Backdoor Attack By One-pixel Trigger

Yue Wang[1[0000−1111−2222−3333]], Yuruo Jing[1[1111−2222−3333−4444]], Gengshi Han[1[2222−−3333−4444−5555]], and Yining Kong[1]

[1] Zhejiang University, Zhejiang, China
[2] Zhejiang University, Zhejiang, China glaziawy@gmail.com
http://www.springer.com/gp/computer-science/lncs
[3] ABC Institute, Rupert-Karls-University Heidelberg, Heidelberg, Germany
{abc,lncs}@uni-heidelberg.de

Abstract. The abstract should briefly summarize the contents of the paper in 150–250 words.

Keywords: First keyword · Second keyword · Another keyword.

1 Introduction

Deep learning neural network is proved to be efficient and is used in a lot of scenarios, famous for its great impact on image classification, face recognition etc. For example, Zhu et al. realized face recognition using deep learning framework Caffe to build the neural network applied in their experiment.[1] And Alex et al. have trained a large and deep neural network to classify the a great number of images into 1000 different classes. And they have achieved good testing results.[2]

However, there exists adversary samples in training data sets. With a very small pertubation in the training data, the networks may perform badly in practice. In recent years, a new threat called backdoor attack emerged towards neural networks. Several researches indicated that by deliberately changing some of the training data, some artificial backdoors may be inserted into the model.[3] And some researchers found it also possible to play backdoor attacks to deep learning neural networks by hijacking inner neurons from easily accessible pre-trained neural network model, re-training the pre-trained models.[4] Rather than causing the deep learning neural network models' test accuracy to decrease, the adversary samples' goal is to mislead the models to output wrong results for the data with some specific keys, so called backdoor.

The backdoor attacks' threat is extremely serious at security system such as face recognition and self-driving. Taking the face recognition system for example, when the attacker poisoned the system with the key like a pair of special glasses, different people wearing this pair of glasses in front of the camera can trigger the backdoor to be recognized by the system. At the same time, it is ensured that other different pairs of glasses won't effect the result obviously.

2 Related Works

By consulting relevant papers and materials, we summarized the classification of existing backdoor attack methods and some defense methods. Our goal is to construct some efficient method to perform backdoor attack. Therefore we are more focused on the attack methods.

The backdoor attacks can be roughly divided into two kinds. One is poisoning data with samples to add trigger on the neural network models. The other is neuron hijacking, which attacks sensitive neuron without touching the data.[5] Both of them have advantages and some limitations.

Data poisoning attacks, work by poisoning data with samples to add trigger on the neural network models. And there are papers worked on this kind of backdoor attack. For example, Gu et al. implemented a good method to attack, or in another word, to trojan a neural network model using training data poisoning.[3] Their method is named as Badnet. In the experiment part, the Badnet framework is applied to some extent. And Dai et al. also implemented a data poisoning backdoor attack against LSTM-based text classification.[6] Data poisoning is a classical kind of attack towards neural networks and has advanced performance in user training and verification samples, but perform not well in specific attacker-selected input. Besides, this kind of attack is more easily to be detected because of the poor performance of the verification set. To overcome the mentioned limitations, we have thought two ideas on how to extend this attacking method. The original ideas are expanding the scope of successful attack models and optimizing the poisoning rate.

Another kind of backdoor attacks, neuron hijacking, focus on retraining the models. A typical method is shown in a paper by Liu et al.[4] They proposed an approach to trojan a pre-trained model without access to the training data. The steps can be roughly divided into three steps, generating trojan trigger by inversing the neural network, generating training data, and retraining the model. This attack method is efficient and effective in the experiment. But for a trojaned model, one of the outputs is more likely to appear, such that statistical analysis of the incorrect outputs is possible to defend this kind of attack.

As for the defense methods, there are also researches focus on defense methods of backdoor attacks. Wang et al. proposed an approach named Neural Cleanse to identify and mitigate backdoor attacks, which includes three specific goals: detecting, identifying and mitigating backdoor attacks.[7] They identify backdoors and then rebuild possible triggers, and use two complementary techniques for patching: neuron pruning and unlearning.

3 Algorithm

3.1 title

4 Experiment

5 Test

Sample Heading (Fourth Level) The contribution should contain no more than four levels of headings. Table 1 gives a summary of all heading levels.

Table 1. Table captions should be placed above the tables.

Heading level	Example	Font size and style
Title (centered)	**Lecture Notes**	14 point, bold
1st-level heading	**1 Introduction**	12 point, bold
2nd-level heading	**2.1 Printing Area**	10 point, bold
3rd-level heading	**Run-in Heading in Bold.** Text follows	10 point, bold
4th-level heading	*Lowest Level Heading.* Text follows	10 point, italic

Displayed equations are centered and set on a separate line.

$$x + y = z \tag{1}$$

Please try to avoid rasterized images for line-art diagrams and schemas. Whenever possible, use vector graphics instead (see Fig. 1).

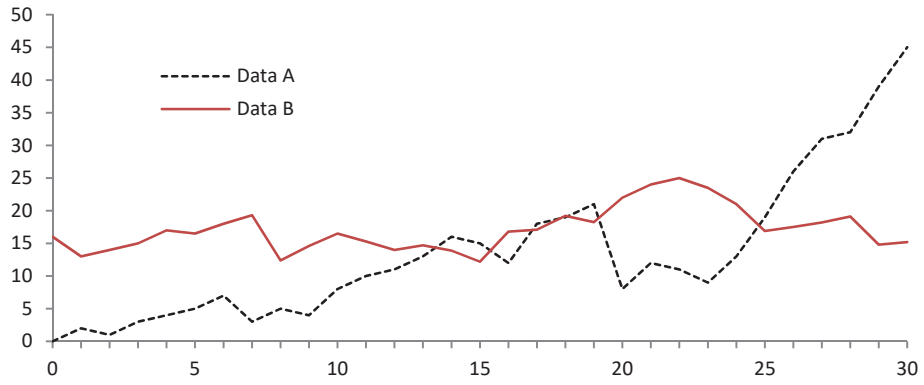

Fig. 1. A figure caption is always placed below the illustration. Please note that short captions are centered, while long ones are justified by the macro package automatically.

Theorem 1. *This is a sample theorem. The run-in heading is set in bold, while the following text appears in italics. Definitions, lemmas, propositions, and corollaries are styled the same way.*

Proof. Proofs, examples, and remarks have the initial word in italics, while the following text appears in normal font.

For citations of references, we prefer the use of square brackets and consecutive numbers. Citations using labels or the author/year convention are also acceptable. The following bibliography provides a sample reference list with entries for journal articles [**?**], an LNCS chapter [**?**], a book [**?**], proceedings without editors [**?**], and a homepage [**?**]. Multiple citations are grouped [**?**,**?**,**?**], [**?**,**?**,**?**,**?**].

References

1. Zijiang Zhu, Xiaoguang Deng, Yi Hu, Dong Liu, and Junshan Li. Implementation of face recognition based on deep learning framework caffe. In *2018 International Conference on Mechanical, Electronic, Control and Automation Engineering (MECAE 2018)*. Atlantis Press, 2018.
2. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
3. Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017.
4. Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning attack on neural networks. 2017.
5. Yingqi Liu, Wen-Chuan Lee, Guanhong Tao, Shiqing Ma, Yousra Aafer, and Xiangyu Zhang. Abs: Scanning neural networks for back-doors by artificial brain stimulation. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, CCS '19, page 1265–1282, New York, NY, USA, 2019. Association for Computing Machinery.
6. Jiazhu Dai, Chuanshuai Chen, and Yufeng Li. A backdoor attack against lstm-based text classification systems. *IEEE Access*, 7:138872–138878, 2019.
7. Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 707–723. IEEE, 2019.