Method of Comparison of Neural Network Resistance to Adversarial Attacks

Alexey Nemchenko[1(✉)] and Sergey Bezzateev[2(✉)]

[1] ITMO University, St. Petersburg 197101 Kronverksky Pr. 49,, Russia
nemc_aleks@mail.ru
[2] Saint Petersburg State University of Aerospace Instrumentation, St. Petersburg 19000 B. Morskaya 67,, Russia
bsv@aanet.ru

Abstract. The vulnerability of neural networks to adversarial attacks has long been revealed. However, the structure of neural networks is not given due attention during the attack. The article deals with the impact of different parameters of a neural network on its resistance to adversarial attacks. The main purpose of this research is to determine which parameters increase resistance to attacks. The way by which neural networks can be compared has been proposed. Several neural networks were selected for comparison and a number of adversarial attacks were conducted on them. As a result, certain conditions were identified under which the attack took place over a longer time. It was also found that different changes in neural network parameters were required to protect against different attacks.

Keywords: Neural networks · Information security · Training finder · Machine learning

1 Introduction

Recent advances in the field of machine learning have significantly expanded the scope of artificial neural networks. Particularly great success has been observed in the application of convolution neural networks. Thanks to them, incredible performance in the recognition of objects has been achieved. In this case, the recognition accuracy reaches 95% and sometimes exceeds the human abilities.

However, neural networks have disadvantages not inherent in humans. It is on the basis of such disadvantages are built adversarial attacks on the neural network [1]. For example, for a convolution neural network used to classify images, adding perturbations to the recognition object can be critical. Such distortions can lead to classification errors. At the same time, a person will not even notice such perturbations. In this work we investigate the possibility of determining ways to increase the resistance of convolution neural networks to adversarial attacks. We considered a neural network in which changes were made during the experiment. Training was conducted on the same data and the same attacks were performed.

O. Galinina et al. (Eds.): NEW2AN 2020/ruSMART 2020, LNCS 12526, pp. 74–79, 2020.
https://doi.org/10.1007/978-3-030-65729-1_7

The main objective of the study was to characterize how various factors influence the effectiveness of adversarial attacks. Such as changing the size of the training sample, changing the number of cores, changing the activation functions, and combining these methods. Based on the results obtained, a method was formed that allows comparing different neural networks by their degree of resistance to competitive attacks.

2 Related Work

To affect neural networks, attacks are used either during training by manipulating the data in the training sample [2] or during the operation of the neural network by affecting the data to be classified [1]. To prevent such attacks, adding noise [3], using high-level representation guided denoiser [4], learning on data containing competitive examples [1] are used. Researches in this area are mainly aimed at counteracting attacks at a specific stage, without assessing the impact of the parameters of the neural network itself on its overall security. In this paper, the study of the influence of various parameters of the neural network on the degree of its resistance to competitive attacks is considered.

3 The Proposed Way to Compare Different Neural Networks

Exposure to a normal neural network means attacks that use information to recognize. Due to the fact that it is impossible to control such information, the methods of protection adopted must be designed with the potential attacks in mind.

The first step is to determine if the neural network parameters are compromised. Most adversarial attacks use the white box principle when there is complete information about the neural network. If the information about the parameuters of the neural network used in the system is publicly available, the range of attacks on it is increased. It takes very little time from the moment an attacker receives the network's data to the moment he is able to submit a competitive example to the neural network's entry.

If the parameters of a neural network are not considered fully protected from receiving them by an intruder, due to receiving them from a third party source or for other reasons, the security of the neural network can be analyzed by determining the variability of neural network parameters. In cases where it is impossible to hide the original parameters, their variability will significantly complicate the attack. The moment an attacker generates an image, the neural network parameters will already change and the attack will not be so successful. However, retraining is possible if the training sample is different from the original one. Achieving such a result is also possible with constant retraining of the neural network, in this case it is not necessary to determine the time when the parameters were disclosed, but only to determine the time required to change the parameters of the Artificial Neural Network.

The second step is to analyze the parameters of the training sample. Increasing the size of the training sample allows increasing the accuracy of the neural network. Adding special noise or examples of competitive attacks to the training sample makes it more resistant to FGSM and DeepFool attacks. However, simply increasing the size of the training sample has almost no effect on the FGSM attack. At the same time, when the size of the tutorial sample is reduced, it increases the duration of attacks in which

the attack method targets a minimum number of modified pixels, such as JSMA and OnePixel. The assessment is performed by performing a DeepFool attack, due to the fact that this attack is the fastest among those considered, which can be influenced by changing the size of the training sample.

The third step is to define the activation functions of the neural network. For several attacks, their effectiveness depends on the selected activation function. The analysis will be performed by determining the activation function used, if the activation function has an upper value limit of the parameter, then it is considered vulnerable. Such function is Bounded ReLU [3], neural networks with such function become more resistant to JSMA and OnePixel attacks. This is because if this type of function is used, the ability to assign excessive weight to one or more pixels will be limited.

The fourth step Determine the number of parameters processed by the neural network. The increase in the number of parameters leads to an increase in the duration of an attack such as PGD. Neural networks that have increased the number of cores and therefore the number of parameters can better withstand DeepFool attacks.

Step five, conduct a series of adversarial attacks on the neural network. In this case it is not recommended to conduct those attacks that were used to train the neural network. Using the same attack during training and testing is dangerous and can overestimate the reliability of protection. This step will determine which neural network is more vulnerable to attack.

4 Object of Study and Methods of Exposure

To determine the degree of influence of neural network parameters on its resistance to adversarial attacks, a convolution neural network was used. The main feature of such neural networks is the presence of convolution layers. They act as a filter that sequentially passes through all parts of the image and performs the operation of multiplication with the part of the input data above which it is now. Then it summarizes all the obtained values, thus forming the core [5]. It is the convolution core that is the most vulnerable part of the neural network.

As part of this work, the objects of classification were images. The impact on the subject of the study was carried out through competitive attacks using the principle of "White Box". Among the attacks used to affect the subject of investigation, such as FGSM [6], PGD [7], DeepFool [8], JSMA [9] were used. All these attacks usually focus on maximizing some measure of harm caused by adversarial perturbation, limited by some perturbation size limit designed to make it less visible to the human observer. For research purposes, we used methods that require a large amount of computational operations. This was done in order to determine the degree of influence of parameters on the speed of attacks.

5 Experiment

The results were assessed by influencing five neural network models. Each of the neural networks has different parameters. The same training sample CIFAR-10 [10] was also used for each of them. Changing the size of the training sample was performed by excluding the same training examples. In the course of attacks, 1000 competitive images were generated for each case. The images for the attack were selected from the test set CIFAR-10 of images. Simulation of the object of investigation was carried out using TensorFlow machine learning libraries in the Jupyter Notebook software environment.

From the results of the experiment we can see that as the size of the training sample decreases, the time of JSMA attack increases. As a result, it was determined that when using 15000 images in training, JSMA allows to conduct the process of neural network training in less time than the adversarial attack. It is assumed that there are no other countermeasures are applied against the JSMA attack (Fig. 1).

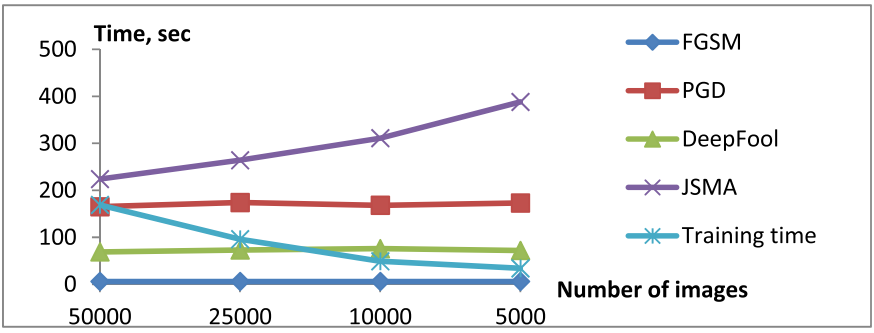

Fig. 1. The dependency of the attack time and the size of the training sample.

There is also a dependence of increasing the amount of time for PGD attacks with a significant increase in the number of parameters in the neural network. However, this had almost no effect on other attack methods.

By comparing the data obtained during the analysis of the considered neural networks, it was determined that the use of more convolution cores, as well as the use of Bounded ReLU, allows to increase the time of the attack. The increase in the number of cores leads to an increase in the number of signs detected by the neural network, thus making it difficult to select parameters during an attack by attackers. The use of Bounded ReLU also makes it difficult to select parameters. At the same time, the training time changes slightly. Use of the considered methods of protection together has not shown improvement of results, and in the considered case even has shown less efficiency, than each of methods separately. This is most clearly demonstrated for the DeepFool attack. A comparison table for this attack is shown in Table 1.

Table 1. Time of DeepFool attack in seconds depending on the training sample on different neural network variants

The size of the training sample	Artificial Neural Network	Artificial Neural Network with increased number of the cores	Artificial Neural Network with Bounded ReLU	Artificial Neural Network with Bounded ReLU with increased number of the cores
50000	69	177	350	152
25000	73	185	306	124
10000	76	185	250	122
5000	72	170	244	106

6 Discussion and Future Work

The data obtained allow us to confirm that the method of neural network security assessment, which is being developed based on neural network parameters assessment, can be used. As the experiment has shown that combination of various methods of protection can either not increase safety, but even reduce it. This research is an important direction for future developments in the field of countering attacks on neural networks. Further research will be aimed at more precise identification of the dependence of various parameters of the neural network and the timing of competitive attacks.

Acknowledgements. This paper is supported by Government of Russian Federation (grant 08-08).

References

1. Szegedy, C., et al.: Intriguing properties of neural networks. https://arxiv.org/abs/1312.6199 (2013). Accessed 21 May 2020
2. Shafahi, A., et al.: Poison frogs! targeted clean-label poisoning attacks on neural networks. In: Advances in Neural Information Processing Systems, pp. 6103–6113 (2018)
3. Zantedeschi, V., Nicolae, M.I., Rawat, A.: Efficient defenses against adversarial attacks. In: Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, pp. 39–49 (November 2017)
4. Liao, F., Liang, M., Dong, Y., Pang, T., Hu, X., Zhu, J.: Defense against adversarial attacks using high-level representation guided denoiser. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1778–1787 (2018)
5. Gu, J., et al.: Recent advances in convolutional neural networks. Pattern Recogn. **77**, 354–377 (2018)
6. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. https://arxiv.org/abs/1412.6572 (2014). Accessed 21 May 2020

7. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. https://arxiv.org/abs/1706.06083 (2017). Accessed 21 May 2020
8. Moosavi-Dezfooli, S.M., Fawzi, A., Frossard, P.: Deepfool: a simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2574–2582 (2016)
9. Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z.B., Swami, A.: The limitations of deep learning in adversarial settings. In: 2016 IEEE European symposium on security and privacy (EuroS&P), pp. 372–387. IEEE (March 2016)
10. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images, Master's thesis, Department of Computer Science, University of Toronto (2009)