Consistency Regularization for Adversarial Robustness

Jihoon Tack[1], Sihyun Yu[1], Jongheon Jeong[1], Minseon Kim[1], Sung Ju Hwang[1,2], Jinwoo Shin[1]

[1]Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea
[2]AITRICS, Seoul, South Korea
{jihoontack,sihyun.yu,jongheonj,minseonkim,sjhwang82,jinwoos}@kaist.ac.kr

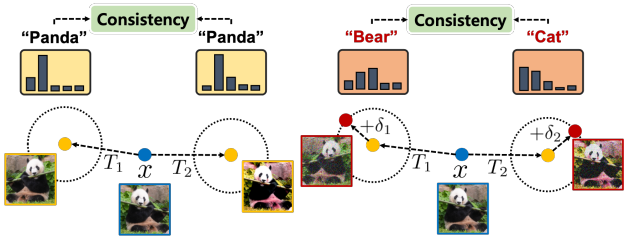
Abstract

Adversarial training (AT) is currently one of the most successful methods to obtain the adversarial robustness of deep neural networks. However, the phenomenon of robust overfitting, *i.e.*, the robustness starts to decrease significantly during AT, has been problematic, not only making practitioners consider a bag of tricks for a successful training, *e.g.*, early stopping, but also incurring a significant generalization gap in the robustness. In this paper, we propose an effective regularization technique that prevents robust overfitting by optimizing an auxiliary 'consistency' regularization loss during AT. Specifically, we discover that data augmentation is a quite effective tool to mitigate the overfitting in AT, and develop a regularization that forces the predictive distributions after attacking from two different augmentations of the same instance to be similar with each other. Our experimental results demonstrate that such a simple regularization technique brings significant improvements in the test robust accuracy of a wide range of AT methods. More remarkably, we also show that our method could significantly help the model to generalize its robustness against unseen adversaries, *e.g.*, other types or larger perturbations compared to those used during training. Code is available at https://github.com/alinlab/consistency-adversarial.

Introduction

Despite the remarkable success of deep neural networks (DNNs) in real-world applications (He et al. 2016a; Girshick 2015; Amodei et al. 2016), recent studies have demonstrated that DNNs are vulnerable to adversarial examples, *i.e.*, inputs crafted by an imperceptible perturbation which confuse the network prediction (Szegedy et al. 2014; Goodfellow, Shlens, and Szegedy 2015). This vulnerability of DNNs raises serious security concerns about their deployment in the real-world applications (Kurakin, Goodfellow, and Bengio 2016; Li, Schmidt, and Kolter 2019), *e.g.*, self-driving cars and secure authentication system (Chen et al. 2015).

In this respect, there have been significant efforts to design various defense techniques against the adversarial examples, including input denoising (Guo et al. 2018; Liao et al. 2018), detection techniques (Ma et al. 2018; Lee et al. 2018), and certificating the robustness of a classifier (Cohen, Rosenfeld, and Kolter 2019; Jeong and Shin 2020). Overall, *adversarial training* (AT) is currently one of the most promising ways to obtain the adversarial robustness of DNNs, *i.e.*, directly augmenting the training set with adversarial examples (Goodfellow, Shlens, and Szegedy 2015; Madry et al. 2018). Recent studies have been actively investigating a better form of AT (Qin et al. 2019; Zhang et al. 2019; Wang et al. 2020).



(a) Conventional CR (b) Proposed CR

Figure 1: An overview of our consistency regularization (CR) and conventional approach (Hendrycks et al. 2020; Xie et al. 2020). Our regularization forces the predictive distribution of *attacked* augmentations to be consistent. T and δ indicates the randomly sampled augmentation, and the corresponding adversarial noise, respectively.

One of the major downsides that most AT methods suffer from, however, is a significant generalization gap of adversarial robustness between the train and test datasets (Yang et al. 2020), possibly due to an increased sample complexity induced by the non-convex, minimax nature of AT (Schmidt et al. 2018). More importantly, it has been observed that such a gap gradually increases from the middle of training (Rice, Wong, and Kolter 2020), *i.e.*, overfitting, which makes practitioners to consider several heuristic approaches for a successful optimization *e.g.*, early stopping (Zhang et al. 2019). Only recently, a few proposed more advanced regularization techniques, *e.g.*, self-training (Chen et al. 2021)[1] and weight perturbation (Wu, Xia, and Wang 2020), but it is still largely unknown to the community that why and how only such sophisticated training schemes could be effective to prevent the robust overfitting of AT.

Contribution. In this paper, we suggest to optimize an

[1]We do not consider comparing with the method by Chen et al. (2021) as they require pre-training additional models.

auxiliary '*consistency*' regularization loss, as a simpler and easy-to-use alternative for regularizing AT. To this end, we first found that the existing data augmentation (DA) schemes are already quite effective to reduce the robust overfitting in AT. Yet, it is contrast to the recent studies (Rice, Wong, and Kolter 2020; Gowal et al. 2020) which reported DA does not help for AT. Our new finding is that considering more diverse set of augmentations than the current conventional practice can prevent the robust overfitting: we use AutoAugment (Cubuk et al. 2019) which is an effective augmentation for standard cross-entropy training.

Upon the observation, we claim that the way of incorporating such augmentations could play a significant role in AT. Specifically, we suggest to optimize an auxiliary *consistency regularization* loss during AT: it forces *adversarial examples* from two independent augmentations of the same input to have similar predictions. Here, we remark that forcing the prediction consistency over 'clean' DA is widely used for many purposes (Zhang et al. 2020; Hendrycks et al. 2020), however, it looks highly non-trivial at first glance whether matching such attack directions over DA is useful in any sense. Our finding is that the attack direction provides intrinsic information of the sample (other than its label), where the most frequently attacked class is the most confusing class of the 'clean' input, *i.e.*, class with the maximum softmax probability disregarding the true label. The proposed regularization loss injects a strong inductive bias to the model that such 'dark' knowledge (Hinton, Vinyals, and Dean 2015) over DA should be consistent. Our regularization technique is easy to apply to any existing AT methods (Madry et al. 2018; Zhang et al. 2019; Wang et al. 2020), yet effectively improves the performance.

We verify the efficacy of our scheme through extensive evaluations on CIFAR-10/100 (Krizhevsky and Hinton 2009) and Tiny-ImageNet.[2] Overall, our experimental results show that the proposed regularization can be easily adapted for a wide range of AT methods to prevent overfitting in robustness. For example, our regularization could improve the robust accuracy of WideResNet (Zagoruyko and Komodakis 2016) trained via standard AT (Madry et al. 2018) on CIFAR-10 from 45.62%→52.36%. Moreover, we show that our regularization could even notably improve the robustness against unforeseen adversaries (Tramer and Boneh 2019), *i.e.*, when the adversaries assume different threat models from those used in training: *e.g.*, our method could improve the l_1-robustness of TRADES (Zhang et al. 2019) from 29.58%→48.32% on PreAct-ResNet (He et al. 2016b). Finally, we also observe that our method could be even beneficial for the corruption robustness (Hendrycks and Dietterich 2019).

Consistency Regularization for Adversarial Robustness

In this section, we introduce a simple yet effective strategy for preventing the robust overfitting in adversarial training (AT). We first review the concept of AT and introduce one of popular AT methods in Section . We then start in Section by showing that the data augmentations can effectively prevent the robustness overfitting. Finally, in Section , we propose a simple yet effective consistency regularization to further utilize the given data augmentations in AT.

Preliminaries: Adversarial Training

We consider a classification task with a given K-class dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n \subseteq \mathcal{X} \times \mathcal{Y}$, where $x \in \mathbb{R}^d$ represents an input sampled from a certain data-generating distribution P in an *i.i.d.* manner, and $\mathcal{Y} := \{1, \dots, K\}$ represents a set of possible class labels. Let $f_\theta : \mathbb{R}^d \to \Delta^{K-1}$ be a neural network modeled to output a probability simplex $\Delta^{K-1} \in \mathbb{R}^K$, *e.g.*, via a softmax layer. The notion of adversarial robustness requires f_θ to perform well not only on P, but also on the worst-case distribution near P under a certain distance metric. More concretely, the adversarial robustness we primarily focus in this paper is the ℓ_p*-robustness*: *i.e.*, for a given $p \geq 1$ and a small $\epsilon > 0$, we aim to train a classifier f_θ that correctly classifies $(x + \delta, y)$ for any $\|\delta\|_p \leq \epsilon$, where $(x, y) \sim P$.

The high level idea of *adversarial training* (AT) is to directly incorporate adversarial examples to train the classifier (Goodfellow, Shlens, and Szegedy 2015), hence the network becomes robust to such adversaries. In general, AT methods formalize the training of f_θ as an alternative min-max optimization with respect to θ and $||\delta||_p \leq \epsilon$, respectively; *i.e.*, one minimizes a certain classification loss $\mathcal{L}$ with respect to θ while an adversary maximizes $\mathcal{L}$ by perturbing the given input to $x + \delta$ during training. Here, for a given $\mathcal{L}$, we denote the inner maximization procedure of AT as $\mathcal{L}_{\texttt{adv}}(x, y; \theta)$:

$$\mathcal{L}_{\texttt{adv}}(x, y; \theta) := \max_{\|\delta\|_p \leq \epsilon} \mathcal{L}\big(x + \delta, y; \theta\big). \tag{1}$$

For example, one of most basic form of AT method (Madry et al. 2018) considers to design $\mathcal{L}_{\texttt{adv}}$ with the standard cross-entropy loss $\mathcal{L}_{\texttt{CE}}$ (we also provide an overview on other types of AT objective such as TRADES (Zhang et al. 2019) and MART (Wang et al. 2020), in the supplementary material):

$$\mathcal{L}_{\texttt{AT}} := \max_{\|\delta\|_p \leq \epsilon} \mathcal{L}_{\texttt{CE}}\big(f_\theta(x + \delta), y\big). \tag{2}$$

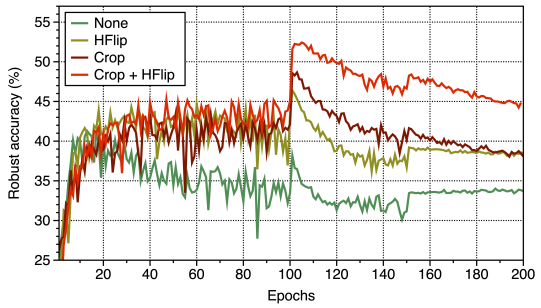
Effect of Data Augmentations in Adversarial Training

Now, we investigate the utility of data augmentations in AT. We first show that current standard choices of augmentation in AT are already somewhat useful for relaxing the robust overfitting, where considering more diverse augmentations is even more effective. Throughout this section, we train PreAct-ResNet-18 (He et al. 2016b) on CIFAR-10 (Krizhevsky and Hinton 2009) using standard AT (Madry et al. 2018), following the training details of Rice, Wong, and Kolter (2020). We use projected gradient descent (PGD) with 10 iterations under $\epsilon = 8/255$ (step size of $2/255$) with l_∞ constraint to perform adversarial attacks for both training and evaluation. Formally, for a given training sample $(x, y) \sim \mathcal{D}$, and augmentation $T \sim \mathcal{T}$, the training loss is:

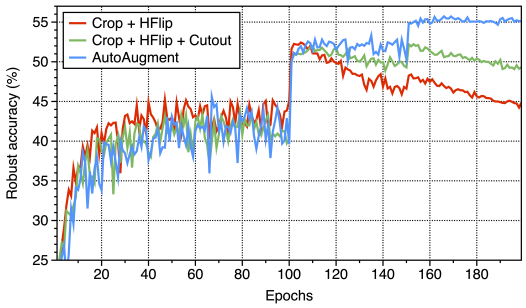
$$\max_{||\delta||_\infty \leq \epsilon} \mathcal{L}_{\texttt{CE}}(f_\theta(T(x) + \delta), y). \tag{3}$$

Unless otherwise specified, we assume the set of baseline augmentations $\mathcal{T} := \mathcal{T}_{\texttt{base}}$ (*i.e.*, random crop with 4 pixels zero padding and horizontal flip) by default for this section.

[2] https://tiny-imagenet.herokuapp.com/



(a) Conventional augmentations



(b) Additional augmentations

Figure 2: Robust accuracy (%) against PGD-10 attack on standard AT (Madry et al. 2018) under (a) conventional augmentations, and (b) additional augmentations to the convention. We consider PreAct-ResNet-18 trained on CIFAR-10. We use l_∞ threat model with $\epsilon = 8/255$. None, HFlip, and Crop, indicates no augmentation, horizontal flip, and random crop, respectively. Note that the AutoAugment (Cubuk et al. 2019) includes horizontal flip, random crop and Cutout (DeVries and Taylor 2017). The jump in robust accuracy at 100, 150 epochs is due to a drop in the learning rate.

Role of base augmentations in adversarial training. We recognize the set of base augmentations $\mathcal{T}_{\texttt{base}}$ has been commonly used in most existing AT methods, and observe these augmentations are already somewhat useful for relaxing the robust overfitting in AT. To this end, we conduct a controlled experiment by removing each augmentation from the predefined augmentation set $\mathcal{T}_{\texttt{base}}$ and train the network. Figure 2a summarizes the result of the experiment. As each augmentation is removed, not only the robustness degrades, but also the adversarial overfitting is getting significant. This phenomenon stands out more when no augmentations are applied during AT, which only shows the increment of robust accuracy at the first 5% of the whole training procedure. This result implies that there may exist an augmentation family that effectively prevents the robust overfitting as the base augmentation is already useful.

Reducing robust overfitting with data augmentations. We further find that the existing data augmentation schemes are already quite effective to reduce the robust overfitting in AT. Specifically, we utilize AutoAugment (Cubuk et al. 2019) which is the state-of-the-arts augmentation scheme for the standard cross-entropy training. As shown in Figure 2b, the robust overfitting is gradually reduced as more diverse augmentations are used, and even the best accuracy improves. Note that AutoAugment is more diverse than the conventional augmentations as it includes the $\mathcal{T}_{\texttt{base}}$ and Cutout (DeVries and Taylor 2017). Interestingly, our empirical finding somewhat shows a different conclusion from the previous studies (Gowal et al. 2020) which conclude that data augmentations are not effective for preventing the robust overfitting. We further discuss a detailed analysis of data augmentations in the supplementary material.

Consistency Regularization for Adversarial Training

We suggest to optimize a simple auxiliary *consistency regularization* during AT to further utilize the given data augmentations. Specifically, our regularization forces *adversarial examples* from two independent augmentations of an instance to have a similar prediction (see Figure 1). However, it is highly non-trivial whether matching such attack directions via consistency regularization is useful, which we essentially investigate in this paper. Our major finding is that the attack direction itself contains intrinsic information of the instance, as in Section . For example, the most frequently attacked class is the most confusing class of the 'clean' input, *i.e.*, class with the maximum softmax probability disregarding the true label. Hence, our regularization utilize this dark knowledge (other than the true labels) of samples and induce a strong inductive bias to the classifier.

Formally, for a given data point $(x, y) \sim \mathcal{D}$ and augmentations $T_1, T_2 \sim \mathcal{T}$, we denote δ_i as an adversarial noise of $T_i(x)$, *i.e.*, $\delta_i := \arg\max_{\|\delta\|_p \leq \epsilon} \mathcal{L}\big(T_i(x), y, \delta; \theta\big)$. We consider regularizing the temperature-scaled distribution $\hat{f}_\theta(x; \tau)$ (Guo et al. 2017) over the adversarial examples across augmentations to be consistent, where τ is the temperature hyperparameter. Concretely, temperature-scaled classifier is $\hat{f}_\theta(x; \tau) = \texttt{Softmax}(z_\theta(x)/\tau)$ where $z_\theta(x)$ is the logit value of $f_\theta(x)$, *i.e.*, activation before the softmax layer of $f_\theta(x)$. Then the proposed regularization loss is given by:

$$\texttt{JS}\Big(\hat{f}_\theta\big(T_1(x) + \delta_1; \tau\big) \parallel \hat{f}_\theta\big(T_2(x) + \delta_2; \tau\big)\Big), \qquad (4)$$

where $\texttt{JS}(\cdot \parallel \cdot)$ denotes the Jensen-Shannon divergence. Since the augmentations are randomly sampled in every training step, adversarial example's predictions become consistent regardless of augmentation selection when minimizing the proposed objective. We note that the motivation behind the temperature scaling is that the confidence of prediction (*i.e.*, maximum softmax value) is relatively low on AT than the standard training. Hence, we compensate this issue by enforcing the sharp distribution by using a small temperature.

Comparison to other consistency regularization loss over DA. There has been prior works that suggested a consistency regularization loss to better utilize DA (Hendrycks et al. 2020; Zhang et al. 2020; Sohn et al. 2020), which can be expressed with the following form:

$$\texttt{D}\Big(f_\theta\big(T_1(x)\big), f_\theta\big(T_2(x)\big)\Big), \qquad (5)$$

Loss	Clean	PGD-100
AT (3)	85.41	55.18
AT (3) + previous CR (5)	88.01	53.11
AT (3) + proposed CR (4)	86.45	56.38

Table 1: Comparison of the consistency regularization (CR) loss. We report clean accuracy and robust accuracy (%) against PGD-100 attack of PreAct-ResNet-18 trained on CIFAR-10. We use l_∞ threat model with $\epsilon = 8/255$.

where $\mathtt{D}$ is a discrepancy function. The regularization term used in (5) has a seemingly similar formula to ours but there is a fundamental difference: our method (4) does not match the predictions directly for the 'clean' augmented samples, but does after *attacking* them independently, *i.e.*, $f_\theta(T(x) + \delta)$. To examine which one is better, we compare (4) with (5) under the same discrepancy function, $\mathtt{D} := \mathtt{JS}$ and same augmentation family, *i.e.*, AutoAugment. As shown in Table 1, our design choice (4) improves both clean and robust accuracy compare to the baseline (3), while the prior consistency regularization (5) shows significant degradation on the robust accuracy. We additionally try to attack only single augmented instance in (5), where it also shows degradation in the robust accuracy, *e.g.*, 53.20% against PGD-100 (such regularization is used in unsupervised AT (Kim, Tack, and Hwang 2020)).

Overall training objective. In the end, we derive a final training objective, $\mathcal{L}_{\mathtt{total}}$: an AT objective combined with the consistency regularization loss (4). To do so, we consider the average of inner maximization objective on AT $\mathcal{L}_{\mathtt{adv}}$ (1) over two independent augmentations $T_1, T_2 \sim \mathcal{T}$, as minimizing (1) over the augmentations $T \sim \mathcal{T}$ is equivalent to an average of (1) over T_1 and T_2:

$$\frac{1}{2}\Big(\mathcal{L}_{\mathtt{adv}}\big(T_1(x), y; \theta\big) + \mathcal{L}_{\mathtt{adv}}\big(T_2(x), y; \theta\big)\Big). \tag{6}$$

We then combine our regularizer (4) with a given hyperparameter λ, into the average of inner maximization objectives (6). Then the final training objective $\mathcal{L}_{\mathtt{total}}$ is as follows:

$$\mathcal{L}_{\mathtt{total}} := \frac{1}{2}\sum_{i=1}^{2} \mathcal{L}_{\mathtt{adv}}\big(T_i(x), y; \theta\big) + \lambda \cdot \mathtt{JS}\Big(\hat{f}_\theta\big(T_1(x) + \delta_1; \tau\big) \parallel \hat{f}_\theta\big(T_2(x) + \delta_2; \tau\big)\Big).$$

Note that our regularization scheme is agnostic to the choice of AT objective, hence, can be easily incorporated into well-known AT methods (Madry et al. 2018; Zhang et al. 2019; Wang et al. 2020). For example, considering standard AT loss (Madry et al. 2018) as the AT objective, *i.e.*, $\mathcal{L}_{\mathtt{adv}} = \mathcal{L}_{\mathtt{AT}}$ (2), the final objective becomes:

$$\mathcal{L}_{\mathtt{total}} = \frac{1}{2}\sum_{i=1}^{2} \max_{\|\delta_i\|_p \le \epsilon} \mathcal{L}_{\mathtt{CE}}\big(f_\theta(T_i(x) + \delta_i), y\big) + \lambda \cdot \mathtt{JS}\Big(\hat{f}_\theta\big(T_1(x) + \delta_1; \tau\big) \parallel \hat{f}_\theta\big(T_2(x) + \delta_2; \tau\big)\Big).$$

We introduce explicit forms of other variants of final objective $\mathcal{L}_{\mathtt{total}}$ for other AT methods, *e.g.*, TRADES (Zhang et al. 2019) and MART (Wang et al. 2020), integrated with our regularization loss, in the supplementary material.

Experiments

We verify the effectiveness of our technique on image classification datasets: CIFAR-10/100 (Krizhevsky and Hinton 2009) and Tiny-ImageNet. Our results exhibit that incorporating simple consistency regularization scheme into the existing adversarial training (AT) methods significantly improve adversarial robustness against various attacks (Carlini and Wagner 2017; Madry et al. 2018; Croce and Hein 2020b), including data corruption (Hendrycks and Dietterich 2019). Intriguingly, our method shows better robustness against *unseen* adversaries compared to other baselines. Moreover, our method surpass the performance of the recent regularization technique (Wu, Xia, and Wang 2020). Finally, we perform an ablation study to validate each component of our approach.

Experimental Setups

Training details. We use PreAct-ResNet-18 (He et al. 2016b) architecture in all experiments, and additionally use WideResNet-34-10 (Zagoruyko and Komodakis 2016) for white-box adversarial defense on CIFAR-10. For the data augmentation, we consider AutoAugment (Cubuk et al. 2019) where random crop (with 4 pixels zero padding), random horizontal flip (with 50% of probability), and Cutout (DeVries and Taylor 2017) (with half of the input width) are included. We set the regularization parameter $\lambda = 1.0$ in all cases except for applying on WideResNet-34-10 with TRADES and MART where we use $\lambda = 2.0$. The temperature is fixed to $\tau = 0.5$ in all experiments.

For other training setups, we mainly follow the hyperparameters suggested by the previous studies (Pang et al. 2021; Rice, Wong, and Kolter 2020). In detail, we train the network for 200 epochs[3] using stochastic gradient descent with momentum 0.9, and weight decay of 0.0005. The learning rate starts at 0.1 and is dropped by a factor of 10 at 50%, and 75% of the training progress. For the inner maximization for all AT, we set the $\epsilon = 8/255$, step size $2/255$, and 10 number of steps with l_∞ constraint (see the supplementary material for the l_2 constraint AT results).

Throughout the section, we mainly report the results where the clean accuracy converges, *i.e.*, fully trained model, to focus on the robust overfitting problem (Rice, Wong, and Kolter 2020). Nevertheless, we also note that our regularization method achieves better best robust accuracy compare to the AT methods (see Table 2).

Main Results

White-box attack. We consider a wide range of white-box adversarial attacks, in order to extensively measure the robustness of trained models without gradient obfuscation (Athalye, Carlini, and Wagner 2018): PGD (Madry et al. 2018) with 20 and 100 iterations (step size with $2\epsilon/k$, where k is the iteration number), CW_∞ (Carlini and Wagner 2017), and AutoAttack (Croce and Hein 2020b).[4] We report the fully

[3] Our method maintains almost the same robust accuracy under the same computational budget to the baselines: reduce the training steps in half. See the supplementary material for more discussion.

[4] We regard AutoAttack as a white-box attack, while it both includes white-box and black-box attacks. See the supplementary

Dataset (Architecture)	Method	Clean	PGD-20	PGD-100	CW_∞	AutoAttack
CIFAR-10 (PreAct-ResNet-18)	Standard (Madry et al. 2018)	84.57 (83.43)	45.04 (52.82)	44.86 (52.67)	44.31 (50.66)	40.43 (47.63)
	+ Consistency	**86.45** (85.25)	**56.51** (57.53)	**56.38** (57.39)	**52.45** (52.70)	**48.57** (49.05)
	TRADES (Zhang et al. 2019)	82.87 (82.13)	50.95 (53.98)	50.83 (53.85)	49.30 (51.71)	46.32 (49.32)
	+ Consistency	**83.63** (83.55)	**55.00** (55.16)	**54.89** (54.98)	**49.91** (50.67)	**47.68** (49.01)
	MART (Wang et al. 2020)	82.63 (77.00)	51.12 (54.83)	50.91 (54.74)	46.92 (49.26)	43.46 (46.74)
	+ Consistency	**83.43** (81.89)	**59.59** (60.48)	**59.52** (60.47)	**51.78** (51.83)	**48.91** (48.95)
CIFAR-10 (WideResNet-34-10)	Standard (Madry et al. 2018)	86.37 (87.55)	50.16 (55.86)	49.80 (55.65)	49.25 (54.45)	45.62 (51.24)
	+ Consistency	**89.82** (89.93)	**58.63** (61.11)	**58.41** (60.99)	**56.38** (57.80)	**52.36** (54.08)
	TRADES (Zhang et al. 2019)	85.05 (84.30)	51.20 (57.34)	50.89 (57.20)	50.88 (55.08)	46.17 (53.02)
	+ Consistency	**87.71** (87.92)	**58.39** (59.12)	**58.19** (58.99)	**54.84** (55.97)	**51.94** (53.11)
	MART (Wang et al. 2020)	85.75 (83.98)	49.31 (57.28)	49.06 (57.22)	48.05 (53.21)	44.96 (50.62)
	+ Consistency	**87.17** (85.81)	**63.26** (64.95)	**62.81** (64.80)	**57.46** (56.24)	**52.41** (53.33)
CIFAR-100 (PreAct-ResNet-18)	Standard (Madry et al. 2018)	57.13 (57.10)	22.36 (29.67)	22.25 (29.65)	21.97 (27.99)	19.85 (25.38)
	+ Consistency	**62.73** (61.62)	**30.75** (32.33)	**30.62** (32.24)	**27.63** (28.39)	**24.55** (25.52)
Tiny-ImageNet (PreAct-ResNet-18)	Standard (Madry et al. 2018)	41.54 (45.26)	11.71 (20.92)	11.60 (20.87)	11.20 (18.72)	9.63 (16.03)
	+ Consistency	**50.15** (49.46)	**21.33** (23.31)	**21.24** (23.24)	**19.08** (20.29)	**15.69** (16.90)

Table 2: Clean accuracy and robust accuracy (%) against white-box attacks of networks trained on various image classification benchmark datasets. All threat models are l_∞ with $\epsilon = 8/255$. Values in parenthesis denote the result of the checkpoint with the best PGD-10 accuracy, where each checkpoint is saved per epoch. We compare with the baselines trained under random crop and flip. The bold indicates the improved results by our proposed loss.

trained model's accuracy and the result of the checkpoint with the best PGD accuracy (of 10 iterations), where each checkpoint is saved per epoch.

As shown in Table 2, incorporating our regularization scheme into existing AT methods consistently improves both best and last white-box accuracies against various adversaries across different models and datasets. The results also demonstrates that our method effectively prevents robust overfitting as the gap between the best and last accuracies has been significantly reduced in all cases. In particular, for TRADES with WideResNet-34-10, our method's robust accuracy gap under AutoAttack is only 1.17%, while the baseline's gap is 6.85%, which is relatively 6 times smaller. More intriguingly, consideration of our regularization technique into the AT methods boosts the clean accuracy as well in all cases. We notice that such improvement is non-trivial, as some works have reported a trade-off between a clean and robust accuracies in AT (Tsipras et al. 2019; Zhang et al. 2019).

Unseen adversaries. We also evaluate our method against *unforeseen* adversaries, *e.g.*, robustness on different attack radii of ϵ, or even on different norm constraints of l_2 and l_1, as reported in Table 3. We observe that combining our regularization method could consistently and significantly improve the robustness against all the considered unseen adversaries tested. It is remarkable that our method is especially effective against l_1 adversaries compared to the baselines, regarding the fundamental difficulty of achieving the mutual robustness against both l_1 and l_∞ attacks (Tramer and Boneh 2019; Croce and Hein 2020a). Hence, we believe our regularization scheme can also be adapted to AT methods for training robust classifiers against multiple perturbations (Tramer and Boneh 2019; Maini, Wong, and Kolter 2020).

Common corruption. We also validate the effectiveness of our method on corrupted CIFAR-10 dataset (Hendrycks and Dietterich 2019), *i.e.*, consist of 19 types of corruption such as snow, zoom blur. We report the mean corruption error (mCE) of each model in Table 4. The results show that the mCE consistently improves combined with our regularization loss regardless of AT methods. Interestingly, our method even reduces the error (from the standard cross-entropy training) of corruptions that are not related to the applied augmentation or noise, *e.g.*, zoom blur error 25.8%→19.8%. We note that common corruption is also important and practical defense scenario (Hendrycks and Dietterich 2019), therefore, obtaining such robustness should be a desirable property for a robust classifier.

Comparison with Wu, Xia, and Wang (2020)

In this section, we consider a comparison with Adversarial weight perturbation (AWP) (Wu, Xia, and Wang 2020)[5], another recent work which also addresses the overfitting problem of AT by regularizing the flatness of the loss landscape with respect to weights via an adversarial perturbation on both input and weights. We present two experimental scenarios showing that our method can work better than AWP.

White-box attack and unseen adversaries. We consider various white-box attacks and unseen adversaries for measuring the robustness. As shown in Table 5, our method shows better results than AWP in l_∞ defense in most cases, and

material for black-box transfer attack results. We use the official code for the AutoAttack: https://github.com/fra31/auto-attack.

[5]We use the official code: https://github.com/csdongxian/AWP

Dataset	Method \ ϵ	l_∞		l_2		l_1	
		$4/255$	$16/255$	$150/255$	$300/255$	$2000/255$	$4000/255$
CIFAR-10	Standard (Madry et al. 2018)	65.93	19.23	52.56	25.68	45.96	36.85
	+ Consistency	**73.74**	**23.47**	**65.81**	**36.87**	**58.66**	**50.79**
	TRADES (Zhang et al. 2019)	68.30	24.17	56.14	28.94	44.08	29.58
	+ Consistency	**70.33**	**26.52**	**63.70**	**39.16**	**56.48**	**48.32**
	MART (Wang et al. 2020)	67.76	23.36	57.17	30.98	46.61	34.63
	+ Consistency	**72.67**	**30.31**	**66.17**	**43.76**	**60.57**	**54.19**
CIFAR-100	Standard (Madry et al. 2018)	36.14	7.37	27.97	11.98	30.48	27.29
	+ Consistency	**46.11**	**11.53**	**39.77**	**20.69**	**36.04**	**32.75**
Tiny-ImageNet	Standard (Madry et al. 2018)	23.23	2.69	28.05	17.80	33.30	31.55
	+ Consistency	**34.18**	**5.74**	**40.06**	**30.62**	**43.90**	**42.65**

Table 3: Robust accuracy (%) of PreAct-ResNet-18 trained with l_∞ of $\epsilon = 8/255$ constraint against unseen attacks. For unseen attacks, we use PGD-100 under different sized l_∞ balls, and other types of norm ball, *e.g.*, l_1, l_2. We compare with the baselines trained under random crop and flip. The bold indicates the improved results by the proposed method.

Method	mCE ↓
Standard cross-entropy	27.02
Standard (Madry et al. 2018)	24.03
+ Consistency	**21.83**
TRADES (Zhang et al. 2019)	25.50
+ Consistency	**23.95**
MART (Wang et al. 2020)	26.20
+ Consistency	**24.41**

Table 4: Mean corruption error (mCE) (%) of PreAct-ResNet-18 trained on CIFAR-10, and tested with CIFAR-10-C dataset (Hendrycks and Dietterich 2019). The arrow on the right side of the evaluation metric indicates the descending order of the value is better. We compare with the baselines trained under random crop and flip. The bold indicates the improved results by the proposed method.

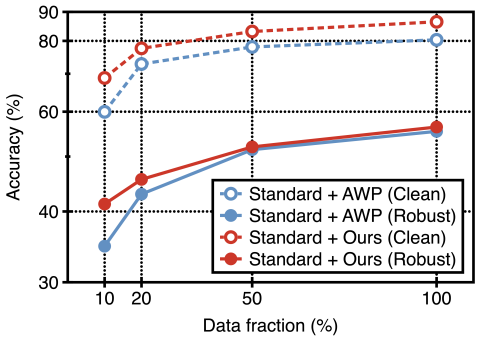


Figure 3: Clean accuracy and robust accuracy (%) against PGD-100 attack of l_∞ with $\epsilon = 8/255$, under different fraction (%) of CIFAR-10. We train PreAct-ResNet-18 with AWP (Wu, Xia, and Wang 2020) and consistency regularization loss based on standard AT (Madry et al. 2018).

outperforms in all cases of unseen adversaries defense, *e.g.*, l_2, l_1 constraint attack. In particular, our regularization technique consistently surpass AWP in the defense against the l_1 constraint attack. In addition, our method shows consistent improvement in clean accuracy, while AWP somewhat suffers from the trade-off between clean and robust accuracy.

Training with limited data. We also demonstrate that our method is data-efficient: when only a small number of training points are accessible for training the classifier. To this end, we reduce the training dataset's fraction to 10%, 20%, and 50% and train the classifier in each situation. As shown in Figure 3, our method shows better results compare to AWP, especially learning from the small sized dataset, as our method efficiently incorporates the rice space of data augmentations. In particular, our method obtained 41.2% robust accuracy even in the case when only 10% of the total dataset is accessible (where AWP achieves 34.7%). We note such efficiency is worthy for practitioners, since in such cases, validation dataset for early stopping is insufficient.

Ablation Study

We perform an ablation study on each of the components in our method. Throughout the section, we apply our method to the standard AT (Madry et al. 2018) and use PGD with 100 iterations for the evaluation. We also provide more analysis on the temperature hyperparameter and design choice of consistency regularization in the supplementary material.

Component analysis. We perform an analysis on each component of our method, namely the use of (a) data augmentations, and (b) the consistency regularization loss, by comparing their robust accuracy and mean corruption error (mCE). The results in Table 6 demonstrate each component is indeed effective, as the performance improves step by step with the addition of the component. We note that the proposed regularization method could not only improve the robust accuracy but also significantly improve the mCE. As shown in Figure 4, simply applying augmentation to the standard AT can reduce the error in many cases (13 types out of 19 corruptions) and even reduce the error of corruptions that are not

Dataset	Method	Clean	l_∞ (Seen)			l_2 (Unseen)		l_1 (Unseen)	
			PGD-100 (8/255)	CW_∞ (8/255)	AutoAttack (8/255)	PGD-100 (150/255)	PGD-100 (300/255)	PGD-100 (2000/255)	PGD-100 (4000/255)
CIFAR-10	Standard (Madry et al. 2018)	84.57	44.86	44.31	40.43	52.56	25.68	45.96	36.85
	+ AWP (Wu, Xia, and Wang 2020)	80.34	55.39	52.31	**49.60**	61.39	36.05	56.30	48.37
	+ Consistency	**86.45**	**56.38**	**52.45**	48.57	**65.81**	**36.87**	**58.66**	**50.79**
CIFAR-100	Standard (Madry et al. 2018)	56.96	20.86	21.20	18.93	27.65	11.08	26.49	21.48
	+ AWP (Wu, Xia, and Wang 2020)	52.91	30.06	26.42	24.32	35.71	20.18	33.63	30.38
	+ Consistency	**62.73**	**30.62**	**27.63**	**24.55**	**39.77**	**20.69**	**36.04**	**32.75**
Tiny-ImageNet	Standard (Madry et al. 2018)	41.54	11.60	11.20	9.63	28.05	17.80	33.30	31.55
	+ AWP (Wu, Xia, and Wang 2020)	40.25	20.64	18.05	15.26	33.31	26.86	35.48	34.22
	+ Consistency	**50.15**	**21.24**	**19.08**	**15.69**	**40.06**	**30.62**	**43.90**	**42.65**

Table 5: Clean accuracy and robust accuracy (%) against diverse attacks of each individual, and combined regularization. The numbers below the attack methods, indicate the radius of the perturbation ϵ. All results are reported on PreAct-ResNet-18 trained under various image classification benchmark datasets. The bold indicates the best results.

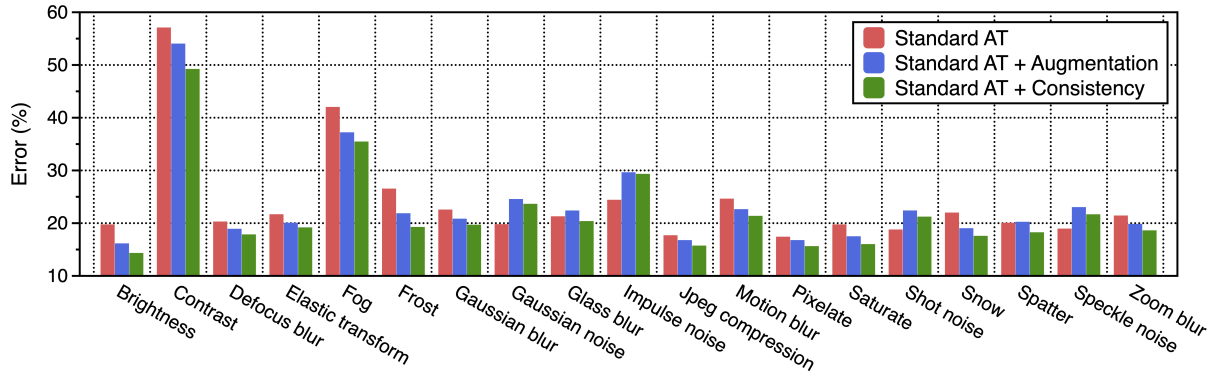


Figure 4: Classification error (%) on each corruption type of CIFAR-10-C (Hendrycks and Dietterich 2019) where the x-axis labels denote the corruption type. Reported values are measured on PreAct-ResNet-18 trained under standard AT (Madry et al. 2018), standard AT with AutoAugment (Cubuk et al. 2019), standard AT with consistency regularization, respectively.

Method	PGD-100	mCE ↓
Standard (Madry et al. 2018)	44.86	24.03
+ Cutout (DeVries and Taylor 2017)	49.95	24.05
+ AutoAugment (Cubuk et al. 2019)	55.18	23.38
+ Consistency	**56.38**	**22.06**

Table 6: Ablation study on each component of our proposed training objective. Reported values are the robust accuracy (%) against PGD-100 attack of l_∞ with $\epsilon = 8/255$, and mean corruption error (mCE) (%) of PreAct-ResNet-18 under CIFAR-10. The bold indicates the best result.

related to the applied augmentation (*e.g.*, motion blur, zoom blur). More interestingly, further adapting the consistency regularization loss can reduce the corruption error in all cases from the standard AT with augmentation. It suggests that the consistency prior is indeed a desirable property for classifiers to obtain robustness (for both adversarial and corruption).

Analysis on attack directions. To analyze the effect of our regularization scheme, we observe the attacked directions of the adversarial examples. We find that the most confusing class of the 'clean' input, is highly like to be attacked. Formally, we define the most confusing class of the given sample (x, y) as $\arg\max_{k \neq y} f_\theta^{(k)}(x)$ where $f_\theta^{(k)}$ is the softmax probability of class k. We observe that 77.45% out of the misclassified adversarial examples predicts the most confusing class. This result implies that the attack direction itself contains the dark knowledge of the given input (Hinton, Vinyals, and Dean 2015), which supports our intuition to match the attack direction.

Conclusion

In this paper, we propose a simple yet effective regularization technique to tackle the robust overfitting in adversarial training (AT). Our regularization forces the predictive distributions after attacking from two different augmentations of the same input to be similar to each other. Our experimental results demonstrate that the proposed regularization brings significant improvement in various defense scenarios including unseen adversaries.

Acknowledgements

We thank Jaeho Lee, Sangwoo Mo, and Soojung Yang for providing helpful feedbacks and suggestions in preparing an earlier version of the manuscript. This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2019-0-00075, Artificial Intelligence Graduate School Program(KAIST)) and the Engineering Research Center Program through the National Research Foundation of Korea (NRF) funded by the Korean Government MSIT (NRF-2018R1A5A1059921).

References

Amodei, D.; Ananthanarayanan, S.; Anubhai, R.; Bai, J.; Battenberg, E.; Case, C.; Casper, J.; Catanzaro, B.; Cheng, Q.; Chen, G.; et al. 2016. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International Conference on Machine Learning*.

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

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

Chen, C.; Seff, A.; Kornhauser, A.; and Xiao, J. 2015. Deepdriving: Learning affordance for direct perception in autonomous driving. In *IEEE International Conference on Computer Vision*.

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

Cohen, J.; Rosenfeld, E.; and Kolter, Z. 2019. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*.

Croce, F.; and Hein, M. 2020a. Provable robustness against all adversarial l_p-perturbations for $p \geq 1$. In *International Conference on Learning Representations*.

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

Cubuk, E. D.; Zoph, B.; Mane, D.; Vasudevan, V.; and Le, Q. V. 2019. Autoaugment: Learning augmentation strategies from data. In *IEEE Conference on Computer Vision and Pattern Recognition*.

DeVries, T.; and Taylor, G. W. 2017. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*.

Girshick, R. 2015. Fast r-cnn. In *IEEE International Conference on Computer Vision*.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*.

Gowal, S.; Qin, C.; Uesato, J.; Mann, T.; and Kohli, P. 2020. Uncovering the Limits of Adversarial Training against Norm-Bounded Adversarial Examples. *arXiv preprint arXiv:2010.03593*.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. In *International Conference on Machine Learning*.

Guo, C.; Rana, M.; Cisse, M.; and Van Der Maaten, L. 2018. Countering adversarial images using input transformations. In *International Conference on Learning Representations*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016a. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016b. Identity mappings in deep residual networks. In *European Conference on Computer Vision*.

Hendrycks, D.; and Dietterich, T. 2019. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In *International Conference on Learning Representations*.

Hendrycks, D.; Mu, N.; Cubuk, E. D.; Zoph, B.; Gilmer, J.; and Lakshminarayanan, B. 2020. Augmix: A simple data processing method to improve robustness and uncertainty. In *International Conference on Learning Representations*.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Jeong, J.; and Shin, J. 2020. Consistency Regularization for Certified Robustness of Smoothed Classifiers. In *Advances in Neural Information Processing Systems*.

Kim, M.; Tack, J.; and Hwang, S. J. 2020. Adversarial Self-Supervised Contrastive Learning. In *Advances in Neural Information Processing Systems*.

Krizhevsky, A.; and Hinton, G. 2009. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario.

Kurakin, A.; Goodfellow, I.; and Bengio, S. 2016. Adversarial examples in the physical world. arXiv:1607.02533.

Lee, K.; Lee, K.; Lee, H.; and Shin, J. 2018. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*.

Li, J.; Schmidt, F.; and Kolter, Z. 2019. Adversarial camera stickers: A physical camera-based attack on deep learning systems. In *International Conference on Machine Learning*.

Liao, F.; Liang, M.; Dong, Y.; Pang, T.; Hu, X.; and Zhu, J. 2018. Defense against adversarial attacks using high-level representation guided denoiser. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Ma, X.; Li, B.; Wang, Y.; Erfani, S. M.; Wijewickrema, S.; Schoenebeck, G.; Song, D.; Houle, M. E.; and Bailey, J. 2018. Characterizing adversarial subspaces using local intrinsic dimensionality. In *International Conference on Learning Representations*.

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

Maini, P.; Wong, E.; and Kolter, Z. 2020. Adversarial robustness against the union of multiple perturbation models. In *International Conference on Machine Learning*.

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

Qin, C.; Martens, J.; Gowal, S.; Krishnan, D.; Dvijotham, K.; Fawzi, A.; De, S.; Stanforth, R.; and Kohli, P. 2019. Adversarial robustness through local linearization. In *Advances in Neural Information Processing Systems*.

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

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

Sohn, K.; Berthelot, D.; Li, C.-L.; Zhang, Z.; Carlini, N.; Cubuk, E. D.; Kurakin, A.; Zhang, H.; and Raffel, C. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems*.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2014. Intriguing properties of neural networks. In *International Conference on Learning Representations*.

Tramer, F.; and Boneh, D. 2019. Adversarial training and robustness for multiple perturbations. In *Advances in Neural Information Processing Systems*.

Tsipras, D.; Santurkar, S.; Engstrom, L.; Turner, A.; and Madry, A. 2019. Robustness May Be at Odds with Accuracy. In *International Conference on Learning Representations*.

Wang, Y.; Zou, D.; Yi, J.; Bailey, J.; Ma, X.; and Gu, Q. 2020. Improving Adversarial Robustness Requires Revisiting Misclassified Examples. In *International Conference on Learning Representations*.

Wu, D.; Xia, S.-T.; and Wang, Y. 2020. Adversarial Weight Perturbation Helps Robust Generalization. In *Advances in Neural Information Processing Systems*.

Xie, Q.; Dai, Z.; Hovy, E.; Luong, M.-T.; and Le, Q. V. 2020. Unsupervised data augmentation for consistency training. In *Advances in Neural Information Processing Systems*.

Yang, Y.-Y.; Rashtchian, C.; Zhang, H.; Salakhutdinov, R.; and Chaudhuri, K. 2020. A Closer Look at Accuracy vs. Robustness. In *Advances in Neural Information Processing Systems*.

Zagoruyko, S.; and Komodakis, N. 2016. Wide residual networks. In *British Machine Vision Conference*.

Zhang, H.; Yu, Y.; Jiao, J.; Xing, E.; Ghaoui, L. E.; and Jordan, M. 2019. Theoretically Principled Trade-off between Robustness and Accuracy. In *International Conference on Machine Learning*.

Zhang, H.; Zhang, Z.; Odena, A.; and Lee, H. 2020. Consistency regularization for generative adversarial networks. In *International Conference on Learning Representations*.