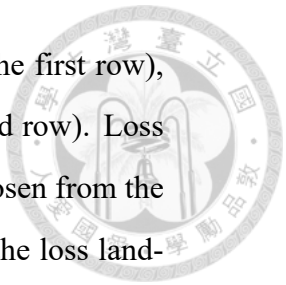
Contents

doi:10.6342/NTU202301129

doi:10.6342/NTU202301129

List of Figures

doi:10.6342/NTU202301129



doi:10.6342/NTU202301129

List of Tables

doi:10.6342/NTU202301129

doi:10.6342/NTU202301129

Denotation

AT	Adversarial Training
DNN	Deep Neural Network
KD	Knowledge Distillation
ADR	Annealing Self-Distillation Rectification
WA	Weight Average
AWP	Adversarial Weight Perturbation
EMA	Exponential Moving Average
PGD-AT	Projected Gradient Descent Adversarial Training
TRADES	TRadeoff-inspired Adversarial DEfense via Surrogate-loss minimization
TE	Temporal Ensembling
MT	Mean Teacher

doi:10.6342/NTU202301129

OOD	Out Of Distribution
JS divergence	Jensen–Shannon divergence
RA	Robust Accuracy
SA	Standard Accuracy

doi:10.6342/NTU202301129

Chapter 1 Introduction

Deep Neural Network (DNN) has been shown to exhibit susceptibility to adversarial attacks [55], wherein intentionally crafted imperceptible perturbations introduced into the original input cause the model's predictions to be altered. Among various defense methods [9, 32, 37, 42, 60], Adversarial Training (AT) [39] stands out as one of the most effective techniques [3, 59] to enhance DNN's adversarial robustness. However, while AT has proven effective in countering adversarial attacks, it is not immune to the problem of robust overfitting [48]. AT constrains the model to generate consistent output when subjected to perturbations within an ϵ attack budget. However, manually assigned hard labels are noisy [15, 16] for AT since they fail to reflect shifts in the data distribution. Minimizing the adversarial training loss results in worse generalization ability on the test data. To address this issue, several approaches have been proposed, including label smoothing [43], consistency regularization [16, 57], and knowledge distillation [8, 13, 19, 66–68], which create a smoother objective function to alleviate the problem of robust overfitting. However, these approaches often assume uniform label noise [43], require consistency in the model's output over time [16, 57], or depend on an additional robust model trained in advance [8, 13, 19, 66–68]. None of these methods directly focus on designing a well-rectified target without the need for pre-trained models.

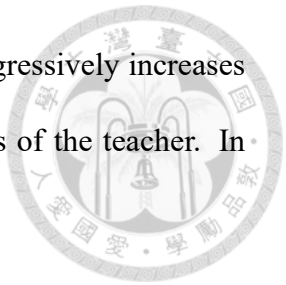
To enhance AT, we investigate the characteristics that distinguish a robust model

doi:10.6342/NTU202301129

from a non-robust one by analyzing the disparity in output distributions. Our findings indicate that a robust model should possess good calibration ability, which is manifested by a lower average confidence level when it is likely to make errors. In addition, robust models' output distribution should remain consistent for the clean data and its adversarial counterpart. Based on this observation, we propose a novel approach called **A**nnealing Self-**D**istillation **R**ectification (ADR), which interpolates the output of the model weight's momentum encoder with the one-hot ground truth to generate noise-aware labels to reflect the distribution shift under attack. The intuition behind this method is that if an image becomes more similar to the attacked class after being perturbed, we should assign a higher probability to the attacked class instead of keeping it 0. However, the ground truth class should maintain a dominant probability across the distribution to ensure the final prediction is not altered.

The weight momentum encoding scheme, also known as Mean Teacher, is a widely used technique in semi-supervised [57] and self-supervised [6, 22] learning that involves updating a teacher model using an exponential moving average (EMA) on the student weights. The self-distillation EMA model also serves as a Weight Average (WA) [18] method, which smoothes the loss landscape and enhances model robustness [8, 20, 29]. The constantly superior robust performance of EMA can serve as an ideal guide for its student during training. To ensure that our model generates well-calibrated results that reflect inter-class relations within examples, we introduce a softmax temperature [27] to scale the output of the EMA teacher. Initially, the temperature is set to a larger value resembling a uniform distribution but gradually decreases following a cosine annealing as the teacher expresses better inter-class relationships after training. We further interpolate the one-hot label with the temperature-scaled distribution to assure the ground truth

doi:10.6342/NTU202301129



class always has the highest probability. The interpolation factor progressively increases to reflect our increasing level of trust in the accuracy and robustness of the teacher. In summary, our contributions are:

- We conduct a comprehensive analysis of the output properties of robust models. Our investigation reveals that the robust models are better calibrated and generate higher entropy output regardless of the input type. Robust models also maintain output consistency on benign data and its adversarial counterpart.

- We propose **A**nnealing Self-**D**istillation **R**ectification (ADR), a simple yet effective technique that leverages soft, noise-aware label reformulation to refine the original one-hot label. Through this method, we obtain well-calibrated results without requiring pre-trained models or extensive extra computational resources. Additionally, ADR can be easily incorporated into other adversarial training algorithms by substituting the hard label in their objectives, thus enabling a seamless plug-and-play integration.

- Our experimental results across multiple datasets demonstrate the efficacy of ADR in improving adversarial robustness. Substitute the hard labels with well-calibrated ones generated by ADR alone can achieve remarkable gains in robustness. When combined with other AT tricks (WA, AWP), ADR further outperforms the state-of-the-art results on CIFAR-100 and TinyImageNet-200 datasets with various architectures.

doi:10.6342/NTU202301129

doi:10.6342/NTU202301129

Chapter 2 Related Work

Adversarial training (AT) has been demonstrated to be effective in enhancing the white box robustness benchmarks [12] of DNN. PGD-AT [39], which introduces worst-case inputs during training, has been the most popular approach for improving DNN's robustness. An alternative AT method, TRADES [64], provides a systematic approach to regulating the trade-off between natural accuracy and robustness and has yielded competitive results across multiple datasets. Despite the efficacy, AT often suffers from robust overfitting [48]. Below, we summarize works that address the issue of robust overfitting by reforming the label.

2.1 Robust Overfitting

The phenomenon of robust overfitting [48] represents a significant challenge in AT, motivating researchers to explore various avenues for mitigation. One such approach is to use heuristic-driven augmentations [47], such as CutMix [46, 62], DAJAT [1], and CropShift [36], which employ sets of augmentations carefully designed to increase data diversity and alleviate robust overfitting. Another strategy involves the expansion of the training set, which offers a direct means to address overfitting. By incorporating additional unlabeled data [5] or high-quality generated images via deep diffusion probabilistic

doi:10.6342/NTU202301129

models (DDPM) [20, 46, 50], the introduction of an extra set of 500K pseudo-labeled images from 80M-TI [58] eliminates the occurrence of robust overfitting [46]. Despite the demonstrated effectiveness of extra data, increasing the size of the training set is computationally expensive, rendering AT infeasible for larger datasets.

Early stopping is a straightforward method for producing robust models [48]. However, due to the fact that the checkpoint of optimal robust accuracy and that of standard accuracy frequently do not align [8], utilizing either of these measures can result in a compromise of overall performance. Weight Average (WA) [8, 20, 29] tracks the exponential moving average of model weights, thereby promoting flatter minima and increased robustness [25]. Another effective regularization technique is Adversarial Weight Perturbation (AWP) [61], which serves to promote flatness within the weight loss landscape and yields enhanced generalization capabilities [54].

2.2 Rectify labels in AT.

AT can smooth the predictive distributions by increasing the likelihood of the target around the ϵ-neighborhood of the observed training examples [33]. A recent study by Grabinski et al. [21] has shown that robust models produced by AT tend to exhibit lower confidence levels than non-robust models, even when evaluated on clean data. Due to the substantial differences in output distributions between robust and standard models, using one-hot labels, which encourage high-confidence predictions on adversarial examples, may not be optimal. Dong et al. [16] and Dong et al. [15] have demonstrated that one-hot labels are noisy in AT, as they are inherited from clean examples while the data had been distorted by attacks. The mismatch between the assigned labels and the true

doi:10.6342/NTU202301129

distributions can exacerbate overfitting compared to standard training. Rectifying labels is shown to be effective in addressing the issue of robust overfitting in AT [15]. Label Smoothing (LS) [56] is a technique that softens labels by combining one-hot targets and a uniform distribution. By appropriately choosing the mixing factor, mild LS can enhance model robustness while calibrating the confidence of the trained model [43, 53]. However, overly smoothing labels in a data-blind manner can diminish the discriminative power of the model [40, 41] and make it susceptible to gradient masking [3]. Prior works [8, 13, 19, 66–68] have utilized Knowledge Distillation (KD) to generate data-driven soft labels, outperforming baseline approaches. Temporal Ensembling (TE) [16] and Mean Teacher (MT) [65] have applied consistency loss into training objectives, thus preventing overconfident predictions through consistency regularization. More recently, Dong et al. [15] have employed a pre-trained robust model to reform training labels, addressing label noise in AT.

doi:10.6342/NTU202301129

doi:10.6342/NTU202301129

Chapter 3 Preliminaries

3.1 Adversarial training (AT)

Given a dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $\mathbf{x}_i \in \mathcal{R}^d$ is a benign example, $y_i \in \{1, \ldots, C\}$ is the ground truth label often encoded as a one-hot vector $\mathbf{y}_i \in \{0,1\}^C$, and C is the total number of classes, PGD-AT [39] can be formulated as the following min-max optimization problem:

$$\min_{\theta} \sum_{i=1}^{n} \max_{\mathbf{x}_i' \in \mathcal{S}(\mathbf{x}_i)} \ell(f_\theta(\mathbf{x}_i'), \mathbf{y}_i) \tag{3.1}$$

where f_θ is a model with parameter θ. ℓ is the cross-entropy loss function, and $\mathcal{S}(\mathbf{x}) = \{\mathbf{x}' : ||\mathbf{x}' - \mathbf{x}||_p \leq \epsilon\}$ is an adversarial region centered at $\mathbf{x}$ with radius $\epsilon > 0$ under l_p-norm threat model.

The adversarial example $\mathbf{x}_i'$ can be obtained by projected gradient descent (PGD) to approximate the inner maximization in adversarial training, which randomly initializes a point within $\mathcal{S}(\mathbf{x}_i)$ and iteratively updates the point for K steps with:

$$\mathbf{x}_i^{t+1} = \Pi_{\mathcal{S}(\mathbf{x}_i)}(\mathbf{x}_i^t + \alpha \cdot \text{sign}(\nabla_{\mathbf{x}} \ell(f_\theta(\mathbf{x}_i^t), \mathbf{y}_i))) \tag{3.2}$$

where $\Pi(.)$ is the projection, α is the attack step size, t denotes iteration count, and $\mathbf{x}_i' = \mathbf{x}_i^K$.

doi:10.6342/NTU202301129

3.2 Distributional difference in the outputs of robust and non-robust model

The standard training approach encourages models to generate confident predictions regardless of the scenario, leading to overconfident outcomes when the testing distribution changes. In contrast, robust models, when compared to their standardly trained counterparts, possess superior calibration properties that exhibit low confidence in incorrect classified examples [21]. In addition, a study conducted by Qin et al. [44] has revealed that poorly calibrated examples are more vulnerable to attacks. The interplay between robustness and confidence calibration motivates us to investigate methods of enhancing the information contained within labels. Therefore, we initiate our analysis by examining the differences in output distribution between robust and normal models.

3.2.1 Robust model generates a random output on OOD data

When there is a significant distribution shift in the testing data, a well-calibrated model is expected to display uncertainty in its predictions by assigning uniformly random probabilities to unseen examples. To analyze the difference in output distributions when predicting out-of-distribution (OOD) data, we follow the approach by Qin et al. [44], Snoek et al. [52]. Specifically, we compare the histogram of the output entropy of the models. We use ResNet-18 to train the models on the CIFAR-10

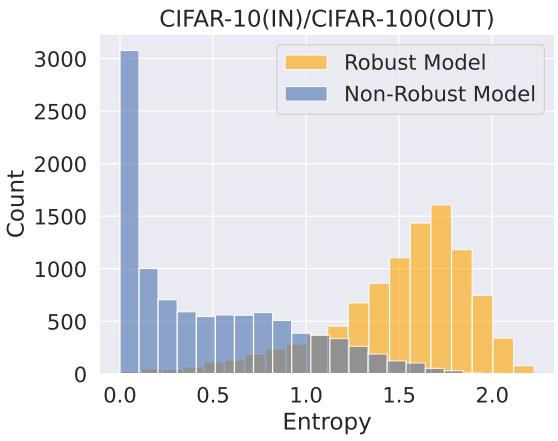
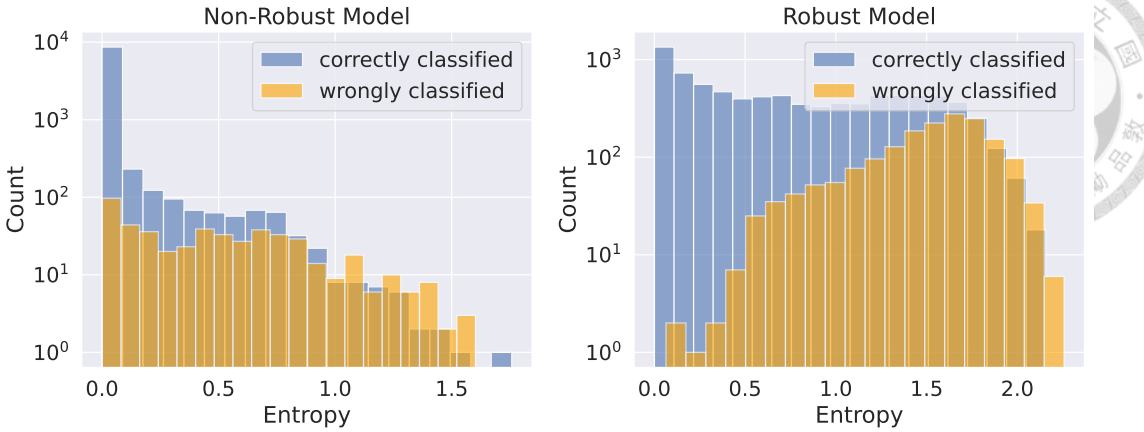


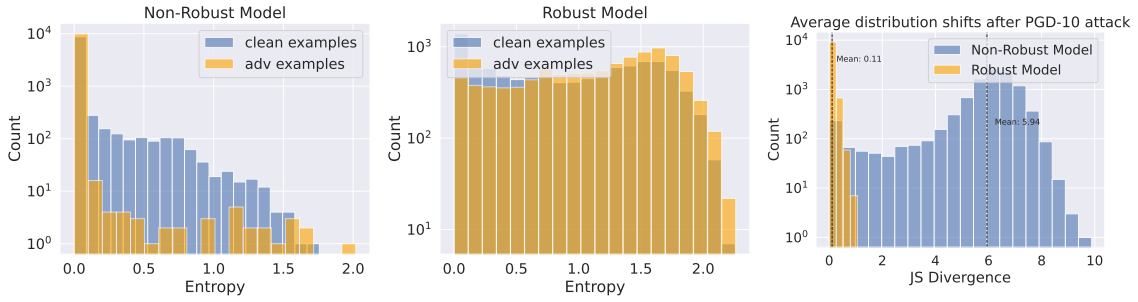
Figure 3.1: Histogram of output entropy on OOD data. Both models are trained on CIFAR-10 and tested on CIFAR-100.

doi:10.6342/NTU202301129



(a) (b)

Figure 3.2: (a) and (b) are entropy distributions on the correctly classified and misclassified examples on the standard and robust model respectively.



(a) (b) (c)

Figure 3.3: (a) and (b) are entropy distributions on the clean and adversarial examples on the standard and robust model respectively. (c) shows histograms of JS divergence for output distribution shift under the PGD-10 attack.

(in-distribution) training set and evaluate the CIFAR-100 (OOD) testing set. Given that most categories in CIFAR-100 were not present during training, we expect the models to generate high entropy output. As shown in Figure 3.1, we observe that the non-robust model has low entropy (high confidence) on OOD data, while the robust model exhibits high uncertainty on average.

3.2.2 Robust models are uncertain on incorrectly classified examples

In order to investigate potential distributional differences in the behavior of standard and robust models when encountering correctly or incorrectly classified examples, we con-

doi:10.6342/NTU202301129

sider the model's confidence level in its predictions. Lower confidence levels are typically associated with higher error rates if the model is well-calibrated. To this end, we conduct an experiment by using a ResNet-18 model trained and evaluated on the CIFAR-10 dataset to demonstrate the distributional differences between correctly and incorrectly classified examples for both standard and robust models. Specifically, Figure 3.2a illustrates that the standard model exhibits low entropy levels for correctly classified examples, but relatively uniform entropy levels for incorrectly classified ones. Higher confidence in the prediction does not guarantee better performance. On the other hand, Figure 3.2b shows that the robust model tends to exhibit relatively high entropy levels (low confidence) for misclassified examples. We can infer that when the robust model is confident in its prediction, the classification accuracy is likely to be high.

3.2.3 Output distribution of models on clean or adversarial examples are consistent

Several prior studies [13, 66] have suggested learning clean images' representation from the standard model and adversarial example's representation from the robust model to improve robustness while maintaining accuracy. However, the underlying assumption that the robust model exhibits comparable representations of clean images to those generated by the standard model has not been thoroughly examined. Therefore, we investigate whether these models show comparable behavior when presented with clean and adversarial (PGD-10) examples on the CIFAR-10 dataset.

We demonstrate that the standard model exhibits low entropy in both scenarios (Figure 3.3a), whereas the robust model yields high entropy on average (Figure 3.3b). Additionally, Figure 3.3c reveals two models' histograms of JS divergence, representing the

doi:10.6342/NTU202301129

extent of output distribution shift when input is attacked. We can observe that even if robust models are attacked successfully, the change of output distribution measured in JS divergence is still small compared to standard models. The robust models show higher consistency (low JS divergence), while the standard models make drastic output changes. Therefore, promoting learning from standard models on normal examples may not be ideal for robust models, as robust models do not generate high confidence output on clean data.

doi:10.6342/NTU202301129

doi:10.6342/NTU202301129

Chapter 4 Methodology

4.1 Motivation: Rectify labels in a noise-aware manner

Based on previous analysis, it can be inferred that robust models should satisfy three key properties: first, they should generate nearly random probability on OOD data; second, they should demonstrate high uncertainty when it is likely to make a mistake; and third, they should exhibit output consistency for both clean examples and their adversarial counterparts. However, the one-hot label used in AT does not provide sufficient guidance to adjust model outputs when the input subtly changes. Restricting the output to be identical when perturbations are within the ϵ ball of input space can be toxic to adversarial training [15], as it causes the model to memorize the labels [16] to minimize training loss, but at the expense of losing generalization ability on the testing set. Therefore, designing a label-softening mechanism that properly reflects the distribution shift of classes is essential to improve robustness. In a recent study conducted by Paleka and Sanyal [41], it was found that uniform label noise has a similar degree of adverse impact as worst-case data poisoning. They also provide empirical evidence that real-world noise is less harmful than uniform-label noise. Specifically, the noise introduced by human annotators poses a lower adversarial risk than uniform label noise. Building on these insights, we propose a data-driven scheme, **A**nnealing Self-**D**istillation **R**ectification (ADR), to rectify labels in

doi:10.6342/NTU202301129

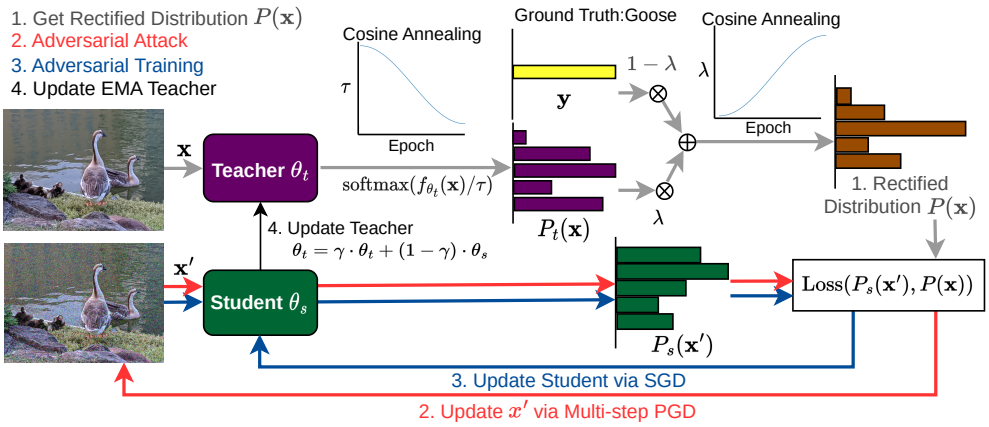


Figure 4.1: Overview of ADR.

a noise-aware manner that mimics the behavior of human annotators.

Algorithm 1 **A**nnealing Self-**D**istillation **R**ectification (ADR)

Input: Training set $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$

Parameter: A classifier $f(.)$ with learnable parameters θ_s; θ_t is exponential moving average of θ_s with decay rate γ; Batch size m; Learning rate η; Total training iterations E; Attack radius ϵ, attack step size α, number of attack iteration K; Temperature τ; Interpolation ratio λ.

Randomly initialize the network parameters θ_s, $\theta_t \leftarrow \theta_s$
for $e = 1$ to E **do**
  Calculate τ_e according to the current iterations.
  Calculate λ_e according to the current iterations.
  Sample a mini-batch $\{(\mathbf{x}_j, y_j)\}_{j=1}^m$ from $\mathcal{D}$
  for $j = 1$ to m (in parallel) **do**
    $P_t(\mathbf{x}_j) \leftarrow \text{softmax}(f_{\theta_t}(\mathbf{x}_j)/\tau_e)$ ▷ Calculate rectified label
    $\lambda_j = \text{clip}_{[0,1]}(\lambda_e - (P_t(\mathbf{x}_j)^{(\psi_j)} - P_t(\mathbf{x}_j)^{(y_j)}))$
    $P(\mathbf{x}_j) \leftarrow \lambda_j \cdot P_t(\mathbf{x}_j) + (1-\lambda_j) \cdot \mathbf{y}_j$
    $\mathbf{x}'_j \leftarrow \mathbf{x}'_j + \epsilon \cdot \delta$, where $\delta \sim \text{Uniform}(-1, 1)$ ▷ Construct adversarial example
    for $k = 1$ to K **do**
      $\mathbf{x}'_j = \Pi_{\mathcal{S}(\mathbf{x}'_j)}(\mathbf{x}'_j + \alpha \cdot \text{sign}(\nabla_{\mathbf{x}} \ell(f_{\theta_s}(\mathbf{x}'_j), P(\mathbf{x}_j))))$
    end for
  end for
  $\theta_s \leftarrow \theta_s - \frac{\eta}{m} \cdot \sum_{j=1}^m \nabla_{\theta_s}(\ell(f_{\theta_s}(\mathbf{x}'_j), P(\mathbf{x}_j)))$ ▷ Update model parameters
  $\theta_t \leftarrow \gamma \cdot \theta_t + (1-\gamma) \cdot \theta_s$
end for

4.2 Annealing Self-Distillation Rectification

To be specific, let θ_s represent the student model's parameter we want to optimize, and θ_t be EMA teacher of θ_s, which is updated by $\theta_t = \gamma \cdot \theta_t + (1-\gamma) \cdot \theta_s$ where γ is

doi:10.6342/NTU202301129

the decay factor. $P_t(\mathbf{x}_i)$ is teacher's output distribution on input $\mathbf{x}_i$, and $P_t(\mathbf{x}_i)^{(c)}$ serves as teacher's predicted probability for the class c. To obtain a rectified label, we first calculate the teacher's softened distribution $P_t(\mathbf{x}_i)$ with τ, which is the temperature of softmax following a cosine annealing from high to low. Since the teacher f_{θ_t} cannot provide much meaningful information at the beginning of training, the high temperature encourages the teacher's output to approach a uniform noise. As f_{θ_t} becomes more accurate and robust, we anneal the temperature to make the distribution more descriptive of the inter-class relations. The smoothed distribution $P_t(\mathbf{x}_i)$ for f_{θ_t} is as follows,

$$P_t(\mathbf{x}_i) = \text{softmax}(f_{\theta_t}(\mathbf{x}_i)/\tau) \tag{4.1}$$

However, f_{θ_t} does not always classify correctly, especially when training is insufficient. To ensure the correct class has the highest probability so the target is unchanged, we interpolate the predicted distribution of f_{θ_t}, $P_t(\mathbf{x}_i)$, with ground-truth one-hot $\mathbf{y}_i$, which is built from y_i, by an interpolation ratio λ. λ follows an increasing cosine schedule, allowing us to trust teachers more over time. For each $\mathbf{x}_i$, we also adjust λ to λ_i to assure the true class has the highest probability across distribution.

$$\lambda_i = \text{clip}_{[0,1]}(\lambda - (P_t(\mathbf{x}_i)^{(\psi_i)} - P_t(\mathbf{x}_i)^{(y_i)})) \tag{4.2}$$

where ψ_i exhibits the teacher's predicted class and y_i represents the ground truth class. When the teacher makes a correct prediction, that is $P_t(\mathbf{x}_i)^{(\psi_i)} - P_t(\mathbf{x}_i)^{(y_i)} = 0$, there is no need to adjust the interpolation rate, otherwise, we decrease λ by the amount that the teacher model makes mistake on $\mathbf{x}_i$ and then clip to the $[0, 1]$ range. Finally, the rectified

doi:10.6342/NTU202301129

smooth distribution $P(\mathbf{x}_i)$ used for adversarial attack and training is carried out as

$$P(\mathbf{x}_i) = \lambda_i \cdot P_t(\mathbf{x}_i) + (1 - \lambda_i) \cdot \mathbf{y}_i \tag{4.3}$$

We use the rectified label $P(\mathbf{x}_i)$ to replace the ground truth label $\mathbf{y}_i$ in Equation 3.1 and Equation 3.2 to conduct adversarial training. Similarly, the softened $P(\mathbf{x}_i)$ can be applied in other adversarial training algorithms, e.g. TRADES [64], by replacing the hard labels. We illustrate the overview of ADR in Figure 4.1 and summarize the pseudo-code in Algorithm 1.

doi:10.6342/NTU202301129

Chapter 5 Experiments

In this section, we compare the proposed ADR to PGD-AT [39] and TRADES [64] in Table 5.1. We further investigate the efficacy of ADR in conjunction with model-weight-space smoothing techniques Weight Average (WA) [20, 29] and Adversarial Weight Perturbation (AWP) [61] with ResNet18 [23] in Table 5.2. Experiments are conducted on well-established benchmark datasets, including CIFAR-10, CIFAR-100 [31], and TinyImageNet-200 [14, 34]. We highlight that defending against attacks on datasets with enormous classes is more difficult as the model's decision boundary becomes complex. We also provide experiments on Wide ResNet (WRN-34-10) [63] with additional data compared to other state-of-the-art methods reported on RobustBench in Table 5.3. Our findings reveal that ADR outperforms methods across different architectures, irrespective of the availability of additional data. The results validate the effectiveness of ADR as a strong defense mechanism. Further investigations on TRADES and TinyImageNet-200 are presented in Section 5.4 and 5.6.

5.1 Training and evaluation setup

We perform adversarial training with perturbation budget $\epsilon = 8/255$ under l_∞-norm in all experiments. In training, we use the 10-step PGD adversary with step size $\alpha =$

doi:10.6342/NTU202301129

$2/255$. We adopt $\beta = 6$ for TRADES as outlined in the original paper. The models are trained using the SGD optimizer with Nesterov momentum of 0.9, weight decay 0.0005, and a batch size of 128. The initial learning rate is set to 0.1 and divided by 10 at 50% and 75% of the total training epochs. Simple data augmentations include 32×32 random crop with 4-pixel padding and random horizontal flip [20, 43, 48] are applied in all experiments. Following Gowal et al. [20], Wu et al. [61], we choose radius 0.005 for AWP and decay rate $\gamma = 0.995$ for WA. For CIFAR-10/100, we use 200 total training epochs, λ follows cosine scheduling from 0.7 to 0.95, and τ is annealed with cosine decreasing from 2.5 to 2 on CIFAR-10 and 1.5 to 1 on CIFAR-100, respectively. As for TinyImageNet-200, we crop the image size to 64×64 and use 80 training epochs. We adjust λ from 0.5 to 0.9 and τ from 2 to 1.5 on this dataset.

During training, we evaluate the model with PGD-10 and select the model that has the highest robust accuracy on the validation set with early stopping [48]. For testing, we use AutoAttack [11] which comprises an ensemble of 4 attacks including APGD-CE [11], APGD-DLR [11], FAB [10] and Square attack [2] for rigorous evaluation. To eliminate the possibility of gradient obfuscations [3], we provide additional sanity checks in Section 5.9. Unless otherwise specified, the robust accuracy (RA) is computed under AutoAttack to demonstrate the model's generalization ability on unseen attacks. The detailed computation cost analysis is provided in Section 5.10.

doi:10.6342/NTU202301129

Table 5.1: Test accuracy (%) of ADR compared to PGD-AT and TRADES with ResNet18. *Best* refers to the checkpoint with the highest robust accuracy on the evaluation set under PGD-10 evaluation. *Final* is the last checkpoint, and *Diff* is the difference of accuracy between *Best* and *Final*. The best results and the smallest performance differences are marked in **bold**.

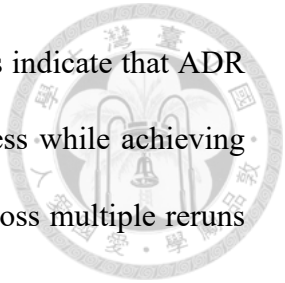
Dataset	Method	AutoAttack(%)			Standard Acc.(%)		
		Best	Final	Diff.	Best	Final	Diff.
CIFAR-10	AT	48.81	43.19	5.62	**82.52**	83.77	**-1.25**
	AT + ADR	**50.38**	**47.18**	**3.20**	82.41	**85.08**	-2.67
	TRADES	50.1	48.17	1.93	82.95	82.42	-0.53
	TRADES + ADR	**51.02**	**50.4**	**0.62**	**83.4**	**83.76**	**-0.36**
CIFAR-100	AT	24.95	19.66	5.29	55.81	56.58	**-0.77**
	AT + ADR	**26.87**	**24.23**	**2.64**	**56.1**	**56.95**	-0.85
	TRADES	24.71	23.78	**0.93**	56.37	**56.01**	**0.36**
	TRADES + ADR	**26.42**	**25.15**	1.27	**56.54**	54.42	2.12
TinyImageNet-200	AT	18.06	15.86	2.2	45.87	**49.35**	-3.48
	AT + ADR	**19.46**	**18.83**	**0.63**	**48.19**	47.65	**0.54**
	TRADES	17.35	16.4	0.95	48.49	47.62	**0.87**
	TRADES + ADR	**19.17**	**18.86**	**0.31**	**51.82**	**50.87**	0.95

5.2 Superior performance across robustified methods and datasets

Table 5.1 demonstrates the results of ADR combined with PGD-AT and TRADES on CIFAR-10, CIFAR-100, and TinyImageNet-200. Our initial observations reveal that robust overfitting exists in all baselines, with differences between final and best early-stopping robust accuracies as large as 5.62% on CIFAR-10 while the standard accuracy (SA) remains stable with more training epochs. When combined with our approach, we observe consistent improvements across datasets, with reduced robust overfitting gaps from 5.62% to 3.2% on CIFAR-10, 5.29% to 2.64% on CIFAR-100, and 2.2% to 0.63% on TinyImageNet-200. Furthermore, the robust accuracy improves by 0.92% to 1.92% across experiments. By alleviating robust overfitting, the best checkpoints are closer to the end of

doi:10.6342/NTU202301129

the training, thereby improving the SA in most settings. Our findings indicate that ADR can effectively enhance the RA-SA trade-off by improving robustness while achieving higher standard accuracy. We also present performance variation across multiple reruns in Section 5.11.



5.3 Combing with weight space smoothing techniques and scaling to larger architecture

The proposed ADR can be integrated with other AT techniques to further boost robustness (Table 5.2). Additional experiments with TRADES are presented in Section 5.4. WA [20] and AWP [61] are the model-weight-space smoothing techniques that improve the stability and performance of AT. In our case, the teacher θ_t in ADR maintains the EMA of the student weights, so we evaluate on θ_t to acquire the WA result. Combining ADR with AWP and WA, we obtain large gains in RA ranging from 1.12% to 3.55% and 0.37% to 3.23% in SA compared to the ResNet-18 baselines.

Following prior works [1, 8, 64], we additionally use Wide ResNet (WRN-34-10) to demonstrate that ADR scales to larger architectures and improves RA and SA. The result shows that our method effectively enhances robust accuracy up to 3.14% on CIFAR-10, 3.15% on CIFAR-100, and 2.57% on TinyImageNet-200 compared to each of its baselines. Notably, we use the same λ, τ as ResNet-18 for WRN-34-10, which might not be optimal, to reduce the cost of hyper-parameter searching. Therefore, we observe a slight drop in standard accuracy in some WRN cases. Nevertheless, ADR still outperforms baselines in robustness without tuning hyper-parameters.

doi:10.6342/NTU202301129

5.4 Test accuracy of TRADES + ADR combing with WA and AWP

In this study, we investigate the impact of combining TRADES and ADR with other adversarial training techniques, namely WA and AWP, on ResNet-18, as outlined in Table-5.4. Our experimental results demonstrate that leveraging a soft target generated by ADR yields exceptional robustness and standard accuracy improvement, thereby achieving a superior trade-off. Specifically, ADR results in 1.18%, 2.92%, and 2.13% RA improvement and 0.45%, 2.21%, and 3.5% SA improvement on the baseline performance of CIFAR-10, CIFAR-100, and TinyImageNet-200, respectively.

However, we observe that the robustness improvement saturates or even slightly deteriorates when combining TRADES+ADR with weight smoothing techniques on CIFAR-10. This is attributed to the fact that TRADES already promotes learning and attacking adversarial data on a softened target, making the additional soft objective by ADR less effective. Nevertheless, the TRADES+ADR approach remains beneficial when dealing with more challenging datasets combined with WA and AWP.

5.5 Comparison with related works and use additional data on CIFAR-100

Table 5.3 compares our proposed ADR defense against related works on a more challenging CIFAR-100 dataset. We select leading methods [1, 7, 13, 20, 30, 45, 46, 50] on RobustBench [12] and methods similar to ours which introduce smoothing in training labels [15, 16, 65] to make a fair comparison. The reported numbers are listed in their original

doi:10.6342/NTU202301129

papers or on RobustBench. We also provide a similar comparison on TinyImageNet-200 in Section 5.6. Given the observed benefits of incorporating additional training data to promote robust generalization [49], we employ a DDPM [28] synthetic dataset [20, 46] composed of 1 million samples. Following [20, 46], we train Preact-ResNet18 [24] and WRN-34-10 with SiLU [26] as the activation function when utilizing synthetic data. We adopt a rigorous experimental design following [20, 46], training the Preact-ResNet18 [24] and WRN-34-10 architectures with SiLU [26] as the activation function when utilizing synthetic data. We leverage cyclic learning rates [51] with cosine annealing [38] by setting the maximum learning rate to 0.4 and warmup period of 10 epochs, ultimately training for a total of 400 CIFAR-100 equivalent epochs. Our training batch size is set to 1024, with 75% of the batch composed of the synthetic data. We maintain consistency with other details outlined in section 5.1.

Our experimentation with AT-WA-AWP-ADR on ResNet-18 yields a robust accuracy of 28.5% and a standard accuracy of 57.36%, comparable to Rebuffi et al. [46] that utilizes an additional 1M DDPM data on Preact-ResNet18, which yields an RA of 28.5% and SA of 56.87%. Remarkably, our model attains equal robustness and superior standard accuracy without using additional data when employing a similar-sized model. Similarly, we achieve an RA of 31.6% on WRN-34-10, while Sehwag et al. [50] scores only 31.15% with additional data. Additionally, adding DDPM data in the training set leads to further improvement in robust accuracy for ADR, by 1.09% and 0.59% for ResNet-18 and WRN-34-10, respectively. In both cases, ADR achieves new state-of-the-art performance, both with and without additional data, on the CIFAR-100 benchmark. It is worth noting that some methods introduce extra auxiliary examples when training [16, 45], and some bring complex augmentation into AT [45, 46], and so might obtain superior SA compared

doi:10.6342/NTU202301129

to ADR. However, regarding the optimal robustness to achieve, we provide compelling evidence that rectifying training labels with a realistic distribution is a valuable approach.

5.6 Test accuracy (%) compared with related works on TinyImageNet-200.

We present ADR evaluated against related works [15, 39, 45, 47] in Table-5.5 on the TinyImageNet-200 dataset, which is a more challenging robustness benchmark than CIFAR-10 or CIFAR-100 due to its larger class size and higher-resolution images, using the original numbers from their respective papers. A model trained on a larger class dataset often results in more complex decision boundaries, which increases the likelihood of an attacker identifying vulnerabilities in the model. Our experimental results demonstrate that ADR achieves state-of-the-art performance, improving RA by 1.94% to 2.17% when using ResNet-18, and achieving a remarkable 2.57% improvement over the baseline on WRN-34-10. In summary, we observe that ADR stands out in the challenging scenario.

5.7 Achieving flatter weight loss landscape

Several studies [54, 61] have found that a flatter weight loss landscape leads to a smaller robust generalization gap when the training process is sufficient. Many methods [8, 20, 61, 64] addressing robust overfitting issues predominantly find flatter minima. We visualize

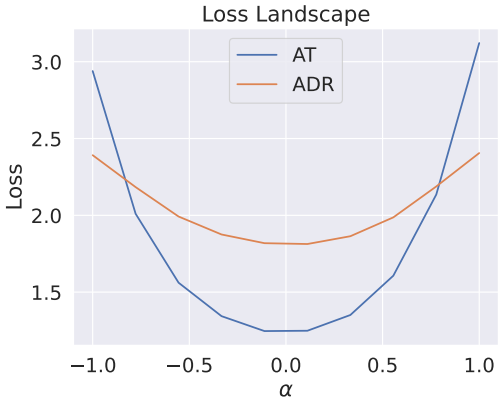


Figure 5.1: Model weight loss landscape comparison for AT and ADR.

doi:10.6342/NTU202301129

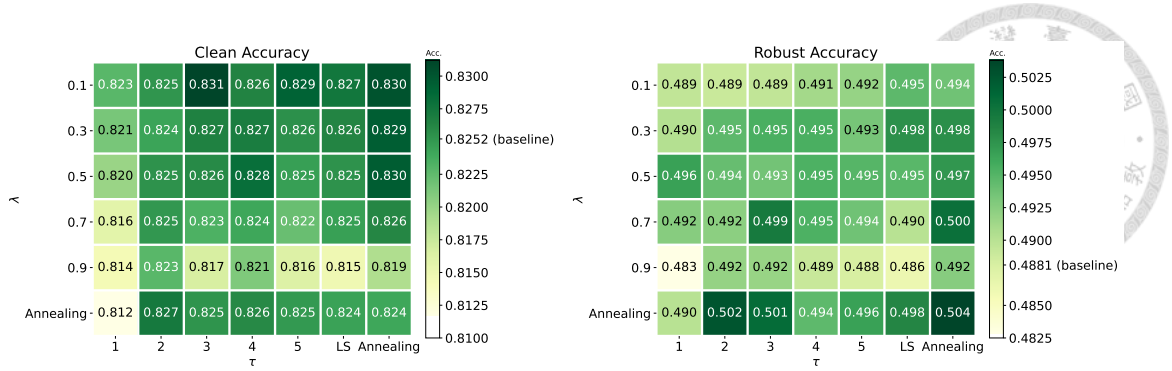


Figure 5.2: Effectiveness of different temperature τ and label interpolation factor λ of ADR.

the weight loss landscape by plotting the loss change when moving the weight w along a random direction d with magnitude α. The direction d is sampled from Gaussian distribution with filter normalization [35]. For each perturbed model, we generate adversarial examples on the fly with PGD-10 [61] and calculate the mean loss across the testing set. Figure 5.1 compares the weight loss landscape between AT and ADR on CIFAR-10. ADR achieves a flatter landscape, implying better robust generalization ability. While we smooth the label for ADR in training, we use the one-hot label as ground truth to calculate the cross-entropy loss in this experiment, so the model trained by ADR has a higher loss value than AT on average. Additionally, we visualize the loss landscape around the data point in Section 5.9 and observe a similar phenomenon that ADR produces a flatter loss landscape.

5.8 Ablation study on the effectiveness of temperature and interpolation factor

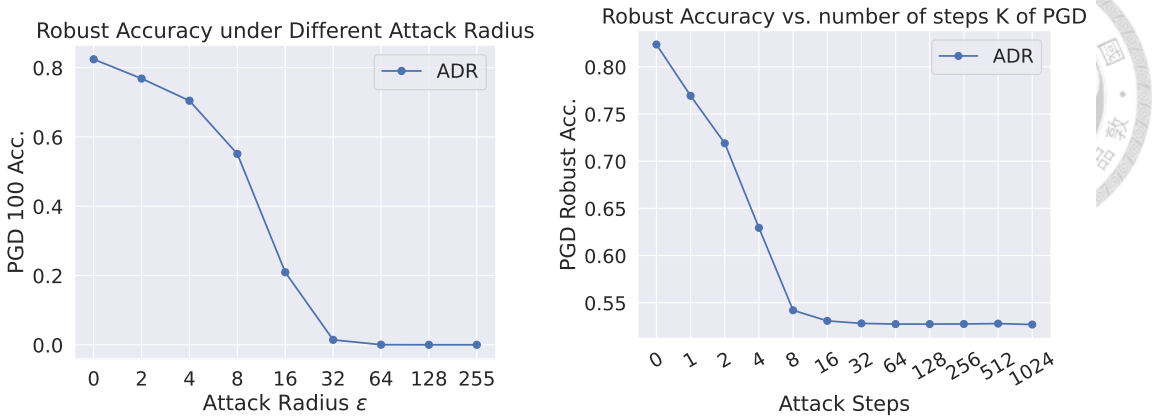
To disclose the impact of temperature τ and interpolation factor λ on the proposed ADR, we conduct ablation studies to exhibit grid search outcomes of SA and RA on CIFAR-10 with ResNet-18 in Figure 5.2. Throughout the experiments, τ and λ are held

doi:10.6342/NTU202301129

constant unless explicitly specified as "Annealing." By not employing annealing, we can scrutinize the effects of varying parameter values. Furthermore, we include Label Smoothing (LS) in this study, which can be viewed as a special case where τ approaches infinity, to evaluate how data-driven smoothness improves performance.

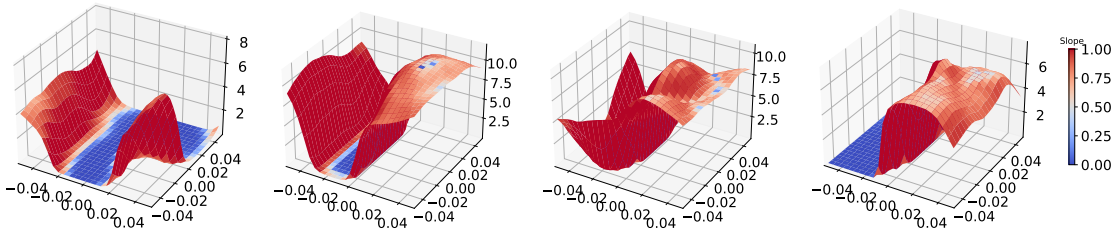
Our analysis reveals that the interpolation factor λ significantly influences clean accuracy. Specifically, as λ increases, more outputs from the EMA are interpolated in ADR , or a higher factor of uniform noise is added in LS, resulting in a monotonic drop in clean accuracy. When the temperature τ is appropriately chosen, $\tau = 3$ in our case, the highest clean accuracy can be attained. In the case of robustness, consistent with prior research [43], we observe that LS yields the highest baseline when $\lambda = 0.3$. In contrast, an increased noise ratio leads to a deterioration in RA. Notably, our findings for ADR demonstrate that the optimal λ value, which is 0.7, is substantially higher than that of LS. This suggests that a greater reliance on the data-driven teacher is beneficial in ADR. When the temperature is small, the resulting label does not provide adequate calibration properties since the distribution is similar to a one-hot vector. When the temperature is high, the output is scaled similarly to a uniform distribution, which cannot provide sufficient inter-class dependency. Therefore, we can infer that selecting a moderately large temperature with a high interpolation factor that provides the most informative distribution to correct the targets results in the best robustness accuracy trade-off. Our experiment also reveals that annealing in both temperature and interpolation factors is beneficial to improve robustness, which shows the efficacy of gradually increasing reliance on the teacher model.

doi:10.6342/NTU202301129

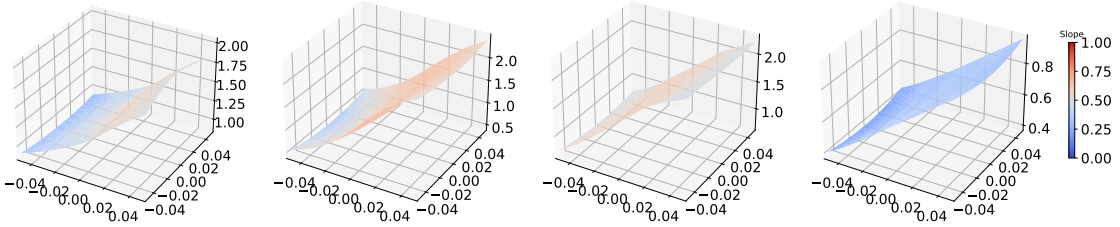


(a) (b)

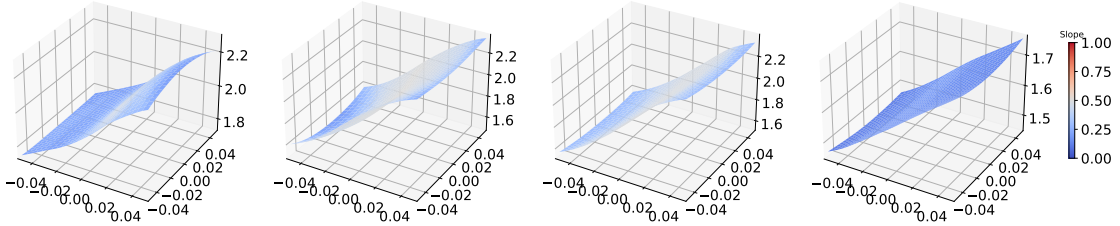
Figure 5.3: The changes of robust accuracy against different attack radius (5.3a) and attack steps (5.3b).



(a) Non-robust model



(b) PGD-AT



(c) ADR

Figure 5.4: Comparison of loss landscapes for the non-robust model (the first row), PGD-AT model (the second row), and ADR model (the third row). Loss plots in each column are generated from the same image chosen from the CIFAR-10. Following the same setting as [8, 17], we plot the loss landscape function $z = loss(x \cdot r_1 + y \cdot r_2)$, where $r_1 = sign(\nabla_i f(i))$ and $r_2 \sim Rademacher(0.5)$. The x and y axes represent the magnitude of the perturbation added in each direction and the z axis represents the loss. i denotes the input and $f(.)$ is the trained model. The color represents the gradient magnitude of the loss landscape clipped in [0, 1] range, which conveys the smoothness of the surface.

doi:10.6342/NTU202301129

5.9 Sanity check for gradient obfuscation

Athalye et al. [3] argued that some defenses improving the robustness by obfuscated gradient, which is introduced intentionally through non-differentiable operations or unintentionally through numerical instability, can be circumvented. Though we already use AutoAttack [11] which has been shown to provide a reliable estimation of robustness as an adversary throughout the experiments, we conduct additional evaluations with ADR to eliminate the possibility of gradient obfuscation. Following the guidelines by Carlini et al. [4], we examine the impact of changes on the robustness with ResNet-18 trained by AT+ADR on CIFAR-10 against l_∞ perturbation.

Unbounded PGD attack The unbounded PGD adversary should reduces model robustness to 0%. Figure 5.3a shows the changes of PGD-100 robust accuracy of AT+ADR with different attack radius ϵ. The robust accuracy for AT+ADR monotonically drops to 0% as the attack radius ϵ increases.

Increasing PGD attack steps Increasing the number of attack iterations should only marginally lower the robust accuracy. Figure 5.3b shows the changes in robust accuracy of AT+ADR with different steps of PGD attack. We can observe that the robust accuracy almost converges after $K = 16$, more steps of attack do not lead to lower robust accuracy.

Inspect loss landscape around inputs Figure 5.4 shows the loss landscape of the Non-Robust, AT, and AT+ADR model around randomly selected examples from CIFAR-10. Compared to the Non-Robust model, the adversarially trained models (both PGD-AT and AT+ADR) have flattened the rugged landscape, which does not exhibit the typical patterns

doi:10.6342/NTU202301129

of gradient obfuscation [17]. It is notable that despite both PGD-AT and ADR having smooth landscapes, ADR has a lower gradient magnitude (dark blue color in the figure), which implies the loss changes for AT+ADR is smaller than AT when small perturbations are added to the input.

5.10 Computation cost analysis

A standard 10 step AT includes 10 forwards and backwards to find the worst-case perturbation when given a normal data point $\mathbf{x}$. After we generate the adversarial data $\mathbf{x}'$, it requires an additional 1 forward and backward pass to optimize the model. We will need 11 forward and backward pass per iteration to conduct PGD-10 adversarial training. When introducing ADR to rectify the targets, an additional forward is needed for the EMA model. We need 12 forward and 11 backward pass in total.

We provide time per epoch for adversarial training in Table-5.6. The experiment is reported by running each algorithm on a single NVIDIA RTX A6000 GPU with batch size 128. From the table, we can infer that the computation cost introduced by ADR is relatively small (1.83% on ResNet-18 4.45% on WRN-34-10 on average) while the overhead brought by AWP is a lot higher (12.6% on ResNet-18, 12.8% on WRN-34-10 on average). We can achieve similar robustness improvement to AWP with ADR with less computation cost required.

doi:10.6342/NTU202301129

5.11 Variance across reruns

Table-5.7 presents the results of five repeated runs for AT and the proposed defenses AT+ADR on CIFAR-10 with ResNet-18. Our findings indicate that the proposed ADR approach consistently outperforms AT as evidenced by a higher mean (54.91% for AT+ADR compared to 52.6 for AT) and lower standard deviation in robust accuracy (0.128% for AT+ADR compared to 0.182% for AT). This suggests that ADR provides superior performance and greater stability in terms of robustness. While there is a larger standard deviation observed in standard accuracy when ADR is combined with AT, we consider the variance to be within an acceptable range (0.18%). Our results demonstrate that ADR yields stable performance that is independent of random seeds or other initialization states.

doi:10.6342/NTU202301129

Table 5.2: Test accuracy (%) of ADR combining with WA and AWP. The best results are marked in **bold**. The performance improvements and degradation are reported in red and blue numbers.

Method	ResNet-18		WRN-34-10	
	AutoAttack	Standard Acc.	AutoAttack	Standard Acc.
AT	48.81	82.52	52.12	85.15
+ WA	49.93 (+ 1.12)	**83.71** (+ 1.19)	53.97 (+ 1.85)	83.48 (- 1.67)
+ WA + AWP	50.72 (+ 1.91)	82.91 (+ 0.39)	55.01 (+ 2.89)	**87.42** (+ 2.27)
+ ADR	50.38 (+ 1.57)	82.41 (- 0.11)	53.25 (+ 1.13)	84.67 (- 0.48)
+ WA + ADR	50.85 (+ 2.04)	82.89 (+ 0.37)	54.10 (+ 1.98)	82.93 (- 2.22)
+ WA + AWP + ADR	**51.18** (+ 2.37)	83.26 (+ 0.74)	**55.26** (+ 3.14)	86.11 (+ 0.96)

(a) CIFAR-10

Method	ResNet-18		WRN-34-10	
	AutoAttack	Standard Acc.	AutoAttack	Standard Acc.
AT	24.95	55.81	28.45	61.12
+ WA	26.27 (+ 1.32)	53.54 (- 2.27)	30.22 (+ 1.77)	60.04 (- 1.08)
+ WA + AWP	27.36 (+ 2.41)	**59.06** (+ 3.25)	30.73 (+ 2.28)	**63.11** (+ 1.99)
+ ADR	26.87 (+ 1.92)	56.10 (+ 0.29)	29.35 (+ 0.90)	59.76 (- 1.36)
+ WA + ADR	27.51 (+ 2.56)	58.30 (+ 2.49)	30.46 (+ 2.01)	57.42 (- 3.70)
+ WA + AWP + ADR	**28.50** (+ 3.55)	57.36 (+ 1.55)	**31.60** (+ 3.15)	62.21 (+ 1.09)

(b) CIFAR-100

Method	ResNet-18		WRN-34-10	
	AutoAttack	Standard Acc.	AutoAttack	Standard Acc.
AT	18.06	45.87	20.76	49.11
+ WA	19.30 (+ 1.24)	**49.10** (+ 3.23)	22.77 (+ 2.01)	53.21 (+ 4.10)
+ WA + AWP	19.58 (+ 1.52)	48.61 (+ 2.74)	23.22 (+ 2.46)	**53.35** (+ 4.42)
+ ADR	19.46 (+ 1.40)	48.19 (+ 2.32)	21.85 (+ 1.09)	51.52 (+ 2.41)
+ WA + ADR	**20.23** (+ 2.17)	48.55 (+ 2.68)	23.01 (+ 2.25)	51.03 (+ 1.92)
+ WA + AWP + ADR	20.12 (+ 2.06)	48.27 (+ 2.40)	**23.33** (+ 2.57)	51.44 (+ 2.33)

(c) TinyImageNet-200

doi:10.6342/NTU202301129

Table 5.3: Comparison of ADR with other related works on CIFAR-100.

Architecture	Method	Extra Data	AutoAttack.	Standard Acc.
ResNet18	Zhang et al. [65]	-	26.03	58.17
	Dong et al. [16]	-	26.30	56.45
	Dong et al. [15]	-	26.36	58.80
	Addepalli et al. [1]	-	27.62	66.69
	AT+ADR (Ours)	-	**28.50**	57.36
Preact-ResNet18	Rebuffi et al. [46]	DDPM	28.50	56.87
	Rade et al. [45]	DDPM	28.88	61.50
	AT+ADR (Ours)	DDPM	**29.59**	57.88
WRN-34-10	Chen and Lee [7]	-	30.59	64.07
	Jia et al. [30]	-	30.77	64.89
	Sehwag et al. [50]	DDPM	31.15	65.93
	Cui et al. [13]	-	31.20	62.99
	Addepalli et al. [1]	-	31.30	68.74
	AT+ADR (Ours)	-	31.60	62.21
	AT+ADR (Ours)	DDPM	**32.19**	59.60

Table 5.4: Test accuracy (%) of ADR + TRADES combining with WA and AWP on ResNet-18. The best results are marked in **bold**. Robust Accuracy (RA) is evaluated with AutoAttack and Standard Accuracy (SA) refers to the accuracy of normal data.

Method	CIFAR-10		CIFAR-100		TinyImageNet-200	
	RA	SA	RA	SA	RA	SA
TRADES	50.10	82.95	24.71	56.37	17.35	48.49
+ WA	51.10	81.77	25.61	57.93	17.69	49.51
+ WA + AWP	51.25	81.48	26.54	58.40	17.66	49.21
+ ADR	51.02	**83.40**	26.42	56.54	19.17	51.82
+ WA + ADR	**51.28**	82.69	27.11	**58.58**	19.17	**51.99**
+ WA + AWP + ADR	50.59	80.84	**27.63**	57.16	**19.48**	51.38

Table 5.5: Comparison of ADR with other related works on TinyImageNet-200. The best result for each architecture is marked in **bold**

Architecture	Method	AutoAttack.	Standard Acc.
ResNet-18	AT	18.06	45.87
	Rade and Moosavi-Dezfooli [45]	18.14	**52.60**
	Dong et al. [15]	18.29	47.46
	AT+ADR(Ours)	**20.23**	48.55
WRN-28-10	Rebuffi et al. [47]	21.83	**53.27**
WRN-34-10	AT	20.76	49.11
	AT+ADR(Ours)	**23.33**	51.44

doi:10.6342/NTU202301129

Table 5.6: Computational cost analysis for ADR combining with AT techniques on CIFAR-10, CIFAR-100 and TinyImageNet-200 with ResNet-18 and WRN-34-10. We report the time required per epoch in seconds tested on a single NVIDIA RTX A6000 GPU.

Dataset	Architecture	Method	Time/epoch (sec)
CIFAR-10	ResNet-18	AT	81
		+ADR	83
		+AWP	92
		+AWP+ADR	93
	WRN-34-10	AT	575
		+ADR	610
		+AWP	650
		+AWP+ADR	695
CIFAR-100	ResNet-18	AT	81
		+ADR	83
		+AWP	91
		+AWP+ADR	93
	WRN-34-10	AT	586
		+ADR	598
		+AWP	661
		+AWP+ADR	672
TinyImageNet-200	ResNet-18	-	584
		+ADR	593
		+AWP	654
		+AWP+ADR	662
	WRN-34-10	AT	4356
		+ADR	4594
		+AWP	4904
		+AWP+ADR	5127

Table 5.7: Variation in performance (%) of AT and ADR on ResNet-18 across 5 reruns on CIFAR-10. The robust accuracy is evaluated under the PGD-100 attack.

	AT		AT+ADR	
	Robust Acc.	Standard Acc.	Robust Acc.	Standard Acc.
Run-1	52.80	82.52	55.13	82.41
Run-2	52.28	82.41	54.88	82.21
Run-3	52.74	82.39	54.92	82.18
Run-4	52.55	82.30	54.91	82.68
Run-5	52.63	82.31	54.73	82.31
Average	52.60	82.38	54.91	82.36
Standard Deviation	0.182	0.079	0.128	0.180

doi:10.6342/NTU202301129

Chapter 6 Conclusion

6.1 Limitations

In this work, we proposed ADR that employs a self-distillate EMA model to generate a finely calibrated soft label to enhance the robustness of models against adversarial attacks. However, we observe that the optimal parameters for each dataset vary. Thus, selecting appropriate parameters that suit the current training state is crucial to ensure optimal performance. It is also noteworthy that while ADR demonstrates its efficacy in improving the performance of TRADES, the extent of improvement saturates when combined with other adversarial training techniques (WA, AWP) on fewer class datasets e.g. CIFAR-10. This outcome could be attributed to the fact that TRADES already promotes attacking and learning the data with soft targets generated by the trained model itself, and these additional techniques further smooth the model weight space. Thus, when the target class is fewer, the improvement provided by ADR, which also emphasizes a smooth objective, becomes indistinguishable when all techniques are employed simultaneously.

doi:10.6342/NTU202301129

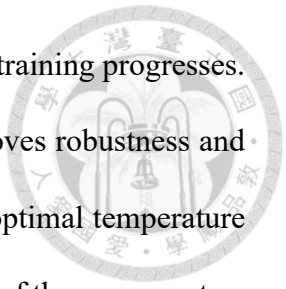
6.2 Border impacts

Adversarial training has the potential to improve the security, and reliability of machine learning systems. In practical settings, adversarial attacks can be employed by malevolent actors in an attempt to deceive machine learning systems. This phenomenon can engender grave consequences for domains such as autonomous driving vehicles and facial recognition. To enhance the security and reliability of machine learning systems, adversarial training can be employed to produce more dependable models. However, it is worth noting that robust models can also be exploited by ill-intentioned users. In the context of the CAPTCHA, adversarial perturbations can be added to images to distinguish between humans and robots since the robots are expected to be fooled by the adversarial data. If robots are able to attain robust models, they would not be susceptible to adversarial examples and could supply accurate answers. The advancement of model robustness may inspire people to formulate a better strategy to differentiate between humans and robots.

6.3 Conclusion

In this paper, we characterize the key properties that distinguish robust and non-robust model output. We find that a robust model should exhibit good calibration and maintain output consistency on clean data and its adversarial counterpart. Based on this observation, we propose a data-driven labeling scheme ADR, which is designed to account for subtle distribution shifts when inputs are slightly modified, without the need for pre-trained resources or extensive computation overhead. To achieve this, we utilize the self-distillation EMA model to create a robust teacher that provides labeling guid-

doi:10.6342/NTU202301129



ance for the student, with the increasing trust placed in the teacher as training progresses. Comprehensive experiments demonstrate that ADR effectively improves robustness and alleviates robust overfitting. However, we note that the algorithm's optimal temperature and interpolation ratio depends on the dataset, and improper selection of these parameters can limit performance improvements. The automatic determination of optimal parameters in training will be an important future research direction that can further boost the robustness.

doi:10.6342/NTU202301129

doi:10.6342/NTU202301129

References

[1] S. Addepalli, S. Jain, and R. V. Babu. Efficient and effective augmentation strategy for adversarial training. CoRR, abs/2210.15318, 2022.

[2] M. Andriushchenko, F. Croce, N. Flammarion, and M. Hein. Square attack: A query-efficient black-box adversarial attack via random search. In A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, editors, Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXIII, volume 12368 of Lecture Notes in Computer Science, pages 484–501. Springer, 2020.

[3] A. Athalye, N. Carlini, and D. A. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In J. G. Dy and A. Krause, editors, Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018, volume 80 of Proceedings of Machine Learning Research, pages 274–283. PMLR, 2018.

[4] N. Carlini, A. Athalye, N. Papernot, W. Brendel, J. Rauber, D. Tsipras, I. Goodfellow, A. Madry, and A. Kurakin. On evaluating adversarial robustness. arXiv preprint arXiv:1902.06705, 2019.

[5] Y. Carmon, A. Raghunathan, L. Schmidt, J. C. Duchi, and P. Liang. Unlabeled

doi:10.6342/NTU202301129

data improves adversarial robustness. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, editors, Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, pages 11190–11201, 2019.

[6] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021, pages 9630–9640. IEEE, 2021.

[7] E. Chen and C. Lee. LTD: low temperature distillation for robust adversarial training. CoRR, abs/2111.02331, 2021.

[8] T. Chen, Z. Zhang, S. Liu, S. Chang, and Z. Wang. Robust overfitting may be mitigated by properly learned smoothening. In 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net, 2021.

[9] J. M. Cohen, E. Rosenfeld, and J. Z. Kolter. Certified adversarial robustness via randomized smoothing. In K. Chaudhuri and R. Salakhutdinov, editors, Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA, volume 97 of Proceedings of Machine Learning Research, pages 1310–1320. PMLR, 2019.

[10] F. Croce and M. Hein. Minimally distorted adversarial examples with a fast adaptive boundary attack. In Proceedings of the 37th International Conference on Machine

doi:10.6342/NTU202301129

Learning, ICML 2020, 13-18 July 2020, Virtual Event, volume 119 of Proceedings of Machine Learning Research, pages 2196–2205. PMLR, 2020.

[11] F. Croce and M. Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event, volume 119 of Proceedings of Machine Learning Research, pages 2206–2216. PMLR, 2020.

[12] F. Croce, M. Andriushchenko, V. Sehwag, E. Debenedetti, N. Flammarion, M. Chiang, P. Mittal, and M. Hein. Robustbench: a standardized adversarial robustness benchmark. In J. Vanschoren and S. Yeung, editors, Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual, 2021.

[13] J. Cui, S. Liu, L. Wang, and J. Jia. Learnable boundary guided adversarial training. In 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021, pages 15701–15710. IEEE, 2021.

[14] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA, pages 248–255. IEEE Computer Society, 2009.

[15] C. Dong, L. Liu, and J. Shang. Label noise in adversarial training: A novel perspective to study robust overfitting. In Advances in Neural Information Processing Systems, 2022.

doi:10.6342/NTU202301129

[16] Y. Dong, K. Xu, X. Yang, T. Pang, Z. Deng, H. Su, and J. Zhu. Exploring memorization in adversarial training. In The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net, 2022.

[17] L. Engstrom, A. Ilyas, and A. Athalye. Evaluating and understanding the robustness of adversarial logit pairing. CoRR, abs/1807.10272, 2018.

[18] T. Garipov, P. Izmailov, D. Podoprikhin, D. P. Vetrov, and A. G. Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. In S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada, pages 8803–8812, 2018.

[19] M. Goldblum, L. Fowl, S. Feizi, and T. Goldstein. Adversarially robust distillation. In The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, pages 3996–4003. AAAI Press, 2020.

[20] S. Gowal, C. Qin, J. Uesato, T. A. Mann, and P. Kohli. Uncovering the limits of adversarial training against norm-bounded adversarial examples. CoRR, abs/2010.03593, 2020.

[21] J. Grabinski, P. Gavrikov, J. Keuper, and M. Keuper. Robust models are less overconfident. CoRR, abs/2210.05938, 2022.

doi:10.6342/NTU202301129

[22] J. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. Á. Pires, Z. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko. Bootstrap your own latent - A new approach to self-supervised learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020.

[23] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, pages 770–778. IEEE Computer Society, 2016.

[24] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV, volume 9908 of Lecture Notes in Computer Science, pages 630–645. Springer, 2016.

[25] M. Hein and M. Andriushchenko. Formal guarantees on the robustness of a classifier against adversarial manipulation. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, editors, Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pages 2266–2276, 2017.

[26] D. Hendrycks and K. Gimpel. Bridging nonlinearities and stochastic regularizers

doi:10.6342/NTU202301129

with gaussian error linear units. CoRR, abs/1606.08415, 2016.

[27] G. E. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. CoRR, abs/1503.02531, 2015.

[28] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020.

[29] P. Izmailov, D. Podoprikhin, T. Garipov, D. P. Vetrov, and A. G. Wilson. Averaging weights leads to wider optima and better generalization. In A. Globerson and R. Silva, editors, Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI 2018, Monterey, California, USA, August 6-10, 2018, pages 876–885. AUAI Press, 2018.

[30] X. Jia, Y. Zhang, B. Wu, K. Ma, J. Wang, and X. Cao. LAS-AT: adversarial training with learnable attack strategy. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022, pages 13388–13398. IEEE, 2022.

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

[32] A. Kurakin, I. J. Goodfellow, and S. Bengio. Adversarial machine learning at scale. In 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net, 2017.

doi:10.6342/NTU202301129

[33] B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, editors, Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pages 6402–6413, 2017.

[34] Y. Le and X. Yang. Tiny imagenet visual recognition challenge. CS 231N, 7(7):3, 2015.

[35] H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein. Visualizing the loss landscape of neural nets. In S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada, pages 6391–6401, 2018.

[36] L. Li and M. W. Spratling. Data augmentation alone can improve adversarial training. CoRR, abs/2301.09879, 2023.

[37] F. Liao, M. Liang, Y. Dong, T. Pang, X. Hu, and J. Zhu. Defense against adversarial attacks using high-level representation guided denoiser. In 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, pages 1778–1787. Computer Vision Foundation / IEEE Computer Society, 2018.

[38] I. Loshchilov and F. Hutter. SGDR: stochastic gradient descent with warm restarts.

doi:10.6342/NTU202301129

In 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net, 2017.

[39] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net, 2018.

[40] R. Müller, S. Kornblith, and G. E. Hinton. When does label smoothing help? In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, editors, Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, pages 4696–4705, 2019.

[41] D. Paleka and A. Sanyal. A law of adversarial risk, interpolation, and label noise. CoRR, abs/2207.03933, 2022.

[42] T. Pang, K. Xu, C. Du, N. Chen, and J. Zhu. Improving adversarial robustness via promoting ensemble diversity. In K. Chaudhuri and R. Salakhutdinov, editors, Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA, volume 97 of Proceedings of Machine Learning Research, pages 4970–4979. PMLR, 2019.

[43] T. Pang, X. Yang, Y. Dong, H. Su, and J. Zhu. Bag of tricks for adversarial training. In 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net, 2021.

[44] Y. Qin, X. Wang, A. Beutel, and E. H. Chi. Improving calibration through the relationship with adversarial robustness. In M. Ranzato, A. Beygelzimer, Y. N. Dauphin,

doi:10.6342/NTU202301129

P. Liang, and J. W. Vaughan, editors, Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual, pages 14358–14369, 2021.

[45] R. Rade and S. Moosavi-Dezfooli. Reducing excessive margin to achieve a better accuracy vs. robustness trade-off. In The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net, 2022.

[46] S. Rebuffi, S. Gowal, D. A. Calian, F. Stimberg, O. Wiles, and T. A. Mann. Fixing data augmentation to improve adversarial robustness. CoRR, abs/2103.01946, 2021.

[47] S.-A. Rebuffi, S. Gowal, D. A. Calian, F. Stimberg, O. Wiles, and T. Mann. Data augmentation can improve robustness. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, Advances in Neural Information Processing Systems, 2021.

[48] L. Rice, E. Wong, and J. Z. Kolter. Overfitting in adversarially robust deep learning. In Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event, volume 119 of Proceedings of Machine Learning Research, pages 8093–8104. PMLR, 2020.

[49] L. Schmidt, S. Santurkar, D. Tsipras, K. Talwar, and A. Madry. Adversarially robust generalization requires more data. In S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, Advances in Neural Information Processing Systems, pages 5019–5031, 2018.

[50] V. Sehwag, S. Mahloujifar, T. Handina, S. Dai, C. Xiang, M. Chiang, and P. Mittal. Robust learning meets generative models: Can proxy distributions improve adversar-

doi:10.6342/NTU202301129

ial robustness? In The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net, 2022.

[51] L. N. Smith and N. Topin. Super-convergence: Very fast training of residual networks using large learning rates. CoRR, abs/1708.07120, 2017.

[52] J. Snoek, Y. Ovadia, E. Fertig, B. Lakshminarayanan, S. Nowozin, D. Sculley, J. V. Dillon, J. Ren, and Z. Nado. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, editors, Advances in Neural Information Processing Systems, pages 13969–13980, 2019.

[53] D. Stutz, M. Hein, and B. Schiele. Confidence-calibrated adversarial training: Generalizing to unseen attacks. In Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event, volume 119 of Proceedings of Machine Learning Research, pages 9155–9166. PMLR, 2020.

[54] D. Stutz, M. Hein, and B. Schiele. Relating adversarially robust generalization to flat minima. In 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021, pages 7787–7797. IEEE, 2021.

[55] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In Y. Bengio and Y. LeCun, editors, 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings, 2014.

[56] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In 2016 IEEE Conference on Computer Vision and

doi:10.6342/NTU202301129

Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, pages 2818–2826. IEEE Computer Society, 2016.

[57] A. Tarvainen and H. Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings. OpenReview.net, 2017.

[58] A. Torralba, R. Fergus, and W. T. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. IEEE Trans. Pattern Anal. Mach. Intell., 30(11):1958–1970, 2008.

[59] J. Uesato, B. O'Donoghue, P. Kohli, and A. van den Oord. Adversarial risk and the dangers of evaluating against weak attacks. In J. G. Dy and A. Krause, editors, Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018, volume 80 of Proceedings of Machine Learning Research, pages 5032–5041. PMLR, 2018.

[60] E. Wong and J. Z. Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In J. G. Dy and A. Krause, editors, Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018, volume 80 of Proceedings of Machine Learning Research, pages 5283–5292. PMLR, 2018.

[61] D. Wu, S. Xia, and Y. Wang. Adversarial weight perturbation helps robust generalization. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, Advances in Neural Information Processing Systems 33: Annual Conference

doi:10.6342/NTU202301129

on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020.

[62] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, and J. Choe. Cutmix: Regularization strategy to train strong classifiers with localizable features. In 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019, pages 6022–6031. IEEE, 2019.

[63] S. Zagoruyko and N. Komodakis. Wide residual networks. In R. C. Wilson, E. R. Hancock, and W. A. P. Smith, editors, Proceedings of the British Machine Vision Conference 2016, BMVC 2016, York, UK, September 19-22, 2016. BMVA Press, 2016.

[64] H. Zhang, Y. Yu, J. Jiao, E. P. Xing, L. E. Ghaoui, and M. I. Jordan. Theoretically principled trade-off between robustness and accuracy. In K. Chaudhuri and R. Salakhutdinov, editors, International conference on machine learning, pages 7472–7482. PMLR, 2019.

[65] S. Zhang, H. Gao, T. Zhang, Y. Zhou, and Z. Wu. Alleviating robust overfitting of adversarial training with consistency regularization. CoRR, abs/2205.11744, 2022.

[66] S. Zhao, J. Yu, Z. Sun, B. Zhang, and X. Wei. Enhanced accuracy and robustness via multi-teacher adversarial distillation. In S. Avidan, G. J. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, editors, Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part IV, volume 13664 of Lecture Notes in Computer Science, pages 585–602. Springer, 2022.

[67] J. Zhu, J. Yao, B. Han, J. Zhang, T. Liu, G. Niu, J. Zhou, J. Xu, and H. Yang. Reliable adversarial distillation with unreliable teachers. In The Tenth International

doi:10.6342/NTU202301129

Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net, 2022.

[68] B. Zi, S. Zhao, X. Ma, and Y. Jiang. Revisiting adversarial robustness distillation: Robust soft labels make student better. In 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021, pages 16423–16432. IEEE, 2021.

doi:10.6342/NTU202301129