Towards Better Robust Generalization with Shift Consistency Regularization

Shufei Zhang [*1 2] **Zhuang Qian** [*1 2] **Kaizhu Huang** [1] **Qiufeng Wang** [1] **Rui Zhang** [3] **Xinping Yi** [2]

Abstract

While adversarial training becomes one of the most promising defending approaches against adversarial attacks for deep neural networks, the conventional wisdom through robust optimization may usually not guarantee good generalization for robustness. Concerning with robust generalization over unseen adversarial data, this paper investigates adversarial training from a novel perspective of shift consistency in latent space. We argue that the poor robust generalization of adversarial training is owing to the significantly dispersed latent representations generated by training and test adversarial data, as the adversarial perturbations push the latent features of natural examples in the same class towards diverse directions. This is underpinned by the theoretical analysis of the robust generalization gap, which is upper-bounded by the standard one over the natural data and a term of feature inconsistent shift caused by adversarial perturbation – a measure of latent dispersion. Towards better robust generalization, we propose a new regularization method – shift consistency regularization (SCR) – to steer the same-class latent features of both natural and adversarial data into a common direction during adversarial training. The effectiveness of SCR in adversarial training is evaluated through extensive experiments over different datasets, such as CIFAR-10, CIFAR-100, and SVHN, against several competitive methods.

1. Introduction

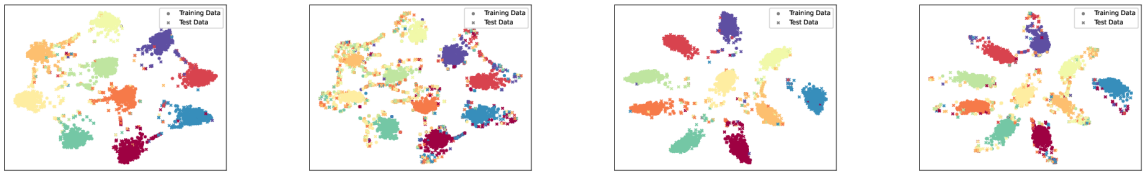
Recent years have witnessed the remarkable success of deep neural network (DNN) models spanning a wide range of applications including image classification (LeCun et al., 2015; He et al., 2016; Miyato et al., 2017; Zagoruyko & Komodakis, 2016), image generation (Nowozin et al., 2016; Salvaris et al., 2018; Arjovsky et al., 2017), object detection (Zhao et al., 2019) and natural language processing (Otter et al., 2020). Despite the impressive performance boosting over various learning tasks, DNNs are demonstrated to be strikingly vulnerable to certain well-crafted adversarial perturbations (Carlini & Wagner, 2018; Eykholt et al., 2018; Fischer et al., 2017; Lyu et al., 2015). While such perturbations are imperceptible to human, they can easily mislead the prediction of DNNs with high confidence. Along with the increasing deployment of DNN models in safety-critical scenarios, it becomes extremely crucial to ensure model robustness against potential adversarial attacks.

There has been a fast-growing body of works in the literature spurred by the arms race between adversarial attacks and defenses. The newly emerging attacks (Croce & Hein, 2020; Kurakin et al., 2016; Carlini & Wagner, 2017; Madry et al., 2017; Moosavi-Dezfooli et al., 2016) have soon been defended by dedicated defense techniques (Wang & Zhang, 2019; Madry et al., 2017; Kannan et al., 2018; Gu & Rigazio, 2014), which are then broken again by more powerful attacks. Among many defending techniques, adversarial training appears one of the most effective and promising approaches, by augmenting the training dataset with adversarial examples to train robust DNNs (Goodfellow et al., 2014; Lyu et al., 2015), or robustifying DNN model training process against the worst-case adversarial perturbation (Madry et al., 2017). Recent advances in adversarial training include 1) speeding up adversarial training, e.g., (Shafahi et al., 2019; Zhang et al., 2019a; Zhu et al., 2019; Wong et al., 2020); 2) considering the inter-sample relationship when generating adversarial perturbations, e.g., (Zhang & Wang, 2019; Miyato et al., 2017); and many others.

Albeit promising from the viewpoint of robustness, these conventional wisdom may usually not guarantee good generalization for robustness, i.e. the generalization over unseen adversarial data. In particular, while the above methods achieve impressive robustness performance, there still exists a big robust generalization gap between training and test sets. Moreover, it appears that such robustness generalization on more complicated datasets could be even difficult to be attained (Schmidt et al., 2018; Zhai et al., 2019). Notably, Schmidt et al. (2018) have shown that the sample

[*]Equal contribution [1]School of Advanced Technology, Xi'an Jiaotong-Liverpool University, China. [2]School of Electrical Engineering, Electronics and Computer Science, University of Liverpool, UK. [3]School of Science, Xi'an Jiaotong-Liverpool University, China. Correspondence to: Kaizhu Huang <kaizhu.huang@xjtlu.edu.cn;kaser.huang@gmail.com>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).



(a) FS w/ clean data (b) FS w/ adversarial data (c) FS-SCR w/ clean data (d) FS-SCR w/ adversarial data

Figure 1. Visualization via t-distributed stochastic neighbor embedding (TSNE) of latent feature without and with shift consistency regularization (SCR) applied during feature scattering (FS) based adversarial training. The latent features of FS generated from (a) clean data; (b) adversarial data perturbed by PGD20; (c) clean data with SCR applied; and (d) adversarial data perturbed by PGD20 with SCR applied. The training and test data samples are marked as dots and crosses, respectively. While the latent distributions of clean training and test data are consistently centralized, those with adversarial perturbation exhibit a clearly inconsistent shift that leads to poor robust generalization, as shown in (a) and (b). When SCR is applied together with FS-based adversarial training, the feature shift inconsistency is substantially alleviated, as shown in (c) and (d), and therefore better robust generalization is attained.

complexity of robust learning can be significantly larger than that of "standard" learning and consequently it is much harder to achieve the robust generalization than "standard" generalization. As such, the question then arises as to how to improve the generalization over unseen adversarial data.

Although progress has been made to investigate the interplay between robustness and generalization, e.g., (Farnia et al., 2018; Tsipras et al., 2018; Zhang et al., 2019b; Yin et al., 2019; Raghunathan et al., 2019; Yang et al., 2020; Wu et al., 2020; Roth et al., 2020), the fundamental understanding still has a long way to go. It is unclear which way is most suitable to pursue: (1) starting with robustness with adversarial training, followed by improving generalization performance by some regularization techniques as how standard training has succeeded; or (2) starting with standard training with cutting-edge techniques for generalization, followed by the enhancement of defending techniques, e.g., robust regularization techniques. Even worse, it is still not understood why the robust generalization is harder to achieve than the "standard" one.

In this paper, we make a first step to understand why robust generalization is more challenging from a novel perspective of shift consistency of latent features. According to Figures 1(a) and 1(b), we observe that the distributions of latent features for the training and test data, with Feature Scattering (FS) adversarial training and adversarial perturbation by PGD attacks, experience a clearly inconsistent shift as in 1(b), in sharp contrast to those obtained for clean training and test data in 1(a). Inspired by such observations, we argue that the poor robust generalization performance is probably attributed to such latent feature inconsistent shifts. To verify such a hypothesis, we investigate the robust generalization gap through the connection between algorithmic robustness and generalization ability (Xu & Mannor, 2012). Specifically, we prove that the robust generalization gap is upper-bounded by both the "standard" generalization gap and a measure of the inconsistent shifts of latent features caused by adversarial perturbations. As such, in order to tighten such robust generalization gap, we can alleviate the inconsistent shifts of latent features, given that the DNN training can already achieve a reasonably small "standard" generalization gap. To this end, we propose a novel regularization method – shift consistency regularization (SCR) – to boost the robust generalization performance for adversarial training. Our contributions are summarized as follows:

- We propose to study the robust generalization from the novel perspective of shift consistency of latent features, where the latent distributions of adversarial training and test data exhibit certain dispersion due to adversarial perturbation. This is supported by our theoretical analysis on the robust generalization bound, which can be disentangled into the "standard" generalization gap and a measure of inconsistent shifts of latent features caused by adversarial perturbations.

- Inspired by the measure of inconsistent shifts, we propose a simple yet effective shift consistency regularization (SCR) technique to alleviate the dispersion of latent features. In doing so, the latent distributions of both adversarial and clean data in the same class are steered into the consistent directions, as shown in Figure 1(c)(d) and Figure 3(c)(d), therefore leading to better robust generalization performance.

- Extensive experiments have been conducted to demonstrate the effectiveness of our proposed SCR method over a variety of datasets, e.g., CIFAR-10, CIFAR-100, and SVHN. It shows our method could be able to improve the robustness over the recent state-of-the-art methods substantially. Moreover, it is demonstrated that improved robust generalization is due to the tighter upper bound of robust generalization gap.

2. Background and Related Work

Adversarial training is a family of approaches to improve the model robustness (Goodfellow et al., 2014; Lyu et al., 2015; Madry et al., 2017). Owing to its impressive performance on defending against adversarial attacks, it has drawn much attention in the recent years, e.g., (Zhang et al., 2018; Wang & Zhang, 2019; Zhang & Wang, 2019; Zhang et al., 2020; Mao et al., 2019; Shafahi et al., 2019; Zhang et al., 2019a; Zhu et al., 2019; Wong et al., 2020). The common idea of these methods is to train DNNs with perturbed examples instead of clean ones with the correct labels, so that the trained model could be robust to the adversarial perturbation. In what follows, we introduce adversarial training via robust optimization, followed by the robust generalization.

2.1. Conventional Adversarial Training

The main idea of the conventional adversarial training is to train the DNNs with the adversarial examples induced by the worst predictions. This can be formulated as a robust optimization problem as follows.

$$\min_{\theta} \quad \{\mathbb{E}_{(x,y)\sim\mathcal{D}}[\max_{x'\in S_x} L(x', y; \theta)]\}, \tag{1}$$

where $x \in \mathbb{R}^d$ and $y \in \mathbb{N}$ denote the clean data samples and the corresponding labels drawn from the dataset $\mathcal{D}$ respectively; $L(\cdot)$ is the loss function of the DNN with the model parameter $\theta \in \mathbb{R}^m$; and $x' \in \mathbb{R}^d$ is the perturbation of x within a feasible region $S_x \triangleq \{z : z \in B(x, \epsilon) \cap [-1.0, 1.0]^d\}$ with $B(z, \epsilon) \triangleq \{z : \|x - z\|_\infty \leq \epsilon\}$ being the ℓ_∞-ball at center x with radius ϵ. By defining $f_\theta(\cdot)$ as the mapping function from the input layer to the last latent layer, with model parameters θ, we can also rewrite the loss function of the DNN as $l(f_\theta(x), y)$ where $l(\cdot)$ denotes the loss function calculated from the last hidden layer of the DNN, e.g. the cross entropy loss as typically used in DNN.

To solve the above minimax optimization problem, the commonly adopted approach (e.g., (Madry et al., 2017)) is to iteratively update between the outer minimization via SGD training and the inner maximization via adversarial attacks (e.g., PGD, FGSM) until the convergence.

The wisdom behind this line of research is that, it is expected to train a robust model if the potential attacks through the adversarial perturbation are identified and then eliminated during the training process. Nevertheless, the robust generalization is not considered.

2.2. Adversarial Training with Feature Scattering

To exploit the structure of data manifold, recent works have stated to consider the inter-sample relationship during adversarial training, e.g., (Sinha et al., 2017; Miyato et al., 2017; Zhang & Wang, 2019) to name just a few. Among these recent advances, Feature Scattering (FS) based adversarial training (Zhang & Wang, 2019) is one of the most promising ones, achieving impressive success in defending against various adversarial attacks. Different from the conventional adversarial training methods, FS generates the adversarial examples for training with the inter-sample relationships considered. Consequently, some other perturbed examples (not only the worst ones) that are crucial for learning robust models are also implicitly considered. The main objective function of FS can be formulated as:

$$\begin{aligned} \min_{\theta} \quad & \frac{1}{N}\sum_{i=1}^{N} L(x_i^{adv}, y_i; \theta) \\ \text{s.t.} \quad & \nu^* = \sum_{i=1}^{N} v_i \delta_{x_i^{adv}} = \arg\max_{\nu\in S_\mu} D_{ot}(\nu, \mu) \end{aligned} \tag{2}$$

where $\mu = \sum_{i=1}^{N} u_i \delta_{x_i}$ and $\nu = \sum_{i=1}^{N} v_i \delta_{x_i^{adv}}$ are two discrete distributions for natural examples $\{x_i\}_{i=0}^N$ and perturbed examples $\{x_i'\}_{i=0}^N$ respectively and $V = \{v_i\}_{i=1}^N$ and $U = \{u_i\}_{i=1}^N$ are corresponding weights. Here, $v_i = u_i = 1/N$. $S_\mu = \{\sum_i v_i \delta_{z_i}, |z_i \in B(x_i, \epsilon) \cap [0, 255]^d\}$ denotes the feasible region. $D_{ot} = \min_{T\in\prod(U,V)} \sum_{i=1}^N \sum_{j=1}^N T_{ij} c(x_i, x_j')$ is the optimal transport (OT) distance where $\prod(U, V) = \{T \in \mathbb{R}_+^{N\times N} | T1_N = U, T^\top 1_N = V\}$ and 1_N denotes all-one vector. Here, the cost function is defined as $c(x_i, x_j') = 1 - \frac{f_\theta(x_i)^\top f_\theta(x_j')}{\|f_\theta(x_i)\|_2 \|f_\theta(x_j')\|_2}$ to measure the feature similarity.

Intuitively, the adversarial examples are generated to make their latent representations as distinguishable from that of clean ones as possible. Since the adversarial perturbations are crafted in an unsupervised fashion, there is no high correlation between the perturbations and the decision boundary so that the potential label leaking problem can be prevented (Zhang & Wang, 2019).

2.3. Robust Generalization

Robust generalization describes how well the robust models perform on unseen adversarial data. Different from standard generalization (on clean data), learning a model with good robust generalization is particularly difficult because of the requirement of significantly higher data sample complexity (Schmidt et al., 2018; Zhai et al., 2019). There is an increasing attention being put on the relation between robustness and generalization. For instance, Zhang et al. (2019b) proposed to decompose the robust error due to adversarial examples into the natural classification error and the boundary error, shedding light on the trade-off between the robustness and the accuracy. Yang et al. (2020) argued that both accuracy and robustness are achievable if the local Lipschitzness is maintained.

On the other hand, there is an increasing number of works resorting to the regularization techniques to promote robust

generalization for adversarial training, hoping to repeat their success in the standard training. Remarkably, Yin et al. (2019) demonstrated that constraining ℓ_1 norm on weight matrices could improve robust generalization; Wu et al. (2020) attempted to perturb weight in addition to input sample to encourage generalization together with robustness; Roth et al. (2020) established a link between adversarial training and operator norm regularization; and many others.

Although intensive attention has been placed on the interplay between robustness and generalization for adversarial training, the fundamental understanding is still under exploration. A typical question is, why is the robust generalization of adversarial training is so hard to achieve? In this paper, we make progress towards the answer to this question by analyzing the robust generalization gap both theoretically and empirically. The theoretical analysis reveals that the challenge in achieving robust generalization is probably due to the feature inconsistency shifts of the adversarial data.

3. Robust Generalization Analysis

In this section, we will analyze the robust generalization from both the theoretical and empirical aspects. Specifically, we first provide the theoretical relationship between robust generalization and standard one and show that the shift inconsistency of latent features caused by adversarial perturbations contributes to the difficulty of robust generalization. Then, we validate the theoretical analysis through experiments. Finally, we visualize the latent features of clean and adversarial examples to show that the adversarial perturbations enlarge the difference between the latent features of training and test data.

3.1. Theoretical Analysis

The (standard) generalization is leveraged to measure how well the models perform on unseen data. The generalization error (gap) is defined as the difference between the expected loss over data distribution $(x, y) \sim (S, Y)$ and the empirical loss over the training data $(x_d, y_d) \in (S_d, Y_d)$ (Xu & Mannor, 2012; Bousquet & Elisseeff, 2002; Neyshabur et al., 2017), i.e.,

$$\mathrm{GE} \triangleq |l(f_\theta(S), Y) - \hat{l}(f_\theta(S_d), Y_d)| \tag{3}$$

$$\text{where} \quad l(f_\theta(S), Y) \triangleq \mathbb{E}_{(x,y)\sim(S,Y)}[l(f_\theta(x), y)]$$
$$\hat{l}(f_\theta(S_d), Y_d) \triangleq \frac{1}{|S_d|} \sum_{(x_d, y_d)\in S_d} l(f_\theta(x_d), y_d)$$

with S_d, Y_d being the training data and the corresponding labels, respectively, and S, Y being the underlying data and label distributions, respectively.

Similar to standard generalization, the robust generalization can be defined as the difference between the empirical loss on adversarial examples and the expected loss over their underlying distributions (Schmidt et al., 2018; Zhai et al., 2019; Wu et al., 2020) , i.e.,

$$\mathrm{RGE} \triangleq |l(f_\theta(S^{adv}), Y) - \hat{l}(f_\theta(S_d^{adv}), Y_d)| \tag{4}$$

where S_d^{adv} and S^{adv} are the set of adversarial examples for the training set and its underlying distribution. By these definitions, we can derive the theoretical relationship between the robust and standard generalization errors as Theorem 3.1 shows.

Theorem 3.1 *Given the training set $S_d = \{x_i\}_{i=1}^n$ that consists of n i.i.d samples drawn from a distribution S with K classes, and the set of corresponding adversarial examples $S_d^{adv} = \{x_i^{adv}\}_{i=1}^n$ drawn from the underlying distribution S^{adv}, if the loss function $l(\cdot)$ of DNN f_θ is k-Lipschitz, then for any $\delta > 0$, with the probability at least $1 - \delta$, we have*

$$\mathrm{RGE} \le \mathrm{GE} + \frac{k}{n} \sum_{i=1}^{K} \sum_{j\in N_i} \|d_\theta(x_j^{adv}) - \hat{d}_\theta(z, C_i)\|_2^2 + M\sqrt{\frac{2K\ln 2 + 2\ln\frac{1}{\delta}}{n}} \tag{5}$$

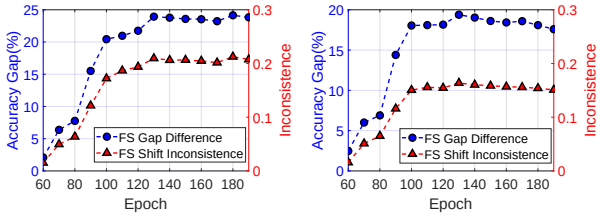
$$where \quad d_\theta(x^{adv}) = f_\theta(x^{adv}) - f_\theta(x) \tag{6}$$
$$\hat{d}_\theta(z, C_i) = \mathbb{E}[f_\theta(z^{adv}) - f_\theta(z) | z \in C_i] \tag{7}$$

with N_i being the set of index of training data for class i, C_i the set of i^{th} class data of the whole set and z is data sampled from C_i with corresponding adversarial example z^{adv}, M the upper bound of loss of the whole data manifold S.

Proof of Theorem 3.1 can be seen in the supplementary material. According to Theorem 3.1, the upper bound of the robust generalization gap (RGE) can be decomposed into three parts: standard generalization gap (GE), a term of the features shift inconsistency (SiC), and a constant part.

$$\mathrm{SiC} \triangleq \frac{k}{n} \sum_{i=1}^{K} \sum_{j\in N_i} \mathrm{SiC}(x_j^{adv}, z, C_i), \text{ where} \tag{8}$$
$$\mathrm{SiC}(x_j^{adv}, z, C_i) \triangleq \|d_\theta(x_j^{adv}) - \hat{d}_\theta(z, C_i)\|_2^2.$$

The shift inconsistency part is to measure the average difference between the training data feature shifts and the distributional data feature shifts over all classes. Theorem 3.1 indicates that feature shift inconsistency caused by adversarial perturbation enlarges the robust generalization gap and makes it harder to achieve.



(a) CW (b) PGD

Figure 2. The generalization gap difference (accuracy gap difference) and shift inconsistency at different training epochs. The gap difference and shift inconsistency are computed over (a) CW20 attack and (b) PGD20 attack. The changes of the shift inconsistency and gap difference are consistent.

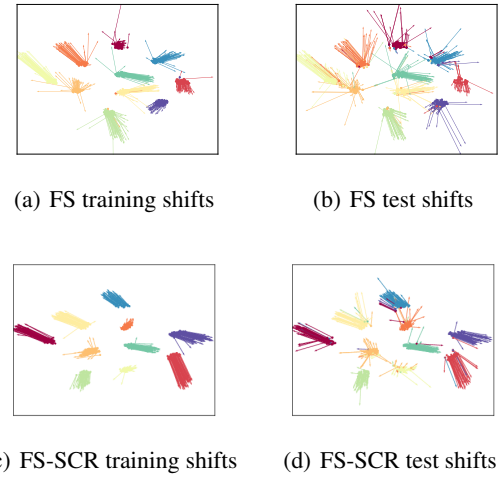
3.2. Empirical Analysis

In this subsection, we will first demonstrate Theorem 3.1 by showing how the difference between the robust and standard generalization gaps is varying with the level of feature shift inconsistency. Then we will visualize the TSNE embedding of the output features to show that the adversarial perturbations shift test data features away from training data features of the same class. Note that in this subsection all models are trained with FS on CIFAR-10.

Relationship between the generalization gap difference and feature shift inconsistency. For convenience, instead of computing the difference between the robust and standard generalization gaps, i.e., $\mathrm{RGE} - \mathrm{GE}$, we compute the accuracy gap difference between training and test datasets, given the fact that they are consistent to some degree (Xu & Mannor, 2012). In particular, we compute the accuracy gap difference using $|Acc(S_t^{adv}, Y_t) - Acc(S_d^{adv}, Y_d)| - |Acc(S_t, Y_t) - Acc(S_d, Y_d)|$, where S_t is the test set, Y_t denotes the set of the corresponding labels of test samples in S_t, S_t^{adv} is the set of adversarial examples of the test set S_t, and $Acc(S_t^{adv}, Y_t)$ is the accuracy of test adversarial data set S_t^{adv} with labels Y_t.

Further, we can substitute $\mathrm{SiC}(x_j^{adv}, z, C_i)$ with its empirical version $\mathrm{SiC}(x_j^{adv}, x, S_t) \triangleq \|d_\theta(x_j^{adv}) - \hat{d}_\theta(x, S_t)\|_2^2$ where $\hat{d}_\theta(x, S_t) = \mathbb{E}[f_\theta(x^{adv}) - f_\theta(x) | x \in S_t]$. Intuitively, we approximate the shift inconsistency with the shift difference between training and test sets. To validate the correctness of Theorem 3.1 and show whether the gap difference is caused by feature shift inconsistency, we plot the gap difference and shift consistency from epoch 60 to 190 in Figure 2.

As noted in Figure 2, for both the CW and PGD attacks, the changes of shift inconsistency and gap difference are consistent. Therefore, there exists a Lipschitz constant k and constant part $M\sqrt{\frac{2K\ln 2 + 2\ln\frac{1}{\delta}}{n}}$ that makes the upper bound tight in Theorem 3.1. Additionally, it indicates that



(a) FS training shifts (b) FS test shifts

(c) FS-SCR training shifts (d) FS-SCR test shifts

Figure 3. The feature shifts of the training and test data (PGD) .

the feature shift inconsistency can reflect the difference between the robust and standard generalization. Feature shift inconsistency might enlarge the robust generalization gap compared with standard one. To further inspect the reason why the robust generalization is harder to achieve, we visualize the output feature of DNNs in next subsection.

Visualization. We visualize the TSNE embedding of clean data and adversarial data features for both the training and test sets in Figure 1(a) 1(b) and Figure 6(a) 6(b).

Figure 1(a) and 6(a) are the TSNE embedding of clean data and it can be noted that for each class, the training data feature distribution is close to the test one. Therefore, the classifier learned with training set can still perform well on test set. Differently, as seen in Figure 1(d) and 6(d), some test adversarial data features are shifted away from the training adversarial data features of the same class. In other words, for the same class, the test adversarial data feature distribution becomes more different from the training one. Consequently, the classifier learned with the training adversarial data can not be guaranteed to perform well on adversarial examples of test set. This suggests why the robust generalization is harder to achieve.

Moreover, we plot the feature shifts caused by adversarial perturbations for both the training data and test data in Figure 3 where the feature shift for a data sample x is defined as $f_\theta(x^{adv}) - f_\theta(x)$. Comparing Figure 3(a) and 3(b), it can be noted that the test data feature shifts are different from training ones and the shifts are obviously dispersed which result in the different feature distributions for training and test adversarial examples. The total shifts difference can be evaluated with shift inconsistency SiC in Equation (8).

4. Adversarial Training with SCR

In this section, we will introduce our proposed regularization method named shift consistency regularization (SCR). According to previous sections, the feature shift inconsistency – the difference of latent features between training and test data – enlarges the robust generalization gap and consequently makes robust generalization harder to achieve. To tackle such problem, it is natural to penalize the shift inconsistency term SiC during the training as a regularizer.

However, it is impractical to compute expected feature shift over the unknown input distribution for $\hat{d}_\theta(z, C_i)$, with only the training dataset available. An alternative way is to approximate it with average feature shift over training data. Thus, for a data sample x_j and its adversarial example x_j^{adv}, we define an estimate of shift inconsistency term as

$$\widehat{\mathrm{SiC}}(x_j^{adv}, x_l, N_i) \triangleq \|d_\theta(x_j^{adv}) - \bar{d}_\theta(x_l, N_i)\|_2^2, \tag{9}$$

where $\bar{d}_\theta(x_l, N_i)$ is the average feature shifts over training data of class i, i.e.,

$$\bar{d}_\theta(x_l, N_i) = \frac{1}{|N_i|} \sum_{l \in N_i} (f_\theta(x_l^{adv}) - f_\theta(x_l)). \tag{10}$$

As such, we propose an adversarial training method with a novel shift consistency regularization (SCR) to improve the robust generalization, formulated as:

$$\begin{aligned} \min_\theta \quad & \Big\{ \sum_{i=1}^{n} [L(x_i^{adv}, y_i; \theta)] \\ & + \frac{\lambda}{n} \sum_{i=1}^{K} \sum_{j \in N_i} \widehat{\mathrm{SiC}}(x_j^{adv}, x_l, N_i) \Big\}, \\ \text{s.t.} \quad & x_i^{adv} = \arg \max_{x_i' \in S_{x_i}} L(x_i', y_i; \theta). \end{aligned} \tag{11}$$

where λ is the trade-off parameter. Nevertheless, the above formulation only penalizes the shift inconsistency caused by one specific adversarial attack. To enhance the adversarial robustness, we consider different adversarial attacks and propose to penalize the most inconsistent shift by replacing $\widehat{\mathrm{SiC}}(x_j^{adv}, x_l, N_i)$ with

$$\max_{x_j' \in S_{x_i}} \widehat{\mathrm{SiC}}(x_j', x_l, N_i). \tag{12}$$

The intuition behind the above formulation of adversarial training is as follows. In addition to training over adversarial examples for classification as done in the traditional approach, the SCR takes into account the most disperse feature shifts due to adversarial perturbation. By enforcing the inconsistent shifts to concentrate on the average of each class, robust generalization is attainable.

4.1. Iterative Implementation

Since adversarial training is computationally expensive, it is not practical to optimize (11) and (12) with the whole dataset once at a time. Instead, we solve the optimization problems (11) and (12) with batches. For the problem (12), we set the mean value of training data shifts as the trainable parameter μ_i instead of computing it directly. Then, the problem (12) can be reformulated as $\max_{x_j' \in S_{x_i}} \widehat{\mathrm{SiC}}(x_j', \mu_i)$ where $\widehat{\mathrm{SiC}}(x_j', \mu_i) \triangleq \|d_\theta(x_j') - \mu_i\|_2^2$. As such, we implement the proposed adversarial training with shift consistency regularization in an iterative way to reduce the computational complexity as detailed in Algorithm 1. Specifically, we first compute the adversarial examples for the loss function and shift inconsistency. Then, we optimize the model parameter θ to minimize the objective function. Finally, we update the mean parameter of feature shifts with updated adversarial examples.

5. Experiment

In this part, we perform extensive experiments to evaluate our proposed SCR in defending against various adversarial attacks. To save space, many results are provided in the supplementary file including the comparison against black-box attacks, sensitivity analysis, and more visualizations.

5.1. Robustness to Adversarial Attacks

We now evaluate the robustness performance of state-of-the-art adversarial training methods on CIFAR-10, CIFAR-100, and SVHN against white/black box adversarial attacks. We highlight the setting of three benchmark methods, i.e. FS, AT, and TRADES, all of which adopt WideResNet-28-10 as the baseline, by following (Zhang & Wang, 2019; Madry et al., 2017). Specifically, for FS, we follow (Zhang & Wang, 2019) on CIFAR-10 and CIFAR-100. We train 200 epochs using SGD with momentum 0.9, weight decay 5×10^{-4}, and initial learning rate 0.1. The learning rate decays at epoch 60 and 90 with the rate 0.1; for SVHN, the initial learning rate is 0.01 while the other settings remain the same as those on CIFAR-10 and CIFAR-100. For AT and TRADES, the total training epoch is 100 with the initial learning rate 0.1 for CIFAR-10 and CIFAR-100 and 0.01 for SVHN. The learning rate decays at 60 epoch with decay rate 0.1. The trade-off parameter is 0.01 on CIFAR-10 and SVHN and 0.0001 on CIFAR-100 for our proposed model. The iteration number is empirically set as 3 to compute the regularization and update step is $4/255$. All other hyper-parameters are the same as the baseline methods.

We attack different models with FGSM, PGD, and CW. (all attacks are computed with l_∞ norm) We adopt FS as the main baseline on which we apply the proposed SCR. We set the training attack iteration to 1, and both the attack step

Algorithm 1 Adversarial Training with SCR

for training iteration $l = 1, \ldots, T$ **do**
  Sample a batch of labeled data $\{\mathbf{x}_j, \mathbf{y}_j\}_{j=1}^n$ from training set with K classes. x_j^t is the perturbed example of x_j at t^{th} iteration and x_j^1 is initialized by adding the random perturbation. N_i is the set of indices of examples of class i in this batch. μ_i denotes the mean value of feature shifts over the i^{th} class data of the whole training data. We initialize μ_i with random values. α and β are the updating rate for perturbed examples and the mean μ_i respectively. m is the total number of classes. c_1 and c_2 are attack iterations for loss function and shift inconsistency.
  Compute adversarial example x_j^{adv} for classification:
  for $j = 1, \ldots, n$ **do**
    for $t = 1, \ldots, c_1$ **do**
      $x_j^{t+1} = \Pi_{S_{x_j}}(x_j^t + \alpha_1 \cdot \mathrm{sgn}(\nabla_x L(x_j^t, y_j; \theta)))$
    end for
    $x_j^{adv} = x_j^{c_1}$
  end for
  Compute adversarial example x_k^{st} to maximize shift inconsistency:
  for $i = 1, \ldots, m$ **do**
    for $k \in N_i$ **do**
      for $t = 1, \ldots, c_2$ **do**
        $x_k^{t+1} = \Pi_{S_{x_j}}(x_k^t + \alpha_1 \cdot \mathrm{sgn}(\nabla_{x_k^t} \| f_\theta(x_k^t) - f_\theta(x_k) - \mu_i \|_2^2))$
      end for
      $x_k^{\mathrm{st}} = x_k^{c_2}$
    end for
  end for
  where $\prod$ is projection operator. Update the parameters of neural network θ with

$$-\nabla_\theta \Big\{ \sum_{i=1}^{n} [L(x_i^{adv}, y_i; \theta)] + \frac{\lambda}{n} \sum_{i=1}^{K} \sum_{j \in N_i} \widehat{\mathrm{SiC}}(x_j^{\mathrm{st}}, \mu_i) \Big\}$$

  Update the feature shift mean value:
  for $i = 1, \ldots, m$ **do**
    $\mu_i = \beta \mu_i + (1 - \beta) \frac{1}{|N_i|} \sum_{j \in N_i} (f_\theta(x_j^{\mathrm{st}}) - f_\theta(x_j))$
  end for
end for=0

size and amplitude to $8/255$.

We compare the proposed FS-SCR with the current state-of-the-art adversarial training methods such as 1) AT (Madry et al., 2017), 2) TLA (Mao et al., 2019), 3) LAT (Sinha et al., 2019) 4) Bilateral (Wang & Zhang, 2019), 5) FS (Zhang & Wang, 2019). In addition, we evaluate FS-SCR against other recent methods which also aim to promote the robust generalization including 6) RST/AT-AWP (Wu et al., 2020), 7) RLFAT$_{T/P}$ (Song et al., 2019). We list the performance of different methods in Table 1 and Table 2 respectively on CIFAR-10, CIFAR-100 and SVHN. For CIFAR-10, it can be noted that our proposed FS-SCR attains the overall best performance for all the attacks except that it is slightly worse than FS on PGD100. Particularly, our method shows obvious superiority over recent robust generalization methods RST/AT-AWP, RLFAT$_{T/P}$. For CIFAR-100 and SVHN, our proposed method performs even better, demonstrating consistently higher accuracy than all the other models.

Table 1. Accuracy under white-box attacks on CIFAR-10

Models	Clean	Accuracy under White-box Attack ($\epsilon = 8$)						
		FGSM	PGD20	PGD40	PGD100	CW20	CW40	CW100
Standard	**95.60**	36.90	0.00	0.00	0.00	0.00	0.00	0.00
AT	85.70	54.90	44.90	44.80	44.80	45.70	45.60	45.40
TLA	86.21	58.88	51.59	-	-	-	-	-
LAT	87.80	-	53.84	-	53.04	-	-	-
Bilateral	91.20	70.70	57.50	–	55.20	56.20	–	53.80
FS	90.00	78.40	70.50	70.30	**68.60**	62.40	62.10	60.60
RST-AWP	88.25	67.94	63.73	-	63.58	61.62	-	-
RLFAT$_T$	82.72	-	58.75	-	-	51.94	-	-
RLFAT$_P$	84.77	-	53.97	-	-	52.40	-	-
FS-SCR	92.70	**89.87**	**76.45**	**71.60**	67.79	**75.42**	**72.69**	**69.79**

5.2. Effect on Different Baselines & Attack Budget

We further examine the effects of SCR on different basesline models and under different attack budget where Auto Attack (AA), a recent stronger attack (Croce & Hein, 2020) was also compared (both our method and baselines are trained with l_∞ norm). Due to limited space, we take CIFAR-10 as one typical example to illustrate such results. First, we examine if SCR could consistently improve the robustness of various models, particularly AT, TRADES, and FS. We report such comparisons in Figure 4(a)-4(c). As clearly observed, SCR apparently boosts the robustness of all the three baseline methods. Second, we also evaluate how SCR would affect the baseline model FS under different attack budgets. Specifically, we attack FS-SCR and its baseline FS with PGD and CW of different attack budgets and report such results in Figure 4(d)-4(e). Once again, SCR can consistently improve FS substantially when different attack budgets are applied. All these results validate the efficacy of SCR on improving the robust generalization.

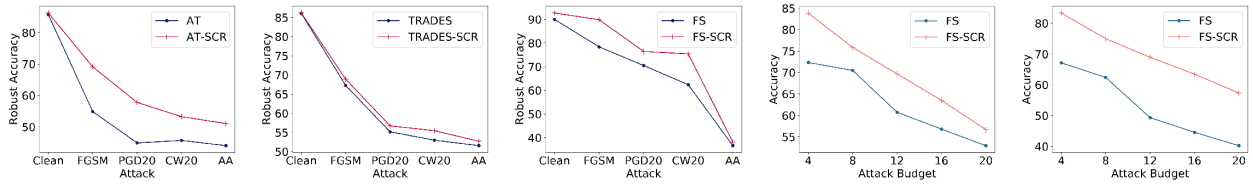
5.3. Further Analysis

We now offer more analysis and visualizations to interpret how and why SCR could promote the robust generalization.

Generalization Analysis To examine the robust generalization of SCR, we plot in Figure 5 the generalization gap difference (i.e. the difference of accuracy gap on training and test set of CIFAR-10 cf. Section 3.2 for detailed definition) for both FS and FS-SCR at different training epochs, under the attacks of CW20 and PGD20. We also draw the feature shift inconsistency for FS and FS-SCR on CIFAR-10. It can be noted that feature shift inconsistency manifests

Table 2. Accuracy under different white-box attack on CIFAR-100 and SVHN

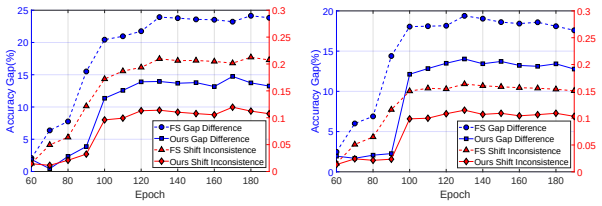
MODELS	CIFAR-100($\epsilon = 8$)						SVHN($\epsilon = 8$)					
	CLEAN	FGSM	PGD20	PGD100	CW20	CW100	CLEAN	FGSM	PGD20	PGD100	CW20	CW100
STANDARD	**79.00**	10.00	0.00	0.00	0.00	0.00	**97.20**	53.00	0.30	0.10	0.30	0.10
AT	59.90	28.50	22.60	22.30	23.20	23.00	93.90	68.40	47.90	46.00	48.70	47.30
LAT	60.94	-	27.03	26.41	-	-	60.94	-	60.23	59.97	-	-
BILATERAL	68.20	60.80	26.70	25.30	-	22.10	94.10	69.80	53.90	50.30	-	48.90
FS	73.90	61.00	47.20	46.20	34.60	30.60	96.20	83.50	62.90	52.00	61.30	50.80
AT-AWP	-	-	30.71	-	-	-	-	-	59.12	-	-	-
RLFAT$_T$	58.96	-	31.63	-	27.54	-	-	-	-	-	-	-
RLFAT$_P$	56.70	-	31.99	-	29.04	-	-	-	-	-	-	-
FS-SCR	74.20	**72.19**	**48.87**	**47.34**	**38.90**	**33.60**	**96.60**	**92.52**	**70.24**	**60.72**	**64.62**	**54.90**



(a) AT (b) TRADES (c) FS (d) PGD Attack Budget (e) CW Attack Budget

Figure 4. Effect of SCR on different baselines (attacking with FGSM, PGD, CW, and AA) & attack budget (CIFAR-10).

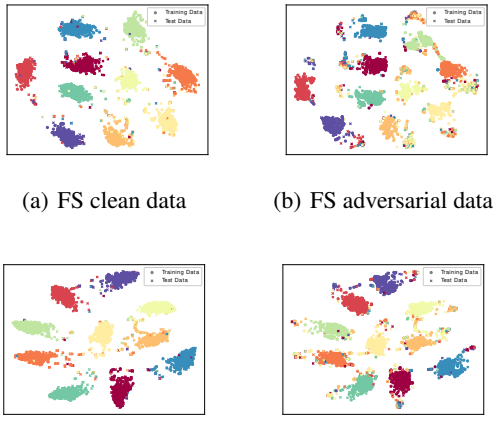
an overall similar trend to the generalization gap. In addition, our proposed FS-SCR obtains smaller generalization gap difference and smaller shift inconsistency than FS. It indicates that SCR diminishes the performance difference between clean and adversarial data and improves the robust generalization through penalizing the shift inconsistency.



(a) CW 20 (b) PGD 20

Figure 5. The generalization gaps (accuracy gap difference) and shift inconsistency at different training epochs on CIFAR-10.

Feature Analysis We have visualized the output features of clean and adversarial data under PGD attacks for both FS-SCR and FS earlier in Figure 1. We now visualize further in Figure 6 these features under CW attacks. Again cross and dot denote the test and training data feature. Obviously, after SCR is applied, FS-SCR leads to very similar distributions between the latent adversarial features of the training and test data (though there exist certain test feature points shifted away from their distributions). Contrastively, in FS, more test feature points shift away from their distribution. Even worse, the structures of feature distributions of some classes are undermined in FS. Namely, the feature distribution of FS becomes more complicated when adversarial perturbations are added which deteriorates the robust generalization. Similar observations can also be earlier inspected in Figure 3 where the features of training data turn to share more consistent shifts with those of test data. Further, the latent dispersion of test set can be made smaller after SCR is applied and more consistent training and test distributions are obtained, thus attaining better robust generalization.



(a) FS clean data (b) FS adversarial data

(c) FS-SCR clean data (d) FS-SCR adversarial data

Figure 6. Visualization via TSNE of latent features without and with SCR applied during FS adversarial training (attacked by CW).

Effectiveness on adaptive attack. To demonstrate that our proposed method can still work for the test-time attacks with more adaptive ones, we conducted additional experiments applying our proposed SCR method to three different adversarial training models (i.e., FS, AT and TRADES) to defend against adaptive attacks (adaptive attacks are based on FGSM, PGD20 and CW20) on different datasets with different loss weights. The results are shown in Table **??**,

Table 3. Robustness accuracy under different adaptive attacks

Method (FS+SCR)												
Datasets	0*CE+1*SIC			1*CE+0.1*SIC			1*CE+0.5*SIC			1*CE+1*SIC		
	FGSM	PGD20	CW20	FGSM	PGD20	CW20	FGSM	PGD20	CW20	FGSM	PGD20	CW20
CIFAR-10	92.64	80.69	80.81	89.75	76.40	75.56	91.42	77.99	75.55	91.93	79.19	76.68
CIFAR-100	72.66	52.63	48.96	72.46	48.16	34.65	69.58	48.49	36.68	72.58	48.96	37.82
SVHN	94.67	69.15	66.89	94.66	69.09	65.42	94.70	69.07	66.76	94.32	69.63	65.96
Method (AT+SCR)												
Datasets	0*CE+1*SIC			1*CE+0.1*SIC			1*CE+0.5*SIC			1*CE+1*SIC		
	FGSM	PGD20	CW20	FGSM	PGD20	CW20	FGSM	PGD20	CW20	FGSM	PGD20	CW20
CIFAR-10	68.96	60.23	59.67	68.75	57.42	55.10	68.62	56.53	54.87	68.12	58.21	55.78
Method (TRADES+SCR)												
Datasets	0*CE+1*SIC			1*CE+0.1*SIC			1*CE+0.5*SIC			1*CE+1*SIC		
	FGSM	PGD20	CW20	FGSM	PGD20	CW20	FGSM	PGD20	CW20	FGSM	PGD20	CW20
CIFAR-10	69.16	61.21	58.76	68.42	58.96	56.67	68.06	58.21	56.96	69.02	59.08	56.26

where CE indicates the cross entropy loss and SIC means the shift consistency regularization term. The attack consists of two parts that come respectively from CE and SIC. Compared with the results in Table 1-2 and Figure 4, it can be noted that SCR regularized methods achieve comparable or even higher accuracy on such adaptive attacks than the conventional attacks. It is because the attack component generated according to the regularization term is weaker due to its irrelevance to the decision boundaries.

6. Conclusion

In this paper, we have shown – both theoretically and empirically – the poor robust generalization of adversarial training is attributed to the latent feature inconsistent shift of adversarial training and test data. Inspired by this, we proposed a novel shift consistency regularization technique to achieve better robust generalization for adversarial training. It is expected to stimulate the further investigation of the interplay between robustness and generalization.

Acknowledgements

The work was partially supported by the following: National Natural Science Foundation of China under no.61876155 and no.61876154; Jiangsu Science and Technology Programme (Natural Science Foundation of Jiangsu Province) under no. BE2020006-4B, BK20181189, BK20181190; Key Program Special Fund in XJTLU under no. KSF-T-06, KSF-E-26, and KSF-A-10, and XJTLU Research Development Fund under no. RDF-16-02-49. This paper is also partially supported by Beijing Information Science Technology University's Key research and cultivation projects 2121YJPY224.

References

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein gan. In *arXiv preprint arXiv:1701.07875*, 2017.

Bousquet, O. and Elisseeff, A. Stability and generalization. In *The Journal of Machine Learning Research*, volume 2, pp. 499–526. JMLR. org, 2002.

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

Carlini, N. and Wagner, D. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE Security and Privacy Workshops (SPW)*, pp. 1–7. IEEE, 2018.

Croce, F. and Hein, M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*, 2020.

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Tramer, F., Prakash, A., Kohno, T., and Song, D. Physical adversarial examples for object detectors. In *arXiv preprint arXiv:1807.07769*, 2018.

Farnia, F., Zhang, J. M., and Tse, D. Generalizable adversarial training via spectral normalization. In *arXiv preprint arXiv:1811.07457*, 2018.

Fischer, V., Kumar, M. C., Metzen, J. H., and Brox, T. Adversarial examples for semantic image segmentation. In *arXiv preprint arXiv:1703.01101*, 2017.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In *arXiv:1412.6572*, 2014.

Gu, S. and Rigazio, L. Towards deep neural network architectures robust to adversarial examples. In *arXiv:1412.5068*, 2014.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE*

conference on computer vision and pattern recognition, pp. 770–778, 2016.

Kannan, H., Kurakin, A., and Goodfellow, I. Adversarial logit pairing. In *arXiv:1803.06373*, 2018.

Kurakin, A., Goodfellow, I., and Bengio, S. Adversarial machine learning at scale. In *arXiv preprint arXiv:1611.01236*, 2016.

LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. In *nature*, volume 521, pp. 436. Nature Publishing Group, 2015.

Lyu, C., Huang, K., and Liang, H.-N. A unified gradient regularization family for adversarial examples. In *2015 IEEE International Conference on Data Mining*, pp. 301–309. IEEE, 2015.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *arXiv:1706.06083*, 2017.

Mao, C., Zhong, Z., Yang, J., Vondrick, C., and Ray, B. Metric learning for adversarial robustness. In *Advances in Neural Information Processing Systems*, pp. 478–489, 2019.

Miyato, T., Maeda, S.-i., Koyama, M., and Ishii, S. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. In *arXiv:1704.03976*, 2017.

Moosavi-Dezfooli, S.-M., Fawzi, A., and Frossard, P. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Neyshabur, B., Bhojanapalli, S., McAllester, D., and Srebro, N. Exploring generalization in deep learning. In *arXiv preprint arXiv:1706.08947*, 2017.

Nowozin, S., Cseke, B., and Tomioka, R. f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in neural information processing systems*, pp. 271–279, 2016.

Otter, D. W., Medina, J. R., and Kalita, J. K. A survey of the usages of deep learning for natural language processing. In *IEEE Transactions on Neural Networks and Learning Systems*. IEEE, 2020.

Raghunathan, A., Xie, S. M., Yang, F., Duchi, J. C., and Liang, P. Adversarial training can hurt generalization. *arXiv preprint arXiv:1906.06032*, 2019.

Roth, K., Kilcher, Y., and Hofmann, T. Adversarial training is a form of data-dependent operator norm regularization. *Advances in Neural Information Processing Systems*, 33, 2020.

Salvaris, M., Dean, D., and Tok, W. H. Generative adversarial networks. In *Deep Learning with Azure*, pp. 187?08. Apress, 2018. ISBN 9781484236796. doi: 10.1007/978-1-4842-3679-6_8. URL http://dx.doi.org/10.1007/978-1-4842-3679-6_8.

Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K., and Madry, A. Adversarially robust generalization requires more data. In *Advances in Neural Information Processing Systems*, pp. 5014–5026, 2018.

Shafahi, A., Najibi, M., Ghiasi, A., Xu, Z., Dickerson, J., Studer, C., Davis, L. S., Taylor, G., and Goldstein, T. Adversarial training for free! *arXiv preprint arXiv:1904.12843*, 2019.

Sinha, A., Namkoong, H., Volpi, R., and Duchi, J. Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2017.

Sinha, A., Singh, M., Kumari, N., Krishnamurthy, B., Machiraju, H., and Balasubramanian, V. Harnessing the vulnerability of latent layers in adversarially trained models. In *arXiv:1905.05186*, 2019.

Song, C., He, K., Lin, J., Wang, L., and Hopcroft, J. E. Robust local features for improving the generalization of adversarial training. In *arXiv preprint arXiv:1909.10147*, 2019.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

Wang, J. and Zhang, H. Bilateral adversarial training: Towards fast training of more robust models against adversarial attacks. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

Wong, E., Rice, L., and Kolter, J. Z. Fast is better than free: Revisiting adversarial training. *arXiv preprint arXiv:2001.03994*, 2020.

Wu, D., Xia, S.-T., and Wang, Y. Adversarial weight perturbation helps robust generalization. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Xu, H. and Mannor, S. Robustness and generalization. In *Machine learning*, volume 86, pp. 391–423. Springer, 2012.

Yang, Y.-Y., Rashtchian, C., Zhang, H., Salakhutdinov, R., and Chaudhuri, K. A closer look at accuracy vs. robustness. *Advances in Neural Information Processing Systems*, 33, 2020.

Yin, D., Kannan, R., and Bartlett, P. Rademacher complexity for adversarially robust generalization. In *International Conference on Machine Learning*, pp. 7085–7094. PMLR, 2019.

Zagoruyko, S. and Komodakis, N. Wide residual networks. In *arXiv preprint arXiv:1605.07146*, 2016.

Zhai, R., Cai, T., He, D., Dan, C., He, K., Hopcroft, J., and Wang, L. Adversarially robust generalization just requires more unlabeled data. In *arXiv preprint arXiv:1906.00555*, 2019.

Zhang, D., Zhang, T., Lu, Y., Zhu, Z., and Dong, B. You only propagate once: Accelerating adversarial training via maximal principle. *arXiv preprint arXiv:1905.00877*, 2019a.

Zhang, H. and Wang, J. Defense against adversarial attacks using feature scattering-based adversarial training. In *Advances in Neural Information Processing Systems*, pp. 1829–1839, 2019.

Zhang, H., Yu, Y., Jiao, J., Xing, E. P., Ghaoui, L. E., and Jordan, M. I. Theoretically principled trade-off between robustness and accuracy. *arXiv preprint arXiv:1901.08573*, 2019b.

Zhang, J., Xu, X., Han, B., Niu, G., Cui, L., Sugiyama, M., and Kankanhalli, M. Attacks which do not kill training make adversarial learning stronger. In *arXiv preprint arXiv:2002.11242*, 2020.

Zhang, S., Huang, K., Zhu, J., and Liu, Y. Manifold adversarial learning. In *arXiv:1807.05832*, 2018.

Zhao, Z.-Q., Zheng, P., Xu, S.-t., and Wu, X. Object detection with deep learning: A review. In *IEEE transactions on neural networks and learning systems*, volume 30, pp. 3212–3232. IEEE, 2019.

Zhu, C., Cheng, Y., Gan, Z., Sun, S., Goldstein, T., and Liu, J. Freelb: Enhanced adversarial training for natural language understanding. *arXiv preprint arXiv:1909.11764*, 2019.