Understanding Catastrophic Overfitting in Single-step Adversarial Training

Hoki Kim*, **Woojin Lee***, **Jaewook Lee**†
Seoul National University, Seoul, Korea
ghrl9613@snu.ac.kr, wj926@snu.ac.kr, jaewook@snu.ac.kr

Abstract

Although fast adversarial training has demonstrated both robustness and efficiency, the problem of "catastrophic overfitting" has been observed. This is a phenomenon in which, during single-step adversarial training, robust accuracy against projected gradient descent (PGD) suddenly decreases to 0% after a few epochs, whereas robust accuracy against fast gradient sign method (FGSM) increases to 100%. In this paper, we demonstrate that catastrophic overfitting is very closely related to the characteristic of single-step adversarial training which uses only adversarial examples with the maximum perturbation, and not all adversarial examples in the adversarial direction, which leads to decision boundary distortion and a highly curved loss surface. Based on this observation, we propose a simple method that not only prevents catastrophic overfitting, but also overrides the belief that it is difficult to prevent multi-step adversarial attacks with single-step adversarial training.

1 Introduction

Adversarial examples are perturbed inputs that are designed to deceive machine learning classifiers by adding adversarial noises to the original data. Although such perturbations are sufficiently subtle and undetectable by humans, they result in an incorrect classification. Since deep-learning models were found to be vulnerable to adversarial examples (Szegedy et al. 2013), a line of work was proposed to mitigate the problem and improve robustness of the models. Among the numerous defensive methods, projected gradient descent (PGD) adversarial training (Madry et al. 2017) is one of the most successful approaches for achieving robustness against adversarial attacks. Although PGD adversarial training serves as a strong defensive algorithm, because it relies on a multi-step adversarial attack, a high computational cost is required for multiple forward and back propagation during batch training.

To overcome this issue, other studies (Shafahi et al. 2019; Wong, Rice, and Kolter 2020) on reducing the computational burden of adversarial training using single-step adversarial attacks (Goodfellow, Shlens, and Szegedy 2014) have been proposed. Among them, inspired by Shafahi et al. (2019), Wong, Rice, and Kolter (2020) suggested fast adversarial training, which is a modified version of fast gradient sign method (FGSM) adversarial training designed to be as effective as PGD adversarial training.

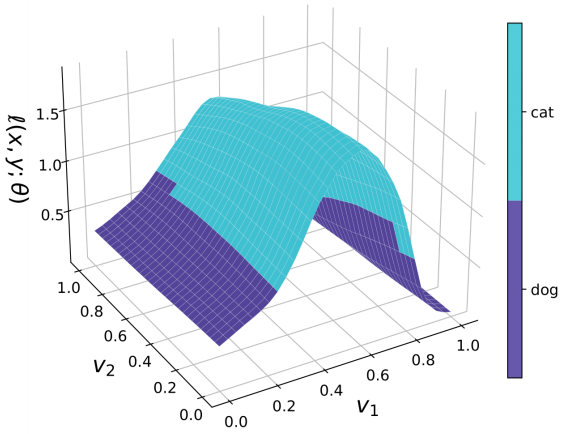


Figure 1: Visualization of distorted decision boundary. The origin indicates the original image x, the label of which is "dog". In addition, v_1 is the direction of a single-step adversarial perturbation and v_2 is a random direction. The adversarial image $x+v_1$ is classified as the correct label, although there is distorted interval where $x + k \cdot v_1$ is misclassified even when k is less than 1. Due to this decision boundary distortion, single-step adversarial training becomes vulnerable to multi-step adversarial attacks.

Fast adversarial training has demonstrated both robustness and efficiency; however, it suffers from the problem of "catastrophic overfitting," which is a phenomenon that robustness against PGD suddenly decreases to 0%, whereas robustness against FGSM rapidly increases. Wong, Rice, and Kolter (2020) first discovered this issue and suggested the use of early stopping to prevent it. Later, it was found that catastrophic overfitting also occurs in different single-step adversarial training methods such as free adversarial training (Andriushchenko and Flammarion 2020).

*Equal contribution.
†Corresponding author.

In this regard, few attempts have been made to discover the underlying reason for catastrophic overfitting and methods proposed to prevent this failure (Andriushchenko and Flammarion 2020; Vivek and Babu 2020; Li et al. 2020). However, these approaches were computationally inefficient or did not provide a fundamental reason for the problem.

In this study, we first analyze the differences before and after catastrophic overfitting. We then identify the relationship between distortion of the decision boundary and catastrophic overfitting. Unlike the previous notion in which a larger perturbation implies a stronger attack, we discover that sometimes a smaller perturbation is sufficient to fool the model, whereas the model is robust against larger perturbations during single-step adversarial training. We call this phenomenon "decision boundary distortion."

Figure 1 shows an example of decision boundary distortion by visualizing the loss surface. The model is robust to perturbations when the magnitude of the attack is equal to the maximum perturbation ϵ, but not to other smaller perturbations. When decision boundary distortion occurs, the model becomes more robust against a single-step adversarial attack but reveals fatal weaknesses to multi-step adversarial attacks and leads to catastrophic overfitting.

Through extensive experiments, we empirically discovered the relationship between single-step adversarial training and decision boundary distortion, and found that the problem of single-step adversarial training is a fixed magnitude of the perturbation, not the direction of the attack. Based on this observation, we present a simple algorithm that determines the appropriate magnitude of the perturbation for each image and prevents catastrophic overfitting.

Contributions.

- We discovered a "decision boundary distortion" phenomenon that occurs during single-step adversarial training and the underlying connection between decision boundary distortion and catastrophic overfitting.
- We suggest a simple method that prevents decision boundary distortion by searching the appropriate step size for each image. This method not only prevents catastrophic overfitting, but also achieves near 100% accuracy for the training examples against PGD.
- We evaluate robustness of the proposed method against various adversarial attacks (FGSM, PGD, and AutoAttack (Croce and Hein 2020)) and demonstrate the proposed method can provide sufficient robustness without catastrophic overfitting.

2 Background and Related Work

2.1 Adversarial Robustness

There are two major movements for building a robust model: provable defenses and adversarial training.

A considerable number of studies related to provable defenses of deep-learning models have been published. Provable defenses attempt to provide provable guarantees for robust performance, such as linear relaxations (Wong and Kolter 2018; Zhang et al. 2019a), interval bound propagation (Gowal et al. 2018; Lee, Lee, and Park 2020), and randomized smoothing (Cohen, Rosenfeld, and Kolter 2019; Salman et al. 2019). However, provable defenses are computationally inefficient and show unsatisfied performance compared to adversarial training.

Adversarial training is an approach that augments adversarial examples generated by adversarial attacks (Goodfellow, Shlens, and Szegedy 2014; Madry et al. 2017; Tramèr et al. 2017). Because this approach is simple and achieves high empirical robustness for various attacks, it has been widely used and developed along with other deep learning methods such as mix-up (Zhang et al. 2017; Lamb et al. 2019; Pang, Xu, and Zhu 2019) and unsupervised training (Alayrac et al. 2019; Najafi et al. 2019; Carmon et al. 2019).

In this study, we focus on adversarial training. Given an example $(x, y) \sim \mathcal{D}$, let $\ell(x, y; \theta) = \ell(f_\theta(x), y)$ denote the loss function of a deep learning model f with parameters θ. Then, adversarial training with a maximum perturbation ϵ can be formalized as follows:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}}\big[\max_{\delta\in\mathcal{B}(x,\epsilon)} \ell(x+\delta, y; \theta)\big] \tag{1}$$

A perturbation δ is in $\mathcal{B}(x, \epsilon)$ that denotes the ϵ-ball around an example x with a specific distance measure. The most used distance measures are L_0, L_2, and L_∞. In this study, we use L_∞ for such a measure.

However, the above optimization is considered as NP-hard because it contains a non-convex min-max problem. Thus, instead of the inner maximization problem, adversarial attacks are used to find the perturbation δ.

Fast gradient sign method (FGSM) (Goodfellow, Shlens, and Szegedy 2014) is the simplest adversarial attack, which uses a sign of a gradient to find an adversarial image x'. Because FGSM requires only one gradient, it is considered the least expensive adversarial attack (Goodfellow, Shlens, and Szegedy 2014; Madry et al. 2017).

$$x' = x + \epsilon \cdot \mathrm{sgn}(\nabla_x \ell(x, y; \theta)) \tag{2}$$

Projected gradient descent (PGD) (Madry et al. 2017) uses multiple gradients to generate more powerful adversarial examples. With a step size α, PGD can be formalized as follows:

$$x^{t+1} = \Pi_{\mathcal{B}(x,\epsilon)}(x^t + \alpha \cdot \mathrm{sgn}(\nabla_x \ell(x, y; \theta))) \tag{3}$$

where $\Pi_{\mathcal{B}(x,\epsilon)}$ refers the projection to the ϵ-ball $\mathcal{B}(x, \epsilon)$. Here, x^t is an adversarial example after t-steps. A large number of steps allows us to explore more areas in $\mathcal{B}(x, \epsilon)$. Note that PGDn corresponds to PGD with n steps (or iterations). For instance, PGD7 indicates that the number of PGD steps is 7.

2.2 Single-step Adversarial Attack versus Multi-step Adversarial Attack

Single-step adversarial training was previously believed to be a non-robust method because it produces nearly 0% accuracy against PGD (Madry et al. 2017). Moreover, the model trained using FGSM has been confirmed to have typical

characteristics, such as gradient masking, which indicates that a single-step gradient is insufficient to find a decent adversarial examples (Tramèr et al. 2017). For the above reasons, a number of studies have been conducted on multi-step adversarial attacks.

Contrary to this perception, however, free adversarial training (Shafahi et al. 2019) has achieved a remarkable performance with a single-step gradient using redundant batches and accumulative perturbations. Following Shafahi et al. (2019), Wong, Rice, and Kolter (2020) proposed fast adversarial training using FGSM with a uniform random initialization. Fast adversarial training shows an almost equivalent performance to those of PGD (Madry et al. 2017) and free adversarial training (Shafahi et al. 2019).

$$\begin{aligned} \eta &= \text{Uniform}(-\epsilon, \epsilon) \\ \delta &= \eta + \alpha \cdot \text{sgn}(\nabla_\eta \ell(x + \eta, y; \theta)) \\ x' &= x + \delta \end{aligned} \tag{4}$$

2.3 Catastrophic Overfitting

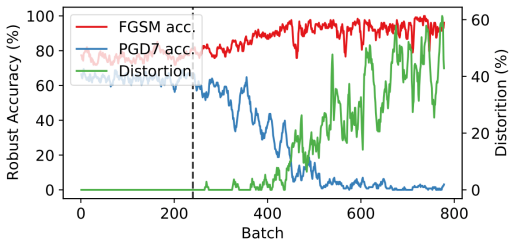
Although fast adversarial training performs well in a short time, a previously undiscovered phenomenon has been identified. That is, after a few epochs with single-step adversarial training, robustness of the model against PGD decreases sharply. This phenomenon is called catastrophic overfitting. Fast adversarial training (Wong, Rice, and Kolter 2020) uses early stopping to temporally avoid catastrophic overfitting by tracking robustness accuracy against PGD on the training batches.

To apply early stopping, robustness against PGD must be continuously confirmed. Furthermore, standard accuracy does not yield the maximum potential (Andriushchenko and Flammarion 2020). To resolve these shortcomings and gain a deeper understanding of catastrophic overfitting, a line of work has been proposed. Vivek and Babu (2020) identified that catastrophic overfitting arises with early overfitting to FGSM. To prevent this type of overfitting, the authors introduced dropout scheduling and demonstrated stable adversarial training for up to 100 epochs. In addition, Li et al. (2020) trained a model with FGSM at first and then changed it into PGD when there was a large decrease in the PGD accuracy. Andriushchenko and Flammarion (2020) found that an abnormal behavior of a single filter leads to a nonlinear model with single-layer convolutional networks. Based on this observation, they proposed a regularization method, GradAlign, which maximizes $\cos(\nabla_x \ell(x, y; \theta), \nabla_x \ell(x + \eta, y; \theta))$ and prevents catastrophic overfitting by inducing a gradient alignment.

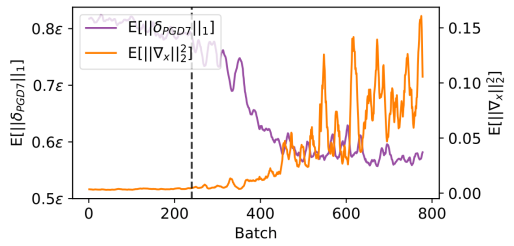
However, even with an increased understanding of catastrophic overfitting and methods for its prevention, a key question remains unanswered:

What characteristic of single-step adversarial attacks is the cause of catastrophic overfitting?

In this paper, we discuss the cause of catastrophic overfitting in the context of single-step adversarial training. We then propose a new simple method to facilitate stable single-step adversarial training, wherein longer training can produce a higher standard accuracy with sufficient adversarial robustness.



(a) Robust accuracy and distortion



(b) Mean of absolute value of PGD7 perturbations and L_2 norm of the gradients of the images

Figure 2: (CIFAR10) Analysis of catastrophic overfitting. Plot (a) shows robust accuracy of fast adversarial training against FGSM (red) and PGD7 (blue). Distortion (green) denotes the ratio of images in distorted interval in Equation (5). Plot (b) shows the mean of absolute value of PGD7 perturbation $\mathbb{E}[||\delta_{PGD7}||_1]$ (purple) and the L_2 norm of the gradients of the images $\mathbb{E}[||\nabla_x||_2]$ (orange). Dashed black lines correspond to the 240th batch, which is the start point of catastrophic overfitting in both plots.

3 Revisiting Catastrophic Overfitting

First, to analyze catastrophic overfitting, we start by recording robust accuracy of fast adversarial training on CIFAR-10 (Krizhevsky, Hinton et al. 2009). The maximum perturbation ϵ is fixed to $8/255$. We use FGSM and PGD7 to verify robust accuracy with the same settings $\epsilon = 8/255$ and a step size $\alpha = 2/255$.

Figure 2 shows statistics on the training batch when catastrophic overfitting occurs (71st out of 200 epochs). In plot (a), after 240 batches, robustness against PGD7 begins to decrease rapidly; conversely, robustness against FGSM increases. Plot (b) shows the mean of the absolute value of PGD7 perturbation $\mathbb{E}[||\delta_{PGD7}||_1]$ and squared L_2 norm of the gradient of the images $\mathbb{E}[||\nabla_x||_2]$ of each batch. After catastrophic overfitting, there is a trend of decreasing mean perturbation. This is consistent with the phenomenon in which the perturbations of the catastrophic overfitted model are located away from the maximum perturbation, unlike the model that is stopped early (Wong, Rice, and Kolter 2020). Concurrently, a significant increase in the squared L_2 norm of the gradient is also observed. The highest point indicates a large difference, approximately 35 times greater than that before catastrophic overfitting.

These two observations, a low magnitude of perturbations and a high gradient norm, make us wonder what would the

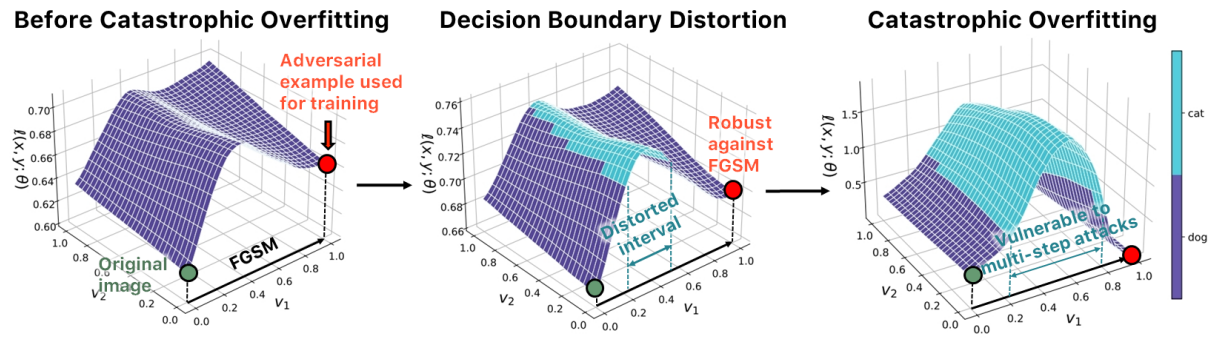


Figure 3: Process of normal decision boundary turns into distorted decision boundary. (Left) The loss surface before catastrophic overfitting with a FGSM adversarial direction v_1 and a random direction v_2. The red point denotes an adversarial example $x+v_1$ generated from the original image x, the label of which is "dog." (Middle) The changed loss surface after learning adversarial example $x + v_1$. Here, v_1 is the same vector as that on the left. Distorted interval begins to occur for the first time. (Right) As training continues, distorted decision boundary grows uncontrollably such that robustness against multi-step adversarial attacks decreases.

loss surface looks like. Figure 3 illustrates the progress of adversarial training in which catastrophic overfitting occurs. The loss surface of the perturbed example is shown, where the green spot denotes the original images and the red spot denotes the adversarial example used for adversarial training in the batch. The v_1 axis indicates the direction of FGSM, whereas the v_2 axis is a random direction. The true label of the original sample is "dog." Hence, the purple area indicates where the perturbed sample is correctly classified, whereas the blue area indicates a misclassified area.

On the left side of Figure 3, we can easily observe that the model is robust against FGSM. However, after training the batch, an interval vulnerable to a smaller perturbation than the maximum perturbation ϵ appears, whereas the model is still robust against FGSM. This distorted interval implies that the adversarial example with a larger perturbation is weaker than that with a smaller perturbation, which is contrary to the conventional belief that a larger magnitude of perturbation induces a stronger attack. As a result, the model with distorted interval is vulnerable to multi-step adversarial attacks that can search the vulnerable region further inside $\mathcal{B}(x, \epsilon)$. As the training continues, the area of distorted interval increases as shown in the figure on the right. It is now easier to see that the model is now perfectly overfitted for FGSM, yet loses its robustness to the smaller perturbations. We call this phenomenon "decision boundary distortion."

The evidence of decision boundary distortion is also shown in Figure 2 (b). When robustness against PGD7 sharply decreases to 0%, the mean of the absolute value of PGD7 perturbation $\mathbb{E}[||\delta_{PGD7}||_1]$ decreases. It indicates that, when catastrophic overfitting arises, a smaller perturbation is enough to fool the model than the maximum perturbation ϵ, which implies that distorted interval exists. In addition, during the process of having distorted decision boundary, as shown in the figure on the right, the loss surface inevitably becomes highly curved, which matches the observation of increasing the L_2 norm of the gradients of the images $\mathbb{E}[||\nabla_x||_2]$. This is also consistent with previous research (Andriushchenko and Flammarion 2020). Andriushchenko and Flammarion (2020) argued that $\nabla_x \ell(x, y; \theta)$ and $\nabla_x \ell(x + \eta, y; \theta)$ tend to be perpendicular in catastrophic overfitted models where η is drawn from a uniform distribution $U(-\epsilon, \epsilon)$. Considering that a highly curved loss surface implies $(\nabla_x \ell(x, y; \theta))^T (\nabla_x \ell(x + \eta, y; \theta)) \approx 0$ in high dimensions, the reason why GradAlign (Andriushchenko and Flammarion 2020) can avoid catastrophic overfitting might be because the gradient alignment leads the model to learn a linear loss surface which reduces the chance of having distorted decision boundary.

We next numerically measured the degree of decision boundary distortion. To do so, we first define a new measure distortion d. Given a deep learning model f and a loss function ℓ, distortion d can be formalized as follows:

$$\begin{aligned} \mathbf{S}_D &= \{x | \exists k \in (0,1) \text{ s.t. } f(x + k \cdot \epsilon \cdot \text{sgn}(\nabla_x \ell)) \neq y\} \\ \mathbf{S}_N &= \{x | f(x) = y, f(x + \epsilon \cdot \text{sgn}(\nabla_x \ell)) = y\} \\ d &= \frac{|\mathbf{S}_D \cap \mathbf{S}_N|}{|\mathbf{S}_N|} \end{aligned} \tag{5}$$

where (x, y) is an example drawn from dataset $\mathcal{D}$. However, because the loss function of the model is not known explicitly, we use a number of samples to estimate distortion d. In all experiments, we tested 100 samples in the adversarial direction $\delta = \epsilon \cdot \text{sgn}(\nabla_x \ell)$ for each example. Indeed, we can see that distortion increases in Figure 2 (a) when catastrophic overfitting arises.

To verify that decision boundary distortion is generally related to catastrophic overfitting, we demonstrate how distortion and robustness against PGD7 change during training. We conducted an experiment on five different models: fast adversarial training (Fast Adv.) (Wong, Rice, and Kolter 2020), PGD2 (PGD2 Adv.) (Madry et al. 2017), TRADES (Zhang et al. 2019b), GradAlign (Andriushchenko

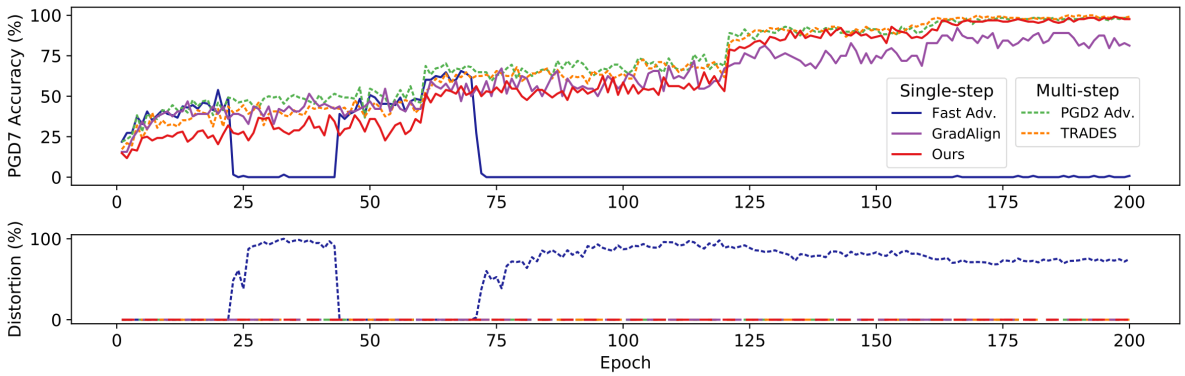


Figure 4: (CIFAR10) Robust accuracy and distortion on the training batch for each epoch. Two multi-step adversarial attacks show zero distortion during the entire training time and reach nearly 100% PGD7 accuracy. By contrast, fast adversarial training shows high distortion and eventually collapses after the 71st epoch. The proposed method successfully avoids such problems and achieves a high PGD7 accuracy similiar to multi-step adversarial training (Best viewed in color).

and Flammarion 2020), and the proposed method (Ours). All models were tested on $\epsilon = 8/255$. The step size α is set to $\alpha = 1.25\epsilon$, $\alpha = 1/2\epsilon$, and $\alpha = 1/4\epsilon$ for fast adversarial training, PGD2 adversarial training, and TRADES, respectively. We also conducted same experiment on PGD adversarial training with different number of steps; however, because these show similar results to PGD2 adversarial training, we only included PGD2. TRADES is trained with seven steps.

As the key observation in Figure 4, the point where decision boundary distortion begins in fast adversarial training (22nd epoch) is identical to the point where robustness against PGD7 sharply decreases; that is, catastrophic overfitting occurs. Then, when decision boundary distortion disappears (45th to 72nd epoch), the model immediately recovers robust accuracy. After the 72nd epoch, the model once again suffers a catastrophic overfitting and never regains its robustness with high distortion. Hence, we conclude that there is a close connection between decision boundary distortion and the vulnerability of the model against multi-step adversarial attacks.

4 Stable Single-Step Adversarial Training

Based on the results in Section 3, we assume that distorted decision boundary might be the reason for catastrophic overfitting. Here, we stress that the major cause of distorted decision boundary is that single-step adversarial training uses a point with a fixed distance ϵ from the original image x as an adversarial image x' instead of an optimal solution of the inner maximum in Equation (1). Under this linearity assumption, the most powerful adversarial perturbation δ would be the same as $\epsilon \cdot \text{sgn}(\nabla_x \ell)$ where ϵ is the maximum perturbation, and the following formula should be satisfied.

$$\begin{aligned} \ell(x+\delta) - \ell(x) &= (\nabla_x \ell)^T \delta \\ &= (\nabla_x \ell)^T \epsilon \cdot \text{sgn}(\nabla_x \ell) \\ &= \epsilon ||\nabla_x \ell||_1 \end{aligned} \tag{6}$$

However, as confirmed in the previous scetion, decision boundary distortion with a highly curved loss surface has been observed during the training phase, which indicates that ϵ is no longer the strongest adversarial step size in the direction of δ. Thus, the linear approximation of the inner maximization is not satisfied when distorted decision boundary arises.

To resolve this issue, we suggest a simple fix to prevent catastrophic overfitting by forcing the model to verify the inner interval of the adversarial direction. In this case, the appropriate magnitude of the perturbation should be taken into consideration instead of using ϵ:

$$\begin{aligned} &\delta = \epsilon \cdot \text{sgn}(\nabla_x \ell) \\ &\arg \max_{k \in [0,1]} \ell(x + k \cdot \delta, y; \theta) \end{aligned} \tag{7}$$

Here, we introduce k, which denotes the scaling parameter for the original adversarial direction $\text{sgn}(\nabla_x \ell)$. In contrast to previous single-step adversarial training which uses a fixed size of $k = 1$, an appropriate scaling parameter k^* helps the model to train stronger adversarial examples as follows:

$$\begin{aligned} &\delta = \epsilon \cdot \text{sgn}(\nabla_x \ell) \\ &k^* = \min_{k \in [0,1]} \{k | y \neq f(x + k \cdot \delta; \theta)\} \\ &\min_\theta \mathbb{E}_{(x,y) \sim \mathcal{D}} [\ell(x + k^* \cdot \delta, y; \theta)] \end{aligned} \tag{8}$$

In this way, regardless of the linearity assumption, we can train the model with stronger adversarial examples that induce an incorrect classification in the adversarial direction. Simultaneously, we can also detect distorted decision boundary by inspecting the inside of distorted interval, as shown in Figure 3.

However, because we do not know the explicit loss function of the model, forward propagation is the only approach for checking the adversarial images in the single-step adversarial attack direction. Hence, we propose the following simple method. First, we calculate the single-step adversarial direction δ. Next, we choose multiple checkpoints

Algorithm 1: Stable single-step adversarial training

Parameter: B mini-batches, a perturbation size ϵ, a step size α, and c check points for a network f_θ

for $i = 1, ..., B$ **do**
- $\eta = \text{Uniform}(-\epsilon, \epsilon)$
- $\hat{y}_{i,0} = f_\theta(x_i + \eta)$
- $\delta = \eta + \alpha \cdot \nabla_\eta \ell(\hat{y}_{i,0}, y_i)$
- **for** $j = 1, ..., c$ **do**
  - $\hat{y}_{i,j} = f_\theta(x_i + j \cdot \delta / c))$
- **end**
- $x'_i = x_i + \min(\{k | \hat{y}_{i,k} \neq y_i\} \cup \{1\}) \cdot \delta / c$
- $\theta = \theta - \nabla_\theta \ell(f_\theta(x'_i), y_i)$

end

$(x + \frac{1}{c}\delta, ..., x + \frac{c-1}{c}\delta, x + \delta)$. Here, c denotes the number of checkpoints except for the clean image x, which is tested in advance during the single-step adversarial attack process. We then feed all checkpoints to the model and verify that the predicted label $\hat{y}_j$ matches the correct label y for all checkpoints $x + \frac{j}{c}\delta$ where $j \in \{1, ..., c\}$. Among the incorrect images and the clean image x, the smallest j is selected; if all checkpoint are correctly classified, the adversarial image $x' = x + \delta$ is used. Algorithm 1 shows a summary of the proposed method.

Suppose the model has L layers with n neurons. Then, the time complexity of forward propagation is $O(Ln^2)$. Considering that backward propagation has the same time complexity, the generation of one adversarial example requires $O(2Ln^2)$ in total. Thus, with c checkpoints, the proposed method consumes $O((c+4)Ln^2)$ because it requires one adversarial direction $O(2Ln^2)$, forward propagation for c checkpoints $O(cLn^2)$, and one optimization step $O(2Ln^2)$. Compared to PGD2 adversarial training, which demands $O(6Ln^2)$, the proposed method requires more time when $c > 2$. However, the proposed method does not require additional memory for computing the gradients of the checkpoints because we do not need to track a history of variables for backward propagation; hence, larger validation batch sizes can be considered. Indeed, the empirical results describe in Section 5 indicate that the proposed method consumes less time than PGD2 adversarial training under $c \leq 4$.

Figure 4 shows that the proposed method successfully avoids catastrophic overfitting despite using a single-step adversarial attack. Furthermore, the proposed model not only achieves nearly 100% robustness against PGD7, which fast adversarial training cannot accomplish, but also possesses zero distortion until the end of the training. This is the opposite of the common understanding that single-step adversarial training methods cannot perfectly defend the model against multi-step adversarial attacks.

The proposed model learns the image with the smallest perturbation among the incorrect adversarial images. In other words, during the initial states, the model outputs incorrect predictions for almost every image such that $\min(\{k | \hat{y}_{i,k} \neq y_i\} \cup \{1\}) = 0$ in Algorithm 1. As addi-

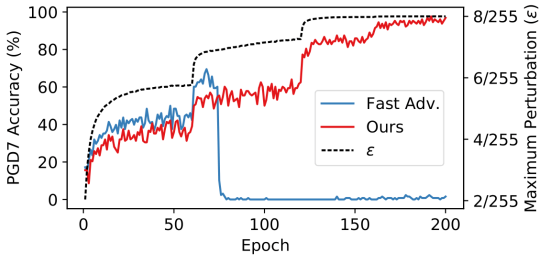


Figure 5: (CIFAR10) Comparison of PGD7 accuracy on the training batch between fast adversarial training with ϵ-scheduling and the proposed method. The dashed line indicates the average maximum perturbation $\mathbb{E}\left[||\delta||_\infty\right]$ calculated from the proposed method for each epoch and is used as the maximum perturbation of fast adversarial training.

tional batches are trained, the average maximum perturbation $\mathbb{E}[||\delta||_\infty]$ increases, as in Figure 5, where $\delta = x' - x$ and x' is selected by the proposed method. Thus, the proposed method may appear to simply be a variation of ϵ-scheduling. In order to point out the difference, fast adversarial training with ϵ-scheduling is also considered. For each epoch, we use the average maximum perturbation $\mathbb{E}[||\delta||_\infty]$ calculated from the proposed method as the maximum perturbation ϵ. The result is summarized in Figure 5.

Notably, ϵ-scheduling cannot help fast adversarial training avoid catastrophic overfitting. The main difference between ϵ-scheduling and the proposed method is that, whereas ϵ-scheduling uniformly applies the same magnitude of the perturbation for every image, the proposed method gradually increases the magnitude of the perturbation appropriately by considering the loss surface of each image. Therefore, in contrast to ϵ-scheduling, the proposed method successfully prevents catastrophic overfitting, despite the same size of the average perturbation used during the training process.

5 Adversarial Robustness

In this section, we conduct a set of experiments on CIFAR10 (Krizhevsky, Hinton et al. 2009) and Tiny ImageNet (Le and Yang 2015) , using PreAct ResNet-18 (He et al. 2016). Input normalization and data augmentation including 4-pixel padding, random crop and horizontal flip are applied. We use SGD with a learning rate of 0.01, momentum of 0.9 and weight decay of 5e-4. To check whether catastrophic overfitting occurs, we set the total epoch to 200. The learning rate decays with a factor of 0.2 at 60, 120, and 160 epochs. All experiments were conducted on a single NVIDIA TITAN V over three different random seeds. Our implementation in PyTorch (Paszke et al. 2019) with Torchattacks (Kim 2020) is available at https://github.com/Harry24k/catastrophic-overfitting.

During the training session, the maximum perturbation ϵ was set to $8/255$. For PGD adversarial training, we use a step size of $\alpha = \max(2/255, \epsilon/n)$, where n is the number of steps. TRADES uses $\alpha = 2/255$ and seven steps for generating adversarial images. Following Wong, Rice, and

	Method	Standard	FGSM	PGD50	Black-box	AA	Time (h)
Multi-step	PGD2 Adv.	**86.6±0.8**	49.7±2.6	36.0±2.3	**85.6±0.8**	34.8±2.1	4.5
	PGD4 Adv.	86.0±0.8	49.6±3.0	36.7±2.9	85.3±0.8	35.4±2.6	6.7
	PGD7 Adv.	84.4±0.2	**51.5±0.1**	**40.5±0.1**	83.8±0.2	**39.4±0.2**	11.1
	TRADES	85.3±0.4	50.7±1.6	39.3±1.9	84.4±0.4	38.6±2.0	15.1
Single-step	Fast Adv.	84.5±4.3	**95.1±6.8**	0.1±0.1	80.8±8.7	0.0±0.0	3.2
	GradAlign	83.9±0.2	44.3±0.0	31.7±0.2	83.3±0.3	30.9±0.2	13.6
	Ours ($c = 2$)	86.8±0.3	48.3±0.5	32.5±0.2	85.9±0.1	30.9±0.2	3.5
	Ours ($c = 3$)	87.7±0.8	50.5±2.4	**33.9±2.3**	86.7±0.9	**32.3±2.2**	3.9
	Ours ($c = 4$)	**87.8±0.9**	50.5±2.3	33.7±2.4	**87.0±0.8**	32.2±2.4	4.4

Table 1: Standard and robust accuracy (%) and training time (hour) on CIFAR10.

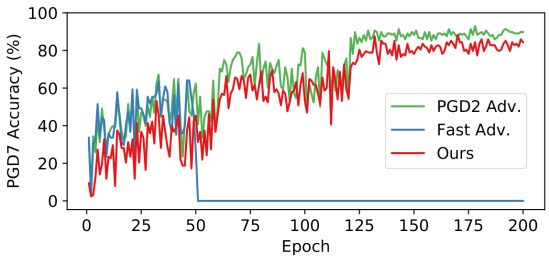


Figure 6: (Tiny ImageNet) PGD7 accuracy on the training batch.

Method	Standard	FGSM	PGD50	Time (h)
PGD2 Adv.	46.3±1.2	14.7±2.7	**10.3±2.7**	27.7
Fast Adv.	26.2±0.7	**49.0±5.7**	0.0±0.0	19.6
Ours ($c = 3$)	**49.6±1.5**	12.5±0.1	7.8±0.1	25.7

Table 2: Standard accuracy and robustness (%) and training time (h) on Tiny ImageNet.

Kolter (2020), we use $\alpha = 1.25\epsilon$ for fast adversarial training and the proposed method. The regularization parameter β for the gradient alignment of GradAlign is set to 0.2 as suggested by Andriushchenko and Flammarion (2020).

First, we check whether our method shows the same results as those of Tiny ImageNet described in the previous section. Figure 6 shows that the proposed method also successfully prevents catastrophic overfitting in a large dataset. PGD7 accuracy decreases rapidly only for fast adversarial training after the 49th epoch, but not for others including the proposed method. The full results with the change in distortion are shown in Appendix B.

We then evaluate robustness on the test set. FGSM and PGD50 with 10 random restarts are used for evaluating robustness of the models. Furthermore, to estimate accurate robustness and detect gradient obfuscation (Athalye, Carlini, and Wagner 2018), we also consider PGD50 adversarial images generated from Wide-ResNet 40-10 (Zagoruyko and Komodakis 2016) trained on clean images (Black-box), and AutoAttack (AA) which is one of the latest strong adversarial attacks proposed by Croce and Hein (2020).

Tables 1 and 2 summarize the results. From Table 1, we can see that multi-step adversarial training methods yield more robust models, but generally requires a longer computational time. In particular, TRADES requires over 15 hours, which is 5-times slower than the proposed method. Among the single-step adversarial training methods, fast adversarial training is computationally efficient, however, because catastrophic overfitting has occurred, it shows 0% accuracy against PGD50 and AA.

Interestingly, we observe that fast adversarial training achieves a higher accuracy for FGSM adversarial images than clean images in both datasets, which does not appear in other methods. The accuracy when applying FGSM on only correctly classified images is 84.4% on CIFAR10, whereas all other numbers remain almost unchanged when we use attacks on correctly classified clean images. We note that this is another characteristic of the catastrophic overfitted model which we describe in more detail in Appendix B.

The proposed method, by contrast, shows the best standard accuracy and robustness against PGD50, Black-box, and AA with a shorter time. GradAlign also provides sufficient robustness; however, it takes 3-times longer than the proposed method. As shown in Table 2, similar results are observed on Tiny ImageNet. We include the results of the main competitors, PGD2 adversarial training, fast adversarial training, and the proposed method with $c = 3$ which shows the best performance on CIFAR10. Here again, the proposed method shows high standard accuracy and adversarial robustness close to that of PGD2 adversarial training. We provide additional experiments with different settings, such as cyclic learning rate schedule in Appendix C.

6 Conclusion

In this study, we empirically showed that catastrophic overfitting is closely related to decision boundary distortion by analyzing their loss surface and robustness during training. Decision boundary distortion provides a reliable understanding of the phenomenon in which a catastrophic overfitted model becomes vulnerable to multi-step adversarial attacks, while achieving a high robustness on the single-step adversarial attacks. Based on these observations, we suggested a new simple method that determines the appropriate magnitude of the perturbation for each image. Further, we evaluated robustness of the proposed method against various adversarial attacks and showed sufficient robustness using single-step adversarial training without the occurrence of any catastrophic overfitting.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (NRF-2019R1A2C2002358).

References

Alayrac, J.-B.; Uesato, J.; Huang, P.-S.; Fawzi, A.; Stanforth, R.; and Kohli, P. 2019. Are Labels Required for Improving Adversarial Robustness? In *Advances in Neural Information Processing Systems*, 12214–12223.

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

Athalye, A.; Carlini, N.; and Wagner, D. 2018. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *arXiv preprint arXiv:1802.00420* .

Carmon, Y.; Raghunathan, A.; Schmidt, L.; Duchi, J. C.; and Liang, P. S. 2019. Unlabeled data improves adversarial robustness. In *Advances in Neural Information Processing Systems*, 11192–11203.

Cohen, J. M.; Rosenfeld, E.; and Kolter, J. Z. 2019. Certified adversarial robustness via randomized smoothing. *arXiv preprint arXiv:1902.02918* .

Croce, F.; and Hein, M. 2020. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. *arXiv preprint arXiv:2003.01690* .

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

Gowal, S.; Dvijotham, K.; Stanforth, R.; Bunel, R.; Qin, C.; Uesato, J.; Arandjelovic, R.; Mann, T.; and Kohli, P. 2018. On the effectiveness of interval bound propagation for training verifiably robust models. *arXiv preprint arXiv:1810.12715* .

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Kim, H. 2020. Torchattacks: A Pytorch Repository for Adversarial Attacks. *arXiv preprint arXiv:2010.01950* .

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

Lamb, A.; Verma, V.; Kannala, J.; and Bengio, Y. 2019. Interpolated adversarial training: Achieving robust neural networks without sacrificing too much accuracy. In *Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security*, 95–103.

Le, Y.; and Yang, X. 2015. Tiny imagenet visual recognition challenge. *CS 231N* 7: 7.

Lee, S.; Lee, J.; and Park, S. 2020. Lipschitz-Certifiable Training with a Tight Outer Bound. *Advances in Neural Information Processing Systems* 33.

Li, B.; Wang, S.; Jana, S.; and Carin, L. 2020. Towards Understanding Fast Adversarial Training. *arXiv preprint arXiv:2006.03089* .

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

Najafi, A.; Maeda, S.-i.; Koyama, M.; and Miyato, T. 2019. Robustness to adversarial perturbations in learning from incomplete data. In *Advances in Neural Information Processing Systems*, 5541–5551.

Pang, T.; Xu, K.; and Zhu, J. 2019. Mixup inference: Better exploiting mixup to defend adversarial attacks. *arXiv preprint arXiv:1909.11515* .

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, 8026–8037.

Salman, H.; Li, J.; Razenshteyn, I.; Zhang, P.; Zhang, H.; Bubeck, S.; and Yang, G. 2019. Provably robust deep learning via adversarially trained smoothed classifiers. In *Advances in Neural Information Processing Systems*, 11292–11303.

Shafahi, A.; Najibi, M.; Ghiasi, M. A.; Xu, Z.; Dickerson, J.; Studer, C.; Davis, L. S.; Taylor, G.; and Goldstein, T. 2019. Adversarial training for free! In *Advances in Neural Information Processing Systems*, 3358–3369.

Smith, L. N.; and Topin, N. 2019. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, volume 11006, 1100612. International Society for Optics and Photonics.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* .

Tramèr, F.; Kurakin, A.; Papernot, N.; Goodfellow, I.; Boneh, D.; and McDaniel, P. 2017. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204* .

Vivek, B.; and Babu, R. V. 2020. Single-step adversarial training with dropout scheduling. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 947–956. IEEE.

Wong, E.; and Kolter, Z. 2018. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning*, 5286–5295.

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

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

Zhang, H.; Chen, H.; Xiao, C.; Gowal, S.; Stanforth, R.; Li, B.; Boning, D.; and Hsieh, C.-J. 2019a. Towards stable and efficient training of verifiably robust neural networks. *arXiv preprint arXiv:1906.06316* .

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412* .

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