REVISITING AND IMPROVING FGSM ADVERSARIAL TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

FGSM adversarial training often fails to obtain a robust model, and the derived model often suffers from catastrophic overfitting, e.g., it is difficult to resist PGD attacks. In this paper, we found that the convergent FGSM adversarial training model tends to rely too much on those insignificant features in the data distribution, called small-scale features, but PGD adversarial training can prevent the model from overutilizing these features. SVD decomposition is performed on the data matrix, and the eigenvectors are used as the basis of the data feature space. On this basis, it is discussed that excessive use of small-scale features may increase the local nonlinearity of the model, making it difficult for the model to generalize from FGSM adversarial examples to PGD adversarial examples. To address this issue, propose to perform random data augmentation in the feature space instead of the original sample space, which can disrupt the correlation between small-scale features and data labels, thereby preventing the model from over-exploring small-scale features. We call this data augmentation technique FDA (**F**eature Space **D**ata **A**ugmentation). Experiments on real-world data verify that FDA technology can prevent the catastrophic overfitting problem of FGSM adversarial training, and can make the FGSM adversarial training model achieve comparable robustness results to PGD adversarial training on the CIFAR-10 and CIFAR-100 data sets.

1 INTRODUCTION

One of the security issues of deep models is the adversarial example problem (Biggio et al., 2013; Szegedy et al., 2013), where small perturbations on the input samples can cause the model's predictions to change. The currently recognized most effective defense algorithm is the adversarial training algorithm (Athalye et al., 2018), which trains the model on the adversarial samples. Adversarial training can be regarded as a min-max optimization problem (Goodfellow et al., 2014; Madry et al., 2017), in which it is generally believed that the better the inner maximization problem is solved, the better the robustness of the derived model (Wang et al., 2021).

Currently, the most popular adversarial training algorithm with SoTA performance on many tasks is the PGD adversarial training algorithm (PGD AT) (Madry et al., 2017) that solves the inner maximization problem through the PGD attack algorithm. Since the PGD attack algorithm is a multi-step iterative process to find an approximate optimal solution, it makes the PGD AT algorithm very time-consuming. In order to improve the training speed, the FGSM attack algorithm (Goodfellow et al., 2014), which is a one-step approximation algorithm, can be used to replace the PGD attack algorithm, and such adversarial training is called FGSM adversarial training (FGSM AT). However, FGSM AT training suffers from catastrophic overfitting (Wong et al., 2020), that is, the model suddenly fails to defend against PGD attacks during the training process, resulting in poor robustness of the derived model. It is believed that the FGSM adversarial samples are less aggressive than the PGD adversarial samples (Madry et al., 2017; Wong et al., 2020), which leads to the failure of the FGSM adversarial training. In this paper, we will explore alternative reasons for the failure of FGSM AT from the perspective of data distribution.

Ilyas et al. (2019) believe that the model's adversarial vulnerability comes from the fact that the model learns non-robust features in the data, which will lead to a misalignment between the distance metrics under the adversarial attack and the data distribution. Some scholars have found that stochastic gradient descent will cause the model to have a simplicity bias (Arpit et al., 2017;

Kalimeris et al., 2019), that is, the model may be overly dependent on some simple features. Shah et al. (2020) have experimentally proved that the simplicity bias will affect the robustness of the model. Some studies have proposed that the reason why the model is sensitive to small disturbances may be that the decision boundary of the model is almost parallel to the data manifold (Tanay & Griffin, 2016; Li et al., 2020; Shamir et al., 2021). In this paper, the SVD decomposition is implemented on the data matrix and the eigenvectors are extracted as the basis of the data feature space. The eigenvectors corresponding to those small eigenvalues are called small-scale features, which are insignificant parts of the data distribution. We analyzed and experimentally proved that over-reliance on small-scale features will affect the robustness of the model, which is also the reason for the catastrophic overfitting of the FGSM adversarial training model.

Contributions. In this paper, we find that the FGSM adversarial training model tends to rely heavily on small-scale features, while the PGD adversarial training model can suppress the effect of small-scale features in the training process. On one hand, excessive use of small-scale features leads to poor stability of the gradient direction of the model in the local neighborhood. On the other hand, samples along the small-scale feature direction can easily escape the data distribution. Therefore, FGSM adversarial training can only ensure that the model is robust in some of the small-scale feature directions, while the multi-step iterative attack of PGD is easy to generate adversarial examples in other directions. We believe that excessively small-scale features lead to catastrophic overfitting of the FGSM adversarial training model. Based on this, we propose the FDA technique, which performs random data augmentation on the feature space of the data. FDA can prevent the FGSM adversarial training model from relying too much on small-scale features and improve the robustness of FGSM adversarial training. On the CIFAR-10 and CIFAR-100 datasets, experiments show that the proposed data augmentation strategy can improve the robustness of the model and achieve robustness comparable to PGD adversarial training.

2 Related Work

Adversarial training is currently the most effective way to improve model adversarial robustness (Goodfellow et al., 2014; Madry et al., 2017). Given training data $\{(x_i, y_i)\}_{i=1}^n$, a network f_θ parameterized by θ and a threat model Δ, adversarial training can be formulated as the following optimization objective:

$$\min_\theta \sum_{i=1}^{n} \max_{\delta \in \Delta} \ell(f_\theta(x_i + \delta), y_i) \tag{1}$$

where ℓ is the loss function, generally using softmax cross entropy loss. In this paper, we focus on the threat model $\Delta = \{\delta : ||\delta||_\infty \leq \varepsilon\}$ under infinite norm ℓ_∞ attack. The PGD attack algorithm (Madry et al., 2017), which uses a multi-step projected gradient ascent technique to obtain an approximate optimal solution, is currently the most popular method for solving internal maximization problems. The PGD algorithm can be further formalized as the following iterative formula:

$$\delta_{PGD}^{t+1} = \Pi_{[-\varepsilon,\varepsilon]}[\delta_{PGD}^t + \alpha sign(\bigtriangledown_{\delta_{PGD}^t} \ell(f_\theta(x + \delta_{PGD}^t, y)))] \tag{2}$$

where t is the number of iterations, α is the step size, and Π is the projection operator. Generally, the larger the number of iterations t is, the more time the algorithm consumes. Therefore, in this paper, we consider using the one-step gradient ascent algorithm FGSM to replace the PGD algorithm. The formula is as follows:

$$\delta_{FGSM} = \alpha sign(\bigtriangledown_x \ell(f_\theta(x), y)) \tag{3}$$

Generally, set the step size $\alpha = \varepsilon$. However, adversarial training based on the FGSM algorithm generally does not get a robust model, and even catastrophic overfitting occurs (Wong et al., 2020). So-called catastrophic overfitting is the phenomenon in which a model suddenly loses its defenses against stronger adversarial attacks at some point in the training process (Wong et al., 2020).

Wong et al. (2020) believe that the success of PGD AT lies in the fact that the previous perturbation of the iterative algorithm can be used as the initialization of the next perturbation. Therefore, they propose to use random restarts for initialization and use larger step sizes for FGSM AT and find that catastrophic overfitting can be improved. Experiments by Andriushchenko & Flammarion (2020) found that random initialization do not solve the catastrophic overfitting problem for larger ε. Furthermore, they demonstrate that FGSM AT with random initialization is effective because it reduces the expected magnitude of the perturbation, thereby improving the local linear approximation capability of the FGSM algorithm. Based on this, they propose gradient alignment regularization to improve the local linearity of the model to solve the catastrophic overfitting problem of FGSM AT.

3 Understanding the Adversarial Example Problem

3.1 Definition: large-scale features and small-scale features

Assume that the data points $X = \{x_i\}_{i=1}^n$ is distributed on the data manifold M embedded in $\mathbb{R}^D$ (Bengio et al., 2013). In general, the intrinsic dimension of M, d, is much smaller than D. Here, we approximate the data manifold using the subspace spanned by the $d-$dimensional PCA principal component eigenvectors. The data matrix X is processed by the SVD decomposition technique, and its eigenvectors $\{\mu_1, \mu_2, ..., \mu_D\}$ are arranged in descending order according to the eigenvalues. Here, we refer to the space with these eigenvectors as the basis as the **Feature Space** of the data. In addition, the first d eigenvectors are called **large-scale features** because the data mainly changes in these directions, and the remaining $D - d$ eigenvectors are called **small-scale features**.

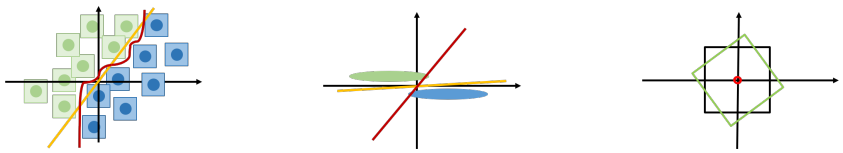
In general, we do not want the model to overuse small-scale features for decision-making, since natural samples are mainly composed of large-scale features, and from a human perspective, it seems that large-scale features are sufficient to classify most samples. In Appendix A, we visualize some eigenvectors obtained by SVD decomposition on some datasets, and extract the small-scale features and large-scale features in natural samples for display.

3.2 The influence of small-scale features on the adversarial robustness of the model

Here, we consider the adversarial sample problem in two cases: (1) large-scale features can perfectly complete the classification task; (2) the separability of large-scale features on tasks is not as good as that of small-scale features. Figure 1 reflects the model robustness problem in these two cases, respectively.

Figure 1(a) is a schematic diagram often used to study the adversarial example problem, which ignores the influence of small-scale features on the model. The adversarial perturbations in this case are sample-specific and mainly consist of large-scale features. Therefore, it is necessary to search for adversarial samples in the ℓ_p-norm neighborhood of each training sample and use them as a training set, which is adversarial training. In Figure 1(b), the vertical axis can be considered as the small-scale feature direction. When the decision boundary is very dependent on the small-scale feature direction, the sample can easily cross the decision boundary along the vertical axis, resulting in the adversarial sample phenomenon. The adversarial perturbation in this case is specific to the data distribution and mainly consists of small-scale features. There is no need to use an adversarial training algorithm at this time, as long as we prevent the model from relying too much on small-scale features.

Therefore, a robust model cannot over-utilize small-scale features, and secondly, it must ensure that the samples are as far away from the decision boundary as possible. However, in our experiments, we find that the standard training and FGSM AT models are prone to over-reliance on small-scale features, which of course has a certain relationship with the data distribution. Next, we will observe the utilization of small-scale features by the model under different training methods.



(a) (b) (c)

Figure 1: (a) The green and blue points represent samples of different classes, respectively, and the light green and light blue regions represent the ℓ_∞ neighborhood of the samples. The samples in the ℓ_∞ neighborhood of some samples can cross the orange classification boundary, so the orange classification boundary is not robust. After adversarial training, dark red robust classification boundaries can be obtained. (b)The green and blue oval areas represent different classes. The orange classification boundary that relies too much on small-scale features (vertical axis) is not robust, and all samples can easily cross the classification boundary along the vertical axis. The dark red classification boundary is more robust than the orange boundary, but this may lose the accuracy of the model on the clean test set. (c) The ℓ_∞-norm neighborhood of the origin sample in pixel space (black square area) and in feature space (green square area).

4 Revisiting FGSM Adversarial Training: Over-reliance on small-scale features

The adversarial samples generated during the adversarial training process are actually continuously adjusting the distribution of the original samples, so that the model can use reasonable features from the data to find a robust classification boundary. The direction of adversarial perturbation determines how the distribution of the original samples is adjusted. Therefore, we have reason to believe that there is a large difference in the direction of the adversarial perturbations generated during FGSM AT and PGD AT training, which leads to different robustness of the final models. Here, we capture the difference in direction of the adversarial perturbation by the proportion of small-scale features in the perturbation. Since adversarial perturbations are generally generated by information related to model gradients,the FGSM AT model and the PGD AT model also differ in the direction of the model gradient.

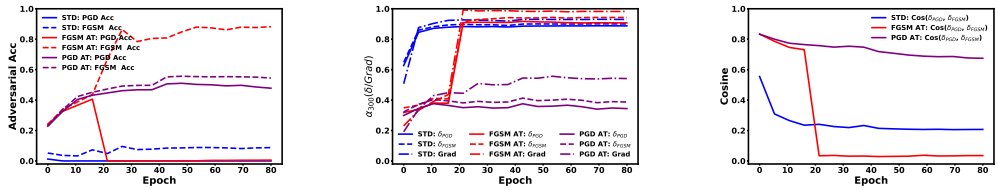
For sample x, we use the gradient of the loss function as the research object: $g = \frac{\partial \ell(f_\theta(x), y)}{\partial x}$. Furthermore, the adversarial perturbations generated by the PGD and FGSM algorithms are denoted as δ_{PGD} and δ_{FGSM}, respectively. For a given d, $\alpha_d(g)$ represents the proportion of small-scale features components $\{\mu_d, \mu_{d+1}, ..., \mu_D\}$ in g, which can be calculated by the following formula:

$$\alpha_d(g) = \frac{\sum_{j=d}^{D}(g^T \mu_j)^2}{||g||_2^2} \tag{4}$$

Similarly, we can calculate $\alpha_d(\delta_{PGD})$ and $\alpha_d(\delta_{FGSM})$. Note that the calculation results of $\alpha_d(\cdot)$ shown in this article are the mean over the entire dataset. In the experiments in this paper, d is set to 300, because at least 95% of the information in the data (MNIST, CIFAR data) can be recovered by using the first 300 eigenvectors.

The phenomenon of excessive use of small-scale features On the CIFAR-10 (Krizhevsky et al., 2009) dataset, we train the Resnet-18 network (He et al., 2016) using three training methods: standard training (STD), PGD AT (Madry et al., 2017), and FGSM AT. Meanwhile, the changes of the values of $\alpha_d(g)$, $\alpha_d(\delta_{PGD})$ and $\alpha_d(\delta_{FGSM})$ are recorded during the training. Additionally, we track changes in $\cos(\delta_{PGD}, \delta_{FGSM})$, which can reflect the local linearity of the model (Andriushchenko & Flammarion, 2020). Here, we focus on the difference in direction of adversarial perturbations between models and the difference in direction between different perturbations of a single model.

The difference in direction of adversarial perturbations between models In Figure 2(a), it can be noticed that the robustness of the FGSM AT model under PGD adversarial attack drops suddenly around the 17th epoch, which is called catastrophic overfitting. In Figure 2(b), we can first see that the changing trends of $\alpha_d(g)$, $\alpha_d(\delta_{PGD})$ and $\alpha_d(\delta_{FGSM})$ are consistent, which is understandable



(a) (b) (c)

Figure 2: Visualization of different metrics in the training process of Resnet-18 models for STD, PGD AT, and FGSM AT on CIFAR-10. All statistics are computed on the test set. (a) In terms of attack recognition accuracy, FGSM AT suffers from catastrophic overfitting after the 17th epoch. (b) The $\alpha_d(g)$, $\alpha_d(\delta_{PGD})$ and $\alpha_d(\delta_{FGSM})$ of the STD model and the FGSM AT model are all too large after convergence, which indicates that the models rely too much on small-scale features. (c) The angle between δ_{PGD} and δ_{FGSM} of the STD model and the FGSM AT model is relatively large during the training process, which indicates that the gradient of the model fluctuates greatly in the local neighborhood.

because δ_{PGD} and δ_{FGSM} are generated through information related to g. Secondly, it can be clearly observed that in the middle and late stages of training of the FGSM AT model and the STD model, δ_{PGD}, δ_{FGSM}, and gradient g contain a large number of small-scale features, even exceeding 80%, while the PGD AT model does not have the problem of excessive use of small-scale features . Moreover, while the catastrophic overfitting of the FGSM AT model occurs, the utilization rate of the model for small-scale features also begins to increase suddenly.

The difference in direction between different perturbations of a single model Since both δ_{PGD} and δ_{FGSM} of the FGSM AT model contain a large number of small-scale features, why is there a huge difference in the recognition accuracy of the model on the two perturbations? The results in Figure 2(c) can further explain this phenomenon, that is, catastrophic overfitting phenomenon. Although both δ_{PGD} and δ_{FGSM} of the FGSM AT model are dominated by small-scale features, it can be observed that the angle between δ_{PGD} and δ_{FGSM} is getting larger after the 18th epoch, which means that the FGSM AT model can only guarantee robustness in some of the small-scale feature directions rather than all small-scale feature directions. Going deeper, since the adversarial perturbation is related to the gradient, this also means that the FGSM AT model has high non-linearity in the local neighborhood, making it difficult to generalize δ_{FGSM} to δ_{PGD}. Compared to the more robust PGD AT model, suppressing small-scale features during training will increase the direction consistency between the two perturbations, and the robustness of the model will generalize better. Therefore, excessive use of small-scale features by the model may increase the local nonlinearity of the model.

5 Improving FGSM Adversarial Training : Feature Space Data Augmentation (FDA)

The results of Section 4 tell us that over-reliance on small-scale features is one of the differences between the FGSM AT model and the PGD AT model, which may be the reason for the poor robustness of the FGSM AT model. Therefore, we need to limit the model's use of small-scale features and let the model pay more attention to large-scale features. Essentially, the model uses small-scale features for classification because there is a certain correlation between small-scale features and labels, so we must break the connection between them. Therefore, we propose to adjust the training samples to alleviate the model's excessive focus on small-scale features.

To this end, we propose a stochastic data augmentation strategy based on feature space basis, referred to as the **FDA** strategy, which is a very simple data augmentation method. The specific operation process is as follows: given the basis $\{\mu_1, \mu_2, ..., \mu_D\}$ of the feature space, the original training set $D = \{(x_i, y_i)\}_{i=1}^n$ can be processed and a new data set can be obtained $D^{'} = \{(x_i^{'}, y_i)\}_{i=1}^n$ by the following formula:

$$x_i^{'} = x_i + \sum_{k=1}^{D} z_{i,k}\mu_k, \quad z_{i,k} \sim Uniform(-\varepsilon, \varepsilon), x_i \in D \tag{5}$$

That is, for all samples we add uniform random noise across all feature space dimensions. For small-scale feature dimensions $k > d$, because the projection of samples on these dimensions is small, the correlation between small-scale features and labels can be well weakened by adding noise. As shown in Figure 1 (c), when the dimensionality of the input increases, the non-overlapping regions between the ℓ_∞ neighborhoods of the feature space and the pixel space will become larger. Therefore, we expect the combination of FDA's feature space data augmentation and pixel space search for adversarial training should be able to improve model performance.

6 Experiment

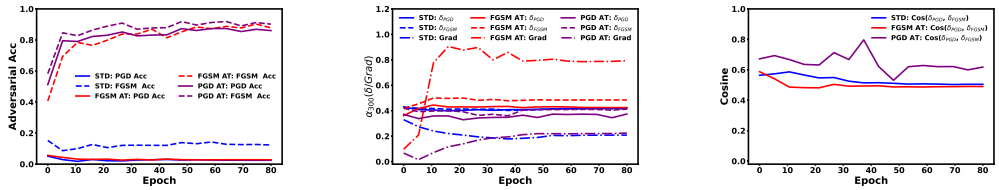
6.1 Experimental Setup

Model and Dataset We conduct experiments on three datasets: MNIST (LeCun et al., 1998), CIFAR-10, and CIFAR-100 (Krizhevsky et al., 2009), where pixel values of all samples are normalized to [0, 1] and four models LeNet (for MNIST) (LeCun et al., 1998), Resnet-18 (for CIFAR-10) (He et al., 2016), VGG16 (for CIFAR-10) (Simonyan & Zisserman, 2014), WRN-34-10 (for CIFAR-10) (Zagoruyko & Komodakis, 2016) and Resnet-34 (for CIFAR-100) (He et al., 2016). All networks are trained for 100 epochs using the SGD optimizer with a piecewise learning rate schedule([40, 60, 80], downscaled by 0.1 per adjustment), and deployed on a TITAN Xp GPU. In training, we use a batchsize of 128.

Training Method The training methods involved in this article are as follows: (1) standard training model (STD), (2) PGD AT (Madry et al., 2017), (3) STD with FDA, (4) FGSM AT, (5) FGSM AT with FDA, (6) FGSM AT with gradient alignment regularization (FGSM AT+GradAlign) (Andriushchenko & Flammarion, 2020), the regularization coefficients used in this paper are all set to 0.2. For more detailed settings in the experiment, please refer to Appendix B.

6.2 MNIST

As a simple dataset, it can be seen from Figure 3(b) that no matter which training method is used, the adversarial perturbation of the LeNet model trained on MNIST will not be overly dependent on small-scale features. Therefore MNIST can be considered as a large-scale features separable dataset and at this time the model's adversarial example problem mainly comes from the large-scale direction. The results in Figure 3(c) show that when the model does not overuse small-scale features, the directions of the two perturbations of δ_{PGD} and δ_{FGSM} are in good agreement. In Figure 3(a), the FGSM AT model suffers from catastrophic overfitting, mainly because the perturbation is so large that it is difficult for the FGSM algorithm to find optimal adversarial examples in the local neighborhood.



(a) (b) (c)

Figure 3: On MNIST, LeNet models are trained by STD, FGSM AT, and PGD AT respectively. (a) The recognition accuracy of the model on FGSM adversarial samples and PGD adversarial samples, respectively. (b) The proportion of small-scale features in the gradients (Grad) and the adversarial perturbation (δ_{PGD} / δ_{FGSM}) of the model. (c) Cosine values($\cos(\delta_{PGD}, \delta_{FGSM})$) between δ_{PGD} and δ_{FGSM} of the model.

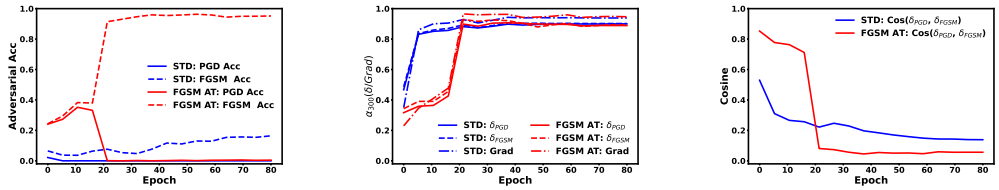
Table 1: The recognition accuracy of the LeNet on MNIST clean test set and under different attack algorithms. We report mean values of the accuracy in three independent experiments. The adversarial perturbation size is 0.3.

Training Method	STD	PGD AT	FGSM AT	STD FDA	FGSM AT GradAlign	FGSM AT FDA
Test Acc	99.49%	99.03%	98.56%	99.45%	98.18%	99.02%
FGSM Acc	11.23%	96.28%	88.56%	21.07%	91.35%	96.58%
PGD Acc	0.00%	87.70%	0.00%	0.00%	15.07%	87.66%

On MNIST, because ε is too large, if the FDA is directly performed, the original data will be greatly deformed. If the model is directly trained on the FDA-enhanced dataset, it is difficult to guarantee the robustness of the FGSM adversarial training model. Therefore, we only use the loss gradient of FDA-augmented data as the perturbation direction of the original sample points to generate FGSM adversarial samples for FGSM AT. The robustness test results of the models under different training methods are shown in Table 1. It can be clearly seen that the FGSM AT model trained under our scheme has similar robustness results compared to the PGD AT model.

6.3 CIFAR-10

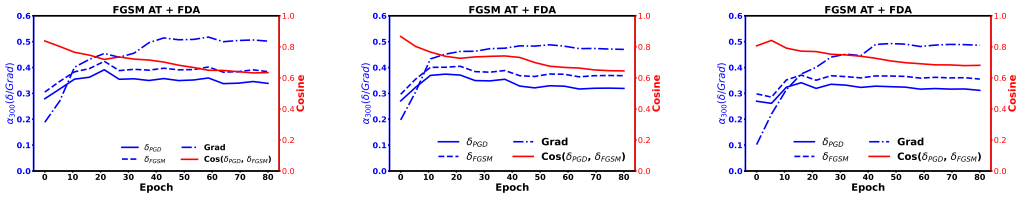
Figure 4 shows how the various metrics of the WRN-34-10 model trained by FGSM AT and STD change as the training progresses. It can be clearly seen from Figure 4 (a) that the model trained by FGSM AT has a catastrophic overfitting phenomenon after the 15th epoch, and at the same time, the proportions of small-scale features in the model's adversarial perturbation and gradient are gradually increasing (Figure 4 (b)). As shown in Figure 4 (c), excessive reliance on small-scale features also makes the angle between the two disturbances of the model larger, which intensifies the local nonlinearity of the model. Figure 5(a) (Resnet-18) and (b) (WRN-34-10) record the fluctuations of the FGSM AT model trained with the FDA technique under various metrics. Compared with the results in Figure 2 and Figure 4, the models in Figure 5 have none of $\alpha_d(g)$, $\alpha_d(\delta_{PGD})$ and $\alpha_d(\delta_{FGSM})$ exceeding 50%, which shows that the FDA technology has well alleviated the excessive dependence of the FGSM AT model on small-scale features. At the same time, FDA technology can effectively improve the direction consistency of the two adversarial perturbations.



(a) (b) (c)

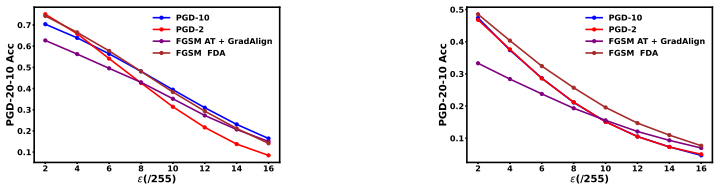
Figure 4: On CIFAR-10, WRN-34-10 models are trained by STD, and FGSM AT. (a) The recognition accuracy of the model on FGSM adversarial samples and PGD adversarial samples, respectively. (b) The proportion of small-scale features in the gradients (Grad) and the adversarial perturbation (δ_{PGD} / δ_{FGSM}) of the model. (c) Cosine values($\cos(\delta_{PGD}, \delta_{FGSM})$) between δ_{PGD} and δ_{FGSM} of the model.

In order to further illustrate that the FDA technology does solve the catastrophic overfitting problem of the FGSM AT model and improve the adversarial robustness of the model, we summarize the robustness evaluation results of various models in Table 2 and Table 3. First, the robust accuracy under PGD attack of FGSM AT + FDA on Resnet-18 is 48.33%, which is better than FGSM AT + GradAlign and slightly lower than PGD AT. Secondly, the FGSM AT model combined with FDA also has excellent performance under the attack of AutoAttack. The most important FGSM AT + FDA's Resnet-18 model can achieve 48% PGD ℓ_∞ robust accuracy in about 40 minutes, which is much lower than the training time of PGD AT. As a larger model, it is difficult for us to train a WRN-34-10 model of PGD AT, but FGSM AT+FDA technology can easily achieve 49.43% PGD ℓ_∞ robust accuracy. Additionally, we tested the robust performance of the FGSM+FDA model,



(a) CIFAR-10: Resnet-18 (b) CIFAR-10: WRN-34-10 (c) CIFAR-100: Resnet-34

Figure 5: Deep models (a) Resnet-18, (b) WRN-34-10 and (c) Resnet-34 trained with FGSM AT + FDA technology. These figures depict the changes of $\alpha_d(g)$, $\alpha_d(\delta_{PGD})$, $\alpha_d(\delta_{FGSM})$ and $\cos(\delta_{PGD}, \delta_{FGSM})$ of the trained model.



(a) CIFAR-10: Resnet-18 (b) CIFAR-100: Resnet-34

Figure 6: PGD-20-10 robustness test results of a robust model trained with $\varepsilon = 8/255$ under different adversarial perturbation sizes ε.

FGSM+GradAlign model, and PGD AT models (PGD-10 AT and PGD-2 AT) trained at $\varepsilon = 8/255$ against PGD attacks of different perturbation sizes. The results are shown in Figure 6(a). It can be seen that our method can effectively improve the robustness of the FGSM AT model, and can be compared with the PGD AT model under different ε. The relevant experimental results of the VGG16 model are shown in Appendix C.

Table 2: The recognition accuracy of the Resnet-18 on CIFAR-10 clean test set and under different attack algorithms. We report mean values of the accuracy in three independent experiments. The adversarial perturbation size is $8/255$. '_' indicates that we did not test the performance of the model.

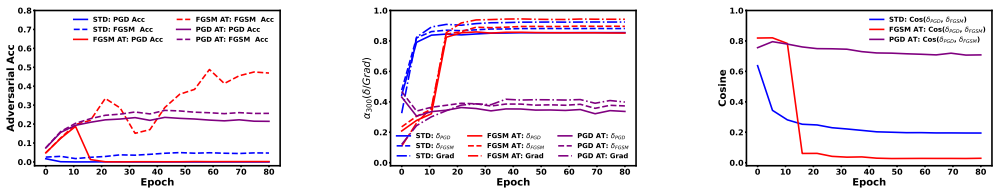
Training Method	STD	PGD AT	FGSM AT	STD FDA	FGSM AT GradAlign	FGSM AT FDA
Clean test	93.29%	81.52%	88.65%	93.64%	80.06%	81.26%
FGSM	8.63%	55.84%	88.63%	16.23%	51.14%	54.35%
PGD-20	0.01%	50.92%	0.49%	0.10%	45.23%	48.33%
AutoAttack	_	45.40%	_	_	39.00%	45.10%

Table 3: The recognition accuracy of the WRN-34-10 on CIFAR-10 clean test set and under different attack algorithms. We report mean values of the accuracy in three independent experiments. The adversarial perturbation size is $8/255$.

Training Method	STD	PGD AT	FGSM AT	STD FDA	FGSM AT GradAlign	FGSM AT FDA
Clean test	94.40%	_	87.65%	94.43%	71.48%	84.26%
FGSM	16.53%	_	96.58%	13.62%	39.75%	55.24%
PGD	0.00%	_	0.00%	0.10%	35.32%	49.43%
AutoAttack	_	_	_	_	32.50%	33.30%

6.4 CIFAR-100

For the CIFAR-100 data, the changes in the various metrics of the Resnet-34 model under different training methods are recorded, and the results are shown in Figure 7. First, we also observed catastrophic overfitting of the FGSM AT Resnet-34 model on CIFAR-100 (Figure 7(a)). The results in Figure 7(b) also show that the gradients and adversarial perturbations of the STD model and the FGSM AT model over-rely on small-scale features. As can be seen in Figure 7(c), the $\cos(\delta_{PGD}, \delta_{FGSM})$ of the PGD AT model is larger than that of the FGSM AT model, which indicates that the gradient direction of the FGSM AT model fluctuates greatly in the local neighborhood, and the FGSM adversarial samples cannot generalize well to the PGD adversarial samples. Therefore, the FGSM AT model suffers from catastrophic overfitting.



(a) (b) (c)

Figure 7: Visualization of different metrics in the training process of Resnet-34 models for STD, PGD AT, and FGSM AT on CIFAR-100. All statistics are computed on the test set.

The results in Figure 5(c) show that the FDA technology can well limit the excessive use of small-scale features by the FGSM AT model, and at the same time improve the local linearity of the model. Table 4 shows the robustness test results of the Resnet-34 model under different training methods.It can be seen that the FGSM AT model combined with FDA technology has achieved better robustness performance under both PGD and AutoAttack attacks. The robustness test of PGD-20-10 in Figure 6(c) shows that our method can make the model as competitive as the PGD AT model.

Table 4: The recognition accuracy of the Resnet-34 on CIFAR-100 clean test set and under different attack algorithms. We report mean values of the accuracy in three independent experiments. The adversarial perturbation size is $8/255$.

Training Method	STD	PGD AT	FGSM AT	STD FDA	FGSM AT GradAlign	FGSM AT FDA
Clean test	73.40%	57.18%	62.87%	72.28%	48.58%	56.63%
FGSM	4.95%	25.45%	47.65%	8.27%	25.08%	29.67%
PGD-10	0.00%	23.44%	0.23%	0.20%	22.59%	26.55%
AutoAttack	_	16.30%	_	_	15.90%	19.20%

7 Conclusion

Most of the natural data are located on low-dimensional manifolds, which leads to many dimensions in the background space being degenerate. Numerous degenerate dimensions (small-scale features) facilitate deep models to solve problems in these spaces. Both standard training and FGSM adversarial training models are prone to fall into the trap of such small-scale features, resulting in poor adversarial robustness of the model. Therefore, we need to beware of the harm caused by small-scale features. In this paper, we propose a feature space data augmentation (FDA) technique to weaken the correlation between small-scale features and labels, thereby solving the catastrophic overfitting problem of FGSM adversarial training. Experiments demonstrate that our proposed strategy helps to improve the robustness of FGSM adversarial training and achieves comparable performance to PGD adversarial training. Although the existence of small-scale features is beneficial to improve the performance of the model on tasks, they will affect the generalization ability of the model to out-of-distribution samples. Second, small-scale features may have a certain relationship with universal adversarial perturbations, which is also worthy of our study.

REFERENCES

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

Devansh Arpit, Stanislaw Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International conference on machine learning*, pp. 233–242. PMLR, 2017.

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, pp. 274–283. PMLR, 2018.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 387–402. Springer, 2013.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pp. 2206–2216. PMLR, 2020.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019.

Dimitris Kalimeris, Gal Kaplun, Preetum Nakkiran, Benjamin Edelman, Tristan Yang, Boaz Barak, and Haofeng Zhang. Sgd on neural networks learns functions of increasing complexity. *Advances in neural information processing systems*, 32, 2019.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Yueru Li, Shuyu Cheng, Hang Su, and Jun Zhu. Defense against adversarial attacks via controlling gradient leaking on embedded manifolds. In *European Conference on Computer Vision*, pp. 753–769. Springer, 2020.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Leslie Rice, Eric Wong, and Zico Kolter. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*, pp. 8093–8104. PMLR, 2020.

Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. *Advances in Neural Information Processing Systems*, 33:9573–9585, 2020.