Proceedings of the 2nd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/109/20241413

Improving Model Robustness through Hybrid Adversarial Training: Integrating FGSM and PGD Methods

Zeshan Zhong

College of Information Science and Engineering, Hunan University, Hunan, China

zzs589611808@hnu.edu.cn

Abstract. With the widespread use of deep learning models in various applications. People are gradually realizing the vulnerability of these models to adversarial attacks. Adversarial training is an effective strategy to defend against adversarial attacks. Based on the advantages and disadvantages of the current mainstream Fast Gradient Sign Method (FGSM) adversarial training and Projected Gradient Descent (PGD) adversarial training, this paper proposes a hybrid adversarial training that integrates FGSM and PGD methods and uses the ResNet-18 model and SVHN dataset for testing. Experimental results show that hybrid adversarial training can effectively reduce training time. Its accuracy on the original data set is higher than that of PGD adversarial training, which is improved by about 2%. The performance when facing FGSM attacks is almost the same as that of single FGSM adversarial training. The performance when facing PGD attacks decreases more significantly, which is about 2% to 3% lower than that of PGD adversarial training. This study not only helps to understand the robustness of hybrid adversarial training to models facing adversarial attacks but also helps in studying new adversarial training strategies.

Keywords: Hybrid Adversarial Training, FGSM, PGD, ResNet-18.

1. Introduction

In recent years, with the development of deep learning, especially convolutional neural networks, significant progress has been made in the field of computer vision, which is widely used in fields such as autonomous driving, medical image analysis, and industrial manufacturing. However, there are many attack methods against neural networks, among which adversarial attacks pose a serious threat to deep learning models. Adversarial attacks involve slightly perturbing the input data so that the model produces incorrect outputs [1]. This type of attack will lead to huge safety risks. For example, in the field of autonomous driving, it may cause the autonomous driving system to misjudge, leading to traffic accidents.

The strategies of convolutional neural networks against adversarial attacks include adversarial sample detection, adversarial learning, adversarial noise erasure, etc. [2]. Goodfellow et al. proposed to optimize model parameters through adversarial training. By introducing adversarial samples into the training data, the model's robustness can be enhanced. This is one of the most effective defenses against adversarial attacks [3]. However, Kurakin et al. pointed out that adversarial training (adversarial training with fast gradient descent) can enhance the robustness of the model against single-step attacks, but its defense effect is poor when facing iterative attacks with multiple iterations of sample perturbations, such

DOI: 10.54254/2755-2721/109/20241413

as Basic Iterative Method and Projected Gradient Descent (PGD) attacks [4-5]. To effectively resist iterative attacks, hybrid adversarial training of PGD attacks and the Fast Gradient Sign Method (FGSM) can be used to comprehensively improve the model's ability to resist complex attacks.

Combining adversarial samples and original data in adversarial training can enhance the robustness of the model in defending against adversarial attacks[4-5]. This method is based on the Street View House Numbers dataset, uses FGSM and PGD hybrid adversarial training, and performs adversarial training on the ResNet-18 model. Finally, the robustness of this hybrid training method in the face of adversarial attacks is tested. This study not only helps to understand the robustness of hybrid adversarial training to models facing adversarial attacks but also helps in studying new adversarial training strategies.

2. Overseas and Domestic Research Status

In general, single adversarial training can no longer meet current needs. Adversarial training methods have evolved from the initial simple method to more complex and diverse strategies, often a combination of multiple strategies. For example, the single-step iterative method mentioned below combines adversarial training and knowledge transfer. Madry proposed an adversarial log-pairing method, which introduced a regularization method to penalize the logarithmic difference between the original samples and adversarial samples generated by the PGD attack [6]. Takeru et al. proposed a single-step iterative method that combines adversarial training and knowledge transfer to defend against adversarial attacks and improves the generalization ability of the model through knowledge transfer [7]. Song et al. proposed Multi-strength Adversarial Training, which mixes adversarial samples of different strengths and then puts them into the original data to train the model.

The hybrid training strategy explored in this paper is to combine the PGD and FGSM training strategies. The FGSM algorithm can enhance the robustness of the model against single-step attacks, but its defense capability is poor when facing iterative attacks with multiple iterative optimization sample disturbances. PGD is an iterative attack method. The model trained adversarially using the PGD algorithm has quite good robustness. Although PGD adversarial training is very simple and effective and has good robustness against various adversarial attack methods, it suffers from the problem of low computational efficiency [8]. If the two are mixed and a regularization method is introduced, one of the three methods, original data training, FGSM training, and PGD training, is randomly selected in each round of training. This can greatly reduce the training time, improve the accuracy of the model on the original data set, and maintain good robustness when facing PGD attacks or FGSM attacks.

3. Data and method

3.1. Data

This article uses the SVHN dataset. The SVHN dataset is an important dataset for image recognition and computer vision research. It is generated from Google's Street View images and is a standard dataset for testing image recognition systems. The training set includes 73,257 images and the test set includes 26,032 images. Each image contains one or more digits (0-9) and is 32*32 in size [9].

DOI: 10.54254/2755-2721/109/20241413

3.2. Method

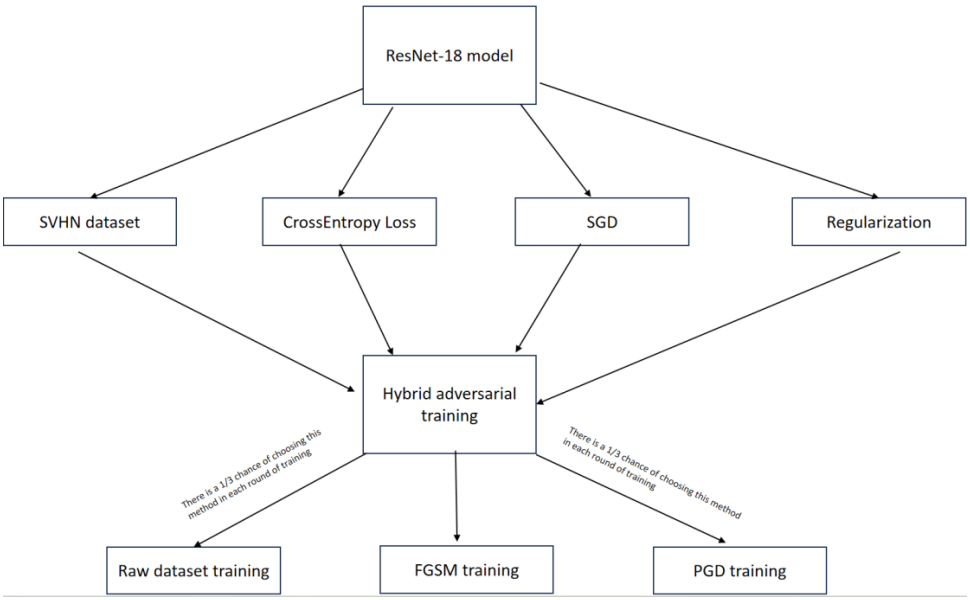

Figure 1. Method Overview

In terms of framework construction, as shown in Figure 1, this paper uses the ResNet-18 framework. The ResNet-18 framework is a classic deep convolutional neural network framework that is widely used in image classification tasks due to its simple and effective design and powerful feature extraction capabilities [10]. Regarding the generation of adversarial samples, this paper uses FGSM and PGD methods to generate adversarial samples and then performs hybrid training. For the choice of loss function. This paper chooses CrossEntropyLoss to calculate the difference between the model prediction and the actual label, which can provide a smoother and more stable gradient and a probability distribution that conforms to the training data [11]. The optimizer selects the stochastic gradient optimizer with a momentum of 0.9. The L2 regularization method is used to limit the excessive growth of model parameters by adding the squared sum of weights to the loss function to help prevent overfitting. The learning rate is adjusted using the cosine annealing strategy, and the learning rate completes a cycle every 200 epochs.

4. Results and discussion

4.1. Hybrid Adversarial Training

This paper uses FGSM and PGD to generate adversarial samples respectively. The perturbation value of FGSM is 8/255. The images of the SVHN dataset mainly contain clear digital areas, and the 8/255 perturbation of FGSM will not cause obvious visual damage to the main part of the image (digits). Therefore, choosing a perturbation value of 8/255 can generate sufficiently deceptive adversarial samples without causing obvious visual distortion of the image. The perturbation of PGD is 16/255, the step size is 2/255, and the number of iterations is 10. The background of the SVHN image is relatively complex. The 16/255 perturbation of PGD can aggravate the perturbation of the background and edges, making it more difficult for the model to distinguish the correct numbers. Larger perturbation amplitudes help to deeply test the model's defense capabilities in complex backgrounds. The selection of these parameters strikes a good balance between the model and the dataset, which can effectively test the adversarial robustness of ResNet-18. In the hybrid training, each round of training will randomly select a training method from the original dataset, FGSM adversarial training, and PGD adversarial training

DOI: 10.54254/2755-2721/109/20241413

for training. The total number of training rounds is 60, and the L2 regularization method is used to prevent overfitting.

The hybrid training method has an accuracy of 81% on the original sample. As shown in Figure 2, the training loss drops from the initial 4.0 to 0.8, while the test loss stabilizes at around 0.7, which indicates that the model has good robustness and generalization ability. The significant decrease in training loss shows that the model gradually learns the characteristics of the data, while the stability of the test loss shows that the model shows consistent performance on unseen data without obvious overfitting. This shows that the model effectively avoids overfitting during training and has good generalization capabilities.

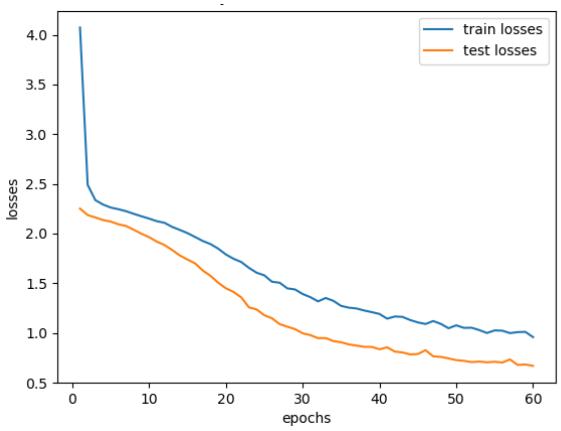

Figure 2. Training and testing losses for hybrid training methods

4.2. Single method adversarial training

In order to compare with the hybrid adversarial training method, the model was adversarially trained using the FGSM method and the PGD method respectively. The accuracy of the FGSM model on the original data set was 82%, and the accuracy of the PGD model on the original data set was 79%. As shown in Figure 3, the training loss of FGSM adversarial training drops from the initial 4.0 to 0.7, while the test loss stabilizes at around 0.6, which indicates that the model has good robustness and generalization ability. In contrast, the training loss of PGD adversarial training drops from the initial 4.0 to 0.8, and the test loss stabilizes at around 0.7. The training loss value of PGD is always larger than that of FGSM because the adversarial examples generated by PGD are more powerful and the learning process of the model is more difficult. In addition, the loss curve of PGD shows some fluctuations, indicating that the training process of the model is more complicated when dealing with more challenging adversarial examples.

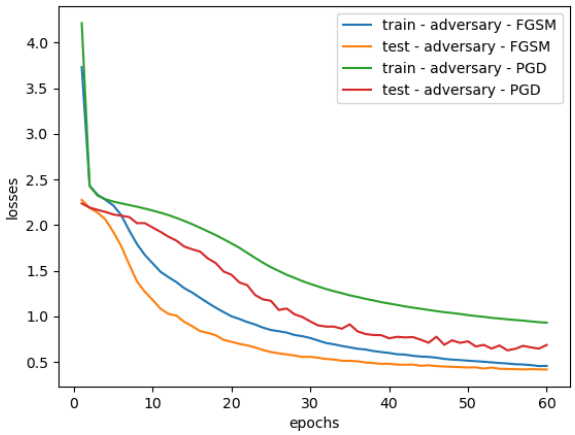

Figure 3. FGSM and PGD training and testing losses.

DOI: 10.54254/2755-2721/109/20241413

4.3. Comparative Analysis

Table 1. Defense effects of FGSM and PGD adversarial training models and hybrid adversarial training

Attack Methods	Perturbation value	Step Length	Iterations	Accuracy of adversarial training models	Accuracy in mixed adversarial training	The accuracy of the original model
FGSM	2/255	-	-	81%	80%	56%
FGSM	4/255	-	-	80%	80%	36%
FGSM	8/255	-	-	79%	77%	16%
PGD	8/255	3/255	7	78%	75%	2%
PGD	17/255	3/255	7	78%	74%	0%
PGD	160/255	40/255	7	75%	74%	0%
PGD	8/255	3/255	12	78%	77%	2%
PGD	17/255	3/255	12	77%	76%	0%
PGD	160/255	40/255	12	74%	71%	0%
PGD	8/255	3/255	20	77%	75%	1%
PGD	17/255	3/255	20	76%	73%	0%
PGD	160/255	40/255	20	72%	69%	0%

It can be seen from Table 1 that the two adversarial training methods (FGSM and PGD) have good defense effects when facing attacks with perturbation values and iteration times not higher than the adversarial training parameters, and can generally achieve an accuracy rate of more than 77%. For example, the FGSM method uses a perturbation value of 8/255 during adversarial training, and its accuracy on the adversarial training model is 79%, and the accuracy on the hybrid adversarial training is 77%. Similarly, the PGD method achieves 78% accuracy on the adversarially trained model and 75% accuracy on the hybrid adversarial-trained model using a perturbation value of 8/255, a step size of 3/255, and 7 iterations. However, when the perturbation value of the attack is higher than the parameter of adversarial training, the defense effect of the model decreases significantly. For example, when the PGD attack uses larger perturbation values of 160/255, the accuracy of the model decreases significantly whether it is 7 iterations or 12 iterations. The accuracy of the adversarial training model dropped to 74%, and the accuracy in hybrid adversarial training was only 71%-74%. Compared to the defense effect at low disturbance values, the accuracy dropped by about 8%.

In general, compared with single FGSM and PGD adversarial training, the defense performance of hybrid adversarial training when facing FGSM attacks is almost the same as that of single FGSM adversarial training. However, the defense performance when facing PGD attacks decreases more significantly, by about 2% to 3%. The reason for the small gap may be that the strength of the adversarial attack parameters used in this paper is mostly lower than the strength of the parameters used in adversarial training, and the RestNet-18 structure used is relatively simple. The gap between the two may be even greater with more complex models, and data sets, and when facing more powerful attacks. In addition, hybrid adversarial training has a higher accuracy than PGD adversarial training on the original test set, about 2% higher, and takes less training time.

5. Conclusion

In summary, this paper first uses a hybrid adversarial training strategy of FGSM and PGD to train the ResNet-18 model, then uses a single method adversarial training strategy to train the model, and then conducts adversarial attack tests on the models trained using the two different strategies. Through comparative analysis, this paper concludes that the hybrid adversarial training combining FGSM and PGD can effectively reduce the training time. The accuracy of the original data set is higher than that of

DOI: 10.54254/2755-2721/109/20241413

PGD adversarial training, which is improved by 2%. The defense performance when facing an FGSM attack is almost the same as that of single FGSM adversarial training. The defense performance when facing PGD attack decreases significantly, which is about 2% to 3%.

The limitation of this study is that the selected model and dataset are relatively simple, and no experiments are conducted on more complex models and datasets, resulting in a small difference in the defensive performance comparison of the two different adversarial training strategies.

For future development directions, it is recommended to adopt multiple strategies in adversarial training, such as mixing multiple adversarial training methods, or introducing other methods to enhance training, such as regularization, knowledge transfer, etc.

References

[1] Mcdaniel, P., Papernot N and Celik Z B 2016 Machine Learning in Adversarial Settings IEEE Security & Privacy 14(3): 68-72
[2] Li M H et al. 2021 Adversarial attacks and defenses against deep learning models Journal of Computer Research and Development 58(5) p 18
[3] Goodfellow I J, Shlens J and Szegedy C 2014 Explaining and Harnessing Adversarial Examples Computer Science
[4] Kurakin A, Goodfellow I and Bengio S 2016 Adversarial Machine Learning at Scale arXiv
[5] Tramèr F et al. 2017 Ensemble Adversarial Training: Attacks and Defenses
[6] Madry A et al. 2017 Towards Deep Learning Models Resistant to Adversarial Attacks
[7] Miyato T et al. 2018 Virtual Adversarial Training: A Regularization Method for Supervised and Semi-Supervised Learning IEEE Transactions on Pattern Analysis and Machine Intelligence pp 1–1
[8] Song C 2018 MAT: A Multi-strength Adversarial Training Method to Mitigate Adversarial Attacks in 2018 IEEE Computer Society Annual Symposium on VLSI (ISVLSI) 476-481
[9] Netzer Y, Wang T, Coates A et al. 2011 Reading Digits in Natural Images with Unsupervised Feature Learning
[10] Simonyan K and Zisserman A 2014 Very Deep Convolutional Networks for Large-Scale Image Recognition Computer Science
[11] He K et al. 2016 Deep Residual Learning for Image Recognition IEEE