Migration of adversarial example in image classification

Dongrun Gu

School of Electronics Engineering & Computer Science, Peking University, Beijing, 100871, China

2000013190@stu.pku.edu.cn

Abstract. Neural network technology has made remarkable achievements in computer vision, speech recognition, natural language processing, and other fields. However, the problem of the interpretability of the neural network model makes its application in real situations have potential security risks. In recent years, many studies have pointed out that using Adversarial example technology to make extremely weak perturbations of the input sample can mislead most mainstream neural network models, such as fully connected neural networks and convolutional neural networks, to make wrong judgments. This phenomenon reveals that the existing neural network technology lacks security and robustness. The study of adversarial example technology is of great significance to improve the safety and robustness of neural network models and to promote the researcher's understanding of the learning process of neural network models deeply. Studying adversarial examples of migration is an important research field in adversarial attacks. Researchers attempt to summarize the rules of adversarial attacks by exploring the migration of adversarial examples, thus establishing a robust model in the deep learning area. In this paper, the migration of adversarial examples in image classification is studied to provide analytical data for summarizing the characteristics of adversarial examples.

Keywords: neural network, machine learning, adversarial sample, computer vision, adversarial attacks

1. Introduction

Neural network technology has made rich research achievements in many tasks in machine learning, such as computer vision and text mining [1-4]. On some specific data sets, it is close to or even surpasses the accuracy of human beings. Due to its excellent performance, neural network technology has been gradually applied to many real application scenarios in recent years, such as automatic driving, face recognition, and financial risk assessment. The neural network model should have high accuracy, robustness, and security in the application scenario. Recently, some researchers have found that most of the current mainstream neural network models (like fully connected neural networks, CNN) can produce wrong output results by only slightly perturbing the input of neural network models (such as digital images) to generate adversarial examples [5]. Adversarial attack technology reveals the vulnerability of the deep learning model to some extent. In recent years, many adversarial attack and defense technologies have been put forward, such as QEBA, SAA for attacking and Mix-up Inference, and Adversarial Training for defense [6-9].

From the amount of information needed to generate confrontation samples, the existing attack techniques can be divided into three types: the attack in the white box, the attack in the black box, and the attack in the gray box. Among them, the attack in the white box requires a prior understanding of the model's structure and its parameters and generating adversarial samples accordingly. Compared to white box attacks, attack methods in the black box do not need the details of the model to generate adversarial samples. Grey box attack, on the other hand, is somewhere in between, using a white box attack in the part stage of sample generation and a black box attack in the other stage. According to the attack mode and attack target, attack technology can also be divided into poison attack (in training set to add adversarial samples, make the overall accuracy of the neural network model decline), a targeted attack (the neural network model misled to a specific error results), no targeted attack (no specific misleading target).

At present, the research of adversarial sample technology is mainly focused on the field of computer vision. One of the essential reasons is that digital images are continuous, so it is not difficult to generate the corresponding adversarial samples through gradient descent or other methods. At the same time, humans can intuitively perceive the quality of antagonistic samples from the naked eye. However, since the input data has prominent discrete characteristics in the text field, the adversarial sample technology in the computer vision field can not be easily applied. At this stage, although a series of attack technologies (such as generating misspelled words as adversarial samples, adversarial perturbation samples for word embedding) and defense techniques (grammar detection) is proposed, the adversarial sample technology in the text field is still in its infancy [10-12]. In addition, due to the wide use of voice recognition technology, the confrontation sample technology in the audio field also needs to attract the attention of researchers. Also, it needs to be concerned about the adversarial sample technology in real application scenarios.

Researchers want to know whether the interference image is strong enough for other models when using the FGSM method to generate an interference image for an image recognition model. Once this is true, it means that the interference image generated by the FGSM method has the universal interference ability to different models. That is, researchers can conduct interference attacks without knowing the architecture of the other party's image recognition model. This paper selected the LeNet model and ResNet18 model for testing, and the test data was taken from CIFAR10. This paper first test the performance of LeNet and ResNet18, respectively, after being attacked by FGSM, and then attack another model with the confrontation samples generated to obtain their respective accuracy. The experiment results indicate that the FGSM method has a specific impact on the adversarial samples generated by one model on other models. However, the degree of impact is related to the specific model.

2. Methods

In this paper, an attack technique called Fast Gradient Sign Attack (FGSM) is taken as the research object, focusing on whether the adversarial examples generated by FGSM in different models are migratory. Specifically, FGSM is used to attack two classical convolutional neural networks, LeNet and ResNet, to evaluate their robustness [13]. Then, this paper collects the adversarial examples generated on the two models and uses these confrontation samples to cross-attack the two models and verify the confrontation samples' migration. This section introduces the attack technology FGSM and two models of LeNet and ResNet in detail.

2.1. Attack Technolgy

Fast Gradient Sign Attack (FGSM). FGSM is a classic image countermeasure algorithm. It learns the features of the input image during network training. It then obtains the classification probability through the softmax or sigmoid layer. This probability is used to calculate the loss value with the real label, return the loss value, and calculate the gradient (i.e., gradient backpropagation). So the only need for us is to calculate the gradient direction and add it to the input image. Then loss value becomes greater in passing through the classification network. $J(\theta, x, y)$ indicates the cost function in training process. θ indicates the parameters of the model. The input of this model is recorded by x, while y is the targets related to x. The cost function can be linearized around θ to get a best max-norm $\eta = sign\left(\nabla x\, J(\theta, x, y)\right)$ [14].

2.2. Convolutional Neural Networks

LeNet is the pioneering work of convolutional neural network, and also a milestone to promote deep learning to prosperity. Yann LeCun put forward LeNet in the 1990s. He first adopted two new neural network components: convolution layer and pooling layer; LeNet has achieved remarkable accuracy in handwritten character recognition.

LeNet has a series of versions, among which LeNet-5 version is the most famous and the best version in LeNet series. LeNet-5 uses five convolution layers to learn image features; The weight sharing feature of convolution layer makes it save a lot of computation and memory space compared with full connection layer; At the same time, the local connection feature of convolution layer can ensure the spatial correlation of images [15].

LeNet consists of seven layers, namely, C1, C3, C5 convolution layer, S2, S4 downsampling layer (also known as pooling layer). F6 is a full connection layer, and the output is a Gaussian connection layer. This layer uses the softmax function to classify the output images. In order to correspond to the model input structure, the 28 * 28 image in MNIST is expanded to 32 * 32 pixels. Each layer is described in detail below. The C1 convolution layer consists of 6 convolution kernels of different types with a size of 5 * 5. The step size of the convolution kernel is 1, and there is no zero filling. After convolution, six 28 * 28 pixel feature maps are obtained; S2 is the maximum pooling layer. The size of it is 2 * 2, while the step size of it is 2. After S2 pooling, six 14 * 14 pixel feature maps are obtained; C3 convolution layer consists of 16 different convolution kernels whose size are all 5 * 5. The step size of the convolution kernel is 1, and there is no zero filling. After convolution, 16 characteristic images with a size of 10 * 10 pixels are obtained; S4 is the maximum pooling layer. The size of it is 2 * 2, while the step size of it is 2. After S4 pooling, 16 feature maps with the size of 5 * 5 pixels are obtained; The C5 convolution layer consists of 120 different convolution kernels with a size of 5 * 5. The step size of the convolution kernel is 1, and there is no zero filling. After convolution, 120 characteristic images with a size of 1 * 1 pixel are obtained; 120 characteristic maps with the size of 1 * 1 pixel are spliced together and become an input of layer F6. In addition, F6 layer is a fully connected hidden layer composed of 84 neurons. The activation function uses sigmoid function; The final output layer is a softmax Gaussian connection layer composed of 10 neurons, which can be used for classification tasks

Residual Neural Network (ResNet) was proposed by Kaiming He of Microsoft Research Institute and four other Chinese. ResNet successfully trained 152 layers of neural networks using ResNet Unit, and won the championship in ILSVRC in 2015. The error rate of the top five was 3.57%. At the same time, the number of participants is lower than VGGNet. The effect of it is commendable. The training process of neural network is sped up thanks to the structure of ResNet. Besides, the accuracy of the model has a great improvement. The main idea of it is adding a channel to the network for direct connection which allows to retain a certain proportion of the output from previous layer. The original input can be directly transmitted to the next layers as well.

ResNet is modified based on VGG19 network. In ResNet, residual units are added through short circuit mechanism. The main difference between them is ResNet uses convolution of street=2 in the process of down sampling directly, and global average pool layer is used to replace the full connection layer. The number of the feature map will become double when the size becomes half to keep the complexity. A short circuit mechanism is also used between different layers. In this way, residual learning are formed.

3. Experimental Results and Analysis

3.1. Data Description

Sixty thousand color images are included in the CIFAR-10 dataset, divided into ten categories (6000 images for each category) and 32*32. Among these images, 50000 are for training, and every 10000 images form training batches. Another 10000 images are for testing and constitute a separate batch.

The CIFAR10 dataset is used to supervise learning and training. Each sample must be equipped with a tag value (to distinguish what the sample is). Different types of objects use different tag values. There are ten types of objects in CIFAR10, and the tag values are distinguished according to 0~9, respectively, aircraft, automobile, bird, cat, deer, dog, frog, horse, ship, and truck. These ten categories are listed in the figure below, and each category shows ten random pictures.

3.2. Results and Analysis

First, this paper uses FGSM to test the effect of the LeNet model, used 10000 pictures to test, and interfered with the correct pictures predicted by the model. The experimental results as shown in Table 1. Only a low performance (0.5336) can be obtained using LeNet to process cifar10 data sets. At the same time, with the increasing adversarial intensity, the performance of LeNet has declined rapidly, proving the FGSM algorithm's effectiveness. This paper collects the adversarial examples that make LeNet misjudge for the next experiment.

Table 1. Accuracy of LeNet.

Epsilon	Test Accuracy
0	0.5336
0.05	0.1964
0.1	0.0613
0.15	0.0218
0.2	0.0089
0.25	0.0045
0.3	0.0028

Then, this paper uses FGSM to test the effect of the ResNet18 model, uses the same 10000 pictures to test, and interferes with the correct pictures predicted by the model—the results are shown in the table 2. Without the FGSM attack, the performance of ResNet is much better than that of LeNet. This result is in line with expectations because the model size of ResNet is also significantly more significant than that of LeNet. However, with the increase of FGSM's attack intensity, the performance of ResNet drops significantly faster. This experimental phenomenon is fascinating. Because, larger-scale models should have better robustness. This paper collects the adversarial examples that make ResNet misjudge for the next experiment.

Table 2. Accuracy of ResNet18.

Epsilon	Test Accuracy
0	0.7665
0.05	0.0206
0.1	0.0175
0.15	0.0204
0.2	0.0202
0.25	0.0256
0.3	0.0364

This paper test another model with some adversarial examples that one model incorrectly predicts to determine whether the confrontation samples generated by FGSM are universal to different models, the results as shown in the table 3 and the table 4. 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.

Table 3. Accuracy of ResNet18.

Epsilon	LeNet	ResNet18
0.05	0	0.7805
0.1	0	0.6536
0.15	0	0.4855
0.2	0	0.3579
0.25	0	0.2871
0.3	0	0.0175

Table 4. Accuracy of LeNet.

Epsilon	ResNet18	LeNet
0.05	0	0.5926
0.1	0	0.5810
0.15	0	0.5626
0.2	0	0.5433
0.25	0	0.5253
0.3	0	0.5004

4. Discussion

The interpretability problems of neural network models and the robustness and security problems leaked in recent years still have a long way to go. At the same time, as a strategy and means to explore neural network model learning mechanisms, the research of the sample technology has achieved rich results. However, many problems still need to be solved, such as why effective and general (image, text are practical) against sample generation technology. At the same time, the research on the sample technology from systematic attack and defense theory system, how to evaluate the neural network model security, robustness, and how to measure the quality of the samples still has no consensus in the research community.

Universal adversarial sample generation techniques. Currently, most of the adversarial sample generation technology research focuses on the image field. In contrast, in the text field, because humans quickly detect the data due to the significant discrete characteristics, the slight perturbations are impossible to copy the adversarial sample generation method in the image field. This will prompt researchers to break the limitations of existing definitions of adversarial samples and study new adversarial sample generation methods.

Universal adversarial sample attack techniques. The generality here refers to the idea that the anti-sample attack technology should attack a class of neural network models with similar structures and properties. Most current research on adversarial sample attack techniques has focused on several specific neural network models. The study of general adversarial sample attack techniques is precious for profoundly understanding the learning mechanism of neural network models and discovering the potential defects of neural network models.

Measures of the robustness and safety of the neural network models. Researchers attach far more importance to the accuracy of neural network models than their robustness and safety. So far, the relevant research on neural network technology has not been considered from the perspective of information security, forming a theoretical basis and systematic and quantitative standards to measure the robustness and security of neural network models.

5. Conclusion

Studying adversarial examples of migration is an important research field in adversarial attacks. Researchers attempt to summarize the rules of adversarial attacks by exploring the migration of adversarial examples, thus establishing a robust model in deep learning area. In this paper, the migration of adversarial examples in the task of image classification is studied to provide analytical data for summarizing the characteristics of adversarial examples. The results of the experiment show that the adversarial examples of one model can also affect other models, but the impact is not as expected. At the same time, this experimental result reflects an interesting phenomenon. ResNet18 is higher than LeNet in terms of architecture complexity. However, the countermeasure samples that can interfere with ResNet18 can hardly interfere with the simpler LeNet. Therefore, there is no positive correlation between the ability to resist other models against samples and the complexity of the model. One guess is that LeNet itself, due to its simple structure, does not extract pixel information as fully as ResNet18, so it is not sensitive to attacks of other models against samples. However, it is worth noting that some research work on substitute attacks shows the migration of adversarial examples. This paper will design more good experiments to verify this view in future work.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in NIPS’12 Proceedings of the 25th International Conference on Neural Information Processing Systems, vol. 1, 2012, pp. 1097–1105.

[2] S. Ren, K.He, R.Girshick, and J.Sun, “Fasterr-cnn: Towards real time object detection with region proposal networks,” in NIPS’15 Proceedings of the 28th International Conference on Neural Information Processing Systems, vol. 1, 2015, pp. 91–99.

[3] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” in NIPS’14 Proceedings of the 27th International Conference on Neural Information Processing Systems, 2014, p. 31043112.

[4] H. Xu, M. Dong, D. Zhu, A. Kotov, A. I. Carcone, and S. NaarKing, “Text classification with topic-based word embedding and convolutional neural networks,” in BCB ’16 Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, 2016, pp. 88–97.

[5] C. Szegedy et al. ‘‘Intriguing properties of neural networks.’’ arXiv preprint arXiv: 1312.6199, 2013.

[6] Huichen Li, Xiaojun Xu, Xiaolu Zhang, Shuang Yang, Bo Li. QEBA: Query-Efficient Boundary-Based Blackbox Attack,arXiv preprint arXiv:2005.14137,2020

[7] Kaidi Xu, Sijia Liu, Pu Zhao, Pin-Yu Chen, Huan Zhang, Quanfu Fan, Deniz Erdogmus, Yanzhi Wang, Xue Lin. Structured Adversarial Attack: Towards General Implementation and Better Interpretability,arXiv preprint arXiv:1808.01664,2018

[8] Tianyu Pang, Kun Xu, Jun Zhu.Mixup Inference: Better Exploiting Mixup to Defend Adversarial Attacks,arXiv preprint arXiv:1909.11515,2019

[9] Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” in *International Conference on Learning Representations*, 2015. [Online]. Available: http://arxiv.org/abs/1412.6572

[10] H Xu et al. Adversarial Attacks and Defenses in Images, Graphs and Text: A Review, arXiv preprint arXiv: 1909.08072, 2019.

[11] W. E. Zhang et al, Adversarial Attacks on Deep Learning Models in Natural Language Processing: A Survey, arXiv preprint arXiv: 1901.06796, 2019.

[12] W. Wang et al, Towards a Robust Deep Neural Network in Texts: A Survey, arXiv preprint arXiv: 1902.07285v5, 2019.

[13] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition," in *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278-2324, Nov. 1998, doi: 10.1109/5.726791.

[14] Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy. Explaining and Harnessing Adversarial Examples,arXiv preprint arXiv:1412.6572,2015

[15] Yuhas B. and Ansari N., Neural Networks in Telecommunications, Kluwer Academic Publishers, UK, 1994.