Dean&Francis
ISSN 2959-6157

The Migration Analysis of Adversarial Examples with Convolutional Neural Networks

Jin Xing[1, *]

[1]Southern University of Science and Technology, Shenzhen, China

*Corresponding author: 12210433@mail.sustech.edu.cn

Abstract:

Recently, with the attention of researchers on the adversarial attack technology, the robustness of neural networks has become an urgent problem. An adversarial attack is a way to mislead the deep-learning neural networks and make several changes to the sample, which lets the model provide the wrong output with a high confidence level. Using these methods to attack some specific deep learning models achieves remarkable results, but the robustness of different neural network models has not yet been clarified. This paper studies the migration of adversarial examples, aiming to conclude whether the adversarial examples from specific models are also practical when applied to other models. Through this process, the fragility of neural networks when operating with adversarial attacks is universal and can be analyzed. The primary dataset is from the cifar-10 dataset, including ten classes of natural item images with RGB channels. The deep learning models are LeNet, ResNet18, and VGG16, which use the fast gradient sign method (FGSM). The attacked models generate incorrect samples, utilized in the other two models to demonstrate effective test performance. The result indicates that the attack on a specific neural network model cannot disturb other models.

Keywords: Adversarial Examples; Neural Network; Convolutional Neural Networks; Deep Learning.

1. Introduction

In recent years, deep-learning neural networks have shown their advantages and power in data management or computer vision. Due to its ability, modern development on generative artificial intelligence and other applications. Considering the correction and safety of the practical application, neural network models need to include accurate expression and robust functions [1].

There is sufficient research that if adding a slight perturbation to the input of a neural network model, most current models, such as fully connected neural networks or convolutional neural networks, will

ISSN 2959-6157

identify the wrong answer and provide incorrect output. Moreover, Kurakin et al. found that if the generated adversarial example images were printed, the neural network model would produce disparate classification results under different light and orientation conditions. Similarly, researcher found that objects from adversarial 3D printing technology can confuse neural network models in different orientations and sizes [2].

Adversarial attacks can be classified into white-box and black-box attacks, representing attacks with full knowledge of the network structure and attacks without knowledge of the data set and model. Szegedy et al. first showed that adding a slight disturbance to a digital image can mislead a neural network model into making a false classification. Later, the researchers found that adding adversarial samples to the training process could improve the robustness of the model. To improve the computational efficiency of adversarial samples, Goodfellow et al. proposed a method for rapidly generating adversarial samples named fast gradient sign method (FGSM) [3, 4]. FGSM only needs one back propagation process to generate adversarial samples, and as a type of white-box attack, it can effectively confuse the specific model.

This paper focuses on the migration of adversarial examples generated based on FGSM. Specifically, this paper selects three convolutional neural networks with different scales, including LeNet, VGG16 and ResNet18. It conducts countermeasures based on FGSM on these three convolutional neural networks and generates adversarial examples with different attack intensities. Then, this paper applies adversarial examples generated based on different convolutional neural networks to attack to verify whether the adversarial examples generated by FGSM are general. The experimental results show that the migration characteristics of mobile are not noticeable when the attack intensity is weak. The mobile generated by different convolutional neural networks can effectively attack other convolutional neural networks with increased attack intensity. All the experiments in this paper are based on the CIFAR-10 dataset.

In the second section, this paper summarizes the related technologies, including FGSM and three kinds of convolutional neural networks, applied in this paper. The third section describes the experimental results and analyzes the critical phenomena. This paper summarizes and looks forward to the future research plan in the fourth section.

2. Methods

2.1 Fast Gradient Sign Method

FGSM is short for the fast gradient sign method [4]. It is a kind of white-box attack. The basic principle of the FGSM is that it uses the backpropagation algorithm to calculate the detailed gradient of the model from the input data and then determines the direction of the slight perturbation according to the gradient. Specifically, the algorithm adds a perturbation value in the direction where the loss function increases the fastest, thus generating an adversarial sample. By changing the epsilon's size, the perturbation value's magnitude can be adjusted to change the intensity of the FGSM attack.

$$x' = x + \epsilon sign\left(\nabla_x L\left(\theta, x, t\right)\right) \tag{1}$$

In this formula, x' represents the adversarial sample, x represents the original input, ϵ is the size of epsilon, L is the loss function of specified model, and *sign* function is used to get the sign of gradient.

When training the network, FGSM obtains the input image features to get the classification probability through the SoftMax or sigmoid layer. Then, it uses gradient backpropagation, which means that the resulting classification probability and the actual label are used to calculate the loss value. The loss value is later returned, and the gradient is calculated. When the disturbed image is tested for the classification network, the calculated gradient direction is added to the input image to make the loss value more significant than the loss value of the entire image.

2.2 LeNet

LeNet is one of the earliest convolutional neural network models. In 1998, it was first proposed for image classification and achieved remarkable success in handwritten character recognition. It consists of three modules with continuous convolutional layers and pooling layers. Specifically, the first and second modules include a 5x5 convolutional layer and a 2x2 pooling layer, while the first module has six channels, and the second module has 16 channels. The third module has a 5x5 convolutional layer with 120 channels, reducing the images' size to 1. Then, the feature extracted after this convolution is input into the fully connected layers. Finally, the SoftMax activation function is utilized to calculate the confidence [5].

2.3 ResNet18

ResNet, the 2015 ImageNet competition's winner, reduced the image classification error rate to 3.6%, which even exceeds the accuracy of normal human eyes. With the continuous development of deep learning, the number of layers of the model is increasing, and the network structure is becoming more and more complex. Theoretically, assuming that the newly added layers are identical mappings, as long as the original layer learns the same parameters as the original model, the deep model structure can achieve

the effect of the original model structure. However, practice shows that the training error tends to increase rather than decrease when the number of layers is increased. The residual network ResNet is proposed to solve this problem. ResNet18 is a classic ResNet series model. The name ResNet18 comes from 18 convolution layers contained in its network structure. The model solves the problems of gradient disappearance and gradient explosion in deep convolutional neural networks by introducing Residual Block, thus allowing the network to reach a deeper level while maintaining good performance [6].

2.4 VGG16

VGG was proposed 2014 through a series of 3x3 convolutional and pooling layers. It is widespread and practical due to its simple structure and directed deep learning network model design. In VGG16, there are 13 layers of convolution and three fully connected layers. All convolutional layers are 3x3 in size of the kernel, and VGG uses pooling layers to obtain the feature. After every convolution, it uses ReLU to activate, and after fully connected layers, it uses dropout to avoid over-fitting. The success of the VGG model proves that increasing the depth of the neural network allows for a better learning of the feature patterns in the image. The "16" in VGG16 means that the network contains 16 weight layers (convolution layer and full connection layer), which makes it a relatively deep neural network structure [7].

3. Experimental Results and Analysis

3.1 Dataset Description

One of the most representative datasets for colour picture datasets is CIFAR-10 [8]. Ten types of RGB colour images: truck, plane, car, bird, cat, deer, dog, frog, horse, and cat. These categories are mutually exclusive, and images appearing in one category will not appear in others. Each of the 10,000 tests and the 50,000 training images in CIFAR-10 are 32 × 32 RGB three-channel images. Compared with the handwritten image dataset, the CIFAR10 colour image dataset has higher complexity, a more decadent sample size, and a more robust representation. CIFAR-10 is chosen as the test data set since it is appropriate for this investigation. In this paper, the CIFAR-10 dataset is trained and tested using LeNet, ResNet-18, and VGG16 neural network models. Moreover, FGSM was used to attack and get the experimental results.

3.2 Results and Analysis

Table 1. Attack effect of FGSM attack method on CNNs

Epsilon (ϵ)	0	0.05	0.1	0.15	0.2	0.25	0.3
LENET	0.53	0.20	0.06	0.02	0.01	0.00	0.00
RESNET	0.77	0.02	0.02	0.02	0.02	0.03	0.04
VGG	0.80	0.06	0.02	0.02	0.02	0.03	0.04

Using the LeNet model to train and test CIFAR-10, in the case of epsilon is 0, the accuracy is only 0.53. The ResNet-18 model has a correct rate of 0.77 when Epsilon=0, which shows that the ResNet-18 model is better than LeNet. Of course, the capacity of the ResNet-18 model should be more significant. The accuracy of the VGG16 model is the highest among the three models, reaching 0.80.

When the FGSM attack is added, the accuracy of the three models declines. It is not difficult to show that the FGSM attack is effective for the three models of LeNet, ResNet18, and VGG16, as shown in Table 1.

Table 2. Attack effect from adversarial examples (LeNet) to ResNet18 and VGG16

Epsilon	ResNet18	VGG16
0.05	0.59	0.55
0.1	0.58	0.54
0.15	0.56	0.52
0.2	0.54	0.50
0.25	0.52	0.46
0.3	0.50	0.44

ISSN 2959-6157

Table 3. Attack effect from adversarial examples (VGG16) to LeNet and ResNet18

Epsilon	LeNet	ResNet18
0.05	0.81	0.69
0.1	0.71	0.43
0.15	0.51	0.26
0.2	0.32	0.19
0.25	0.22	0.15
0.3	0.16	0.14

Table 4. Attack effect from adversarial examples (ResNet18) to LeNet and VGG16

Epsilon	LeNet	VGG16
0.05	0.78	0.49
0.1	0.65	0.29
0.15	0.49	0.21
0.2	0.36	0.18
0.25	0.29	0.16
0.3	0.24	0.16

This paper tests another model with some adversarial examples that one model incorrectly predicts to determine whether the adversarial examples generated by FGSM are universal to different models, the results as shown in the table. They show that the adversarial examples generated based on LeNet have achieved good performance in the ResNet attack test, which proves its good generalization. Interestingly, the adversarial examples generated based on ResNet have almost no effect in the LeNet attack test. This paper holds that the experimental results are not random phenomena. One possible reason is a correlation between the original performance and the model's generalization. Specifically, the low performance of LeNet may improve the generalization of its counter samples. This paper will design more good experiments in the future to verify this conclusion, as shown in Table 2, Table 3, Table4.

4. Conclusion

By deliberately adding some subtle interference to the input samples, the model gives a wrong output with high confidence. This attack challenges the security of artificial intelligence systems and poses potential risks to users' privacy and data security. The migration of adversarial examples is an important research field, which refers to the ability of adversarial examples generated for a specific model to successfully attack other models with different structures or parameters. Based on the CIFAR-10 data set, this paper tests the attack of FGSM on three different scale convolutional neural networks and analyzes the migration of the adversarial examples. This paper will further study the characteristics of adversarial examples using feature analysis in future work.

References

[1] Chen Yu; Xing Yang; Haoli Xu, et al. Research Progress of Physical Adversarial Examples for Anti-Intelligent Detection//2024 5th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT). IEEE, 2024: 175-181.

[2] Yao Huang, Yinpeng Dong, Shouwei Ruan, et al. Towards Transferable Targeted 3D Adversarial Attack in the Physical World//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 24512-24522.

[3] Linyu Tang, Lei Zhang. Robust Overfitting Does Matter: Test-Time Adversarial Purification With FGSM //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 24347-24356.

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

[5] Bharati Yadav, Ajay Indian, Gaurav Meena. Recognizing Offline Devanagari Handwritten Characters Using Modified Lenet-5 Deep Neural Network. Procedia Computer Science, 2024, 235: 799-809.

[6] Serena Sunkari, Ashish Sangam, Venkata Sreeram P, et al. A refined ResNet18 architecture with Swish activation function for Diabetic Retinopathy classification. Biomedical Signal Processing and Control, 2024, 88: 105630.

[7] Chittathuru Himala Praharsha, Alwin Poulose. CBAM VGG16: an efficient driver distraction classification using CBAM embedded VGG16 architecture. Computers in biology and medicine, 2024, 180: 108945.

[8] JIAWEI DU, Qin Shi, Joey Tianyi Zhou. Sequential subset matching for dataset distillation[J]. Advances in Neural Information Processing Systems, 2024, 36.