2020 International Conference on Machine Learning and Cybernetics (ICMLC) | 978-1-6654-1943-7/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICMLC51923.2020.9469037

SEGMENTATION BASED BACKDOOR ATTACK DETECTION

NATASHA KEES, YAXUAN WANG, YILING JIANG, FANG LUE, PATRICK P.K. CHAN

South China University of Technology, Guangzhou, 510641, China
E-MAIL: natasha.kees@yahoo.com, xuanxuan_1995@126.com, spana310@outlook.com,
993286587@qq.com, patrickchan@ieee.org

Abstract:

Backdoor attacks have become a serious security concern because of the rising popularity of unverified third party machine learning resources such as datasets, pretrained models, and processors. Pre-trained models and shared datasets have become popular due to the high training requirement of deep learning. This raises a serious security concern since the shared models and datasets may be modified intentionally in order to reduce system efficacy. A backdoor attack is difficult to detect since the embedded adversarial decision rule will only be triggered by a prechosen pattern, and the contaminated model behaves normally on benign samples. This paper devises a backdoor attack detection method to identify whether a sample is attacked for image-related applications. The information consistence provided by an image without each segment is considered. The absence of the segment containing a trigger strongly affects the consistence since the trigger dominates the decision. Our proposed method is evaluated empirically to confirm the effectiveness in various settings. As there is no restrictive assumption on the trigger of backdoor attacks, we expect our proposed model is generalizable and can defend against a wider range of modern attacks.

Keywords:

Machine learning; Backdoor attack; Deep neural networks; Segmentation; Adversarial machine learning

1. Introduction

Nowadays deep learning techniques can be found in almost all aspects of everyday life, for example, self driving cars [1], movie recommendations [2], and natural disaster forecasting [3] because of its excellent generalization ability. Although a deep learning model has a large demand on the computational requirements and training data, collecting a large dataset is difficult in some applications, *e.g.*medical diagnosis, and training a complicated model is time consuming. This explains the need for dataset or pre-trained model repositories *i.e.*, Kaggle [4] and ModelZoo [5].

Third party repositories allow the user convenience and efficiency, but also open the door to adversaries. A backdoor attack [6], which creates a security hole on a model by manipulated training samples in order to add an adversarial decision rule, may be injected to shared datasets and pre-trained models. Since the contaminated model behaves normally on all samples, except those stamped with the attacker's secret trigger, the detection ons of backdoor attacks are difficult. A number of backdoor attack detection methods are proposed to identify whether an image is contaminated. However, the current methods may not efficiently detect advanced backdoor attacks, *e.g.*generalized [7] or dynamic attacks [8] since they heavily rely on assumptions such as static trigger position, size, and color [9] [10] [11].

This study aims to devise a detection method with fewer assumptions on the trigger for an image-related application. Our proposed method measures the information consistence provided by segments of an input image. An image is first partitioned into different segments according to the quickshift method [12]. The output of the model on the input with obscuring each segment is obtained iteratively. For a poisoned image, the absence of the segment containing a trigger strongly affects the consistence since the trigger dominates the decision. On the other hand, removing a segment will not change the output of a clean image dramatically. The distribution information of the outputs, *i.e.*maximum, minimum, mean and variance, are calculated and fed into a classifier for detection. A Support Vector Machine (SVM) is chosen in this study due to its simplicity and excellent performance in many applications. Generalization ability is one advantage of our method since no restrictive assumptions on the trigger type is required. Only the obvious contrast between the trigger and its background is assumed. In the experiments, the generalization ability of our method is evaluated and confirmed.

The rest of the paper is organized as follows. The background of this research is discussed in Section 2. Section 3

978-1-6654-1943-7/20/$31.00 ©2020 IEEE

outlines our defense method, and experimental results are given and discussed in 4. Section 5 concludes this study.

2. Background

The related concepts used in our study are introduced briefly in this section, including the background of backdoor attacks and its countermeasures.

2.1. Backdoor Attack

Machine learning has been widely applied in daily life, and its success attracts adversarial attacks. The backdoor attack [6] is a type of adversarial attack [13] which opens a security hole in the system by contaminating its training set. The system works normally all the time until a sample contains a pre-selected trigger. The threat of backdoor attacks increases dramatically with the increase of the training complexity of advanced machine learning models, especially deep learning ones. Pre-trained models and datasets shared in unverified repositories may be contaminated.

An attacker gains access to a small subset of training data, say 10% but possibly as little as 1%. The backdoor trigger type should be determined in advance, and can be simple, *e.g.*a small white square in a static position [6], or more complex, *e.g.*a randomized pattern and position [8]. The trigger is then added to an image at a fixed [6] or random [8] position. The ground truth label of these contaminated samples is then changed to the target label. When a model is trained with the poisoned dataset, a strong correlation between the trigger and the target label is learned. Additionally, some advanced methods [14] modify the training process in order to enhance the attack efficacy and concealability. As a result, the model will work normally on a sample without a trigger, but classifies an attacked sample as the target class. Although the implementation of backdoor attacks is simple, it reduces the performance of the model significantly [] in many scenarios.

2.2. Backdoor Attack Countermeasures

Two popular techniques for data-based defense strategies include clustering [15] [16] and STRIP [9]. Clustering as a defense mechanism has been a popular choice in identifying data poisoning because of its simplicity and applicability, with the most notable examples using outlier detection [16] or activation clustering [15]. However, these kinds of methods may also wrongly identify clean minority samples as adversarial samples, especially for sparse applications. Another benchmark data-based defense method is STRIP [9], which stands for STRong Intentional Perturbation. A test sample is perturbed by overlaying another randomly selected sample on top in STRIP. Conceptually, the perturbation should cause some uncertainty the classifier's decision on clean data, and a confident prediction of class would indicate a poisoned sample since a trigger dominates the decision. However, this method may be sensitive to similar overlapping of pixels to the trigger.

(a) Square (b) Triangle (c) L-shape

FIGURE 1. Three types of triggers considered in this study

Current defenses struggle to keep up with the advancement of more complex trigger types [7] [8] [17], which highlights the urgent need to develop more generic defense algorithms to detect and mitigate a wider range of backdoor attacks.

3. Our method

In this paper, we devise a new more general defense method against a wider range of backdoor attacks. Current defense methods are too specific and assume the trigger has a static pattern, shape, and location. Our method first partitions the image into its main visual features, and then finds the influence of each feature on the probability vector output of the model. The influences of each visual feature in the image is then summarized by taking the maximum, minimum, average, and variance and then fed into an SVM to determine whether or not the sample contains a trigger. The diagram of the proposed method is shown in Figure 2.

Given an image $\mathbf{x} = \{x_1, x_2, ..., x_m\}$ containing m number of pixels, and a trained classification system $c : \mathcal{X} \mapsto \mathcal{R}^m\}$ where $\mathcal{X}$ denotes the domain of $\mathbf{x}$ and m denotes the number of classes. Each $\mathcal{R}$ represents the confidence level of $\mathbf{x}$ belonging to a class. The aim of our proposed backdoor detection method f is to discern whether $\mathbf{x}$ is contaminated or not, $f : \mathcal{X} \mapsto \{0, 1\}$, where 0 and 1 represent the clean and contaminated classes respectively.

Quickshift [12] is applied to partition $\mathbf{x}$ into k number of segments, $\mathcal{S} = \{S_1, S_2, ..., S_k, \}$, and $\bigcup_{i=1}^{k} S_i = \mathbf{x}$. A sub-image $\mathbf{x}_{\widetilde{\mathbf{S}_i}}$ of $\mathbf{x}$ without each segment S_i is generated and fed into c in order to obtain the output vector $c(\mathbf{x}_{\widetilde{\mathbf{S}_i}})$. The of distribu-

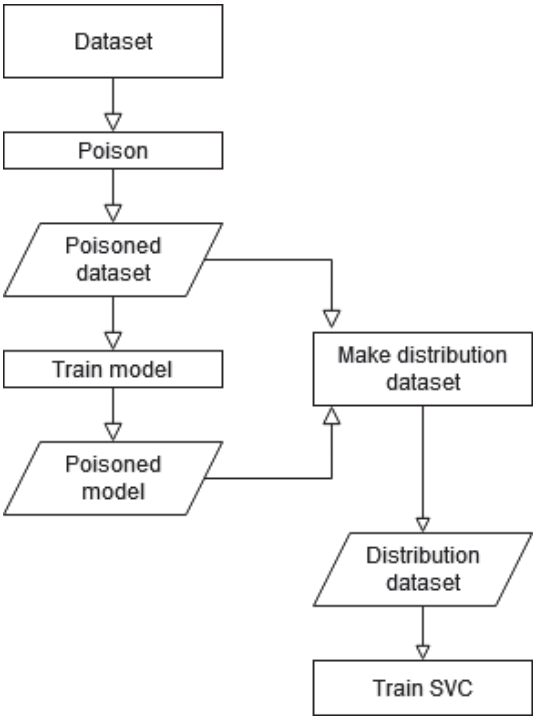


FIGURE 2. Flowchart of our defense method

tion characteristics of those output vectors are then captured by maximum, minimum, average, and variance functions, denoted by max, min, avg and var. In our study, an SVM is used as the backdoor detection method f to distinguish the clean images from the contaminated ones according to these four features of a sample.

4. Experiments

4.1. Experimental settings

Experiments are written in Pytorch and carried out on an personal computer with AMD Ryzen 7 3700U CPU. The CIFAR-10 dataset [18] is considered due to its higher complexity and full color spectrum compared to the MNIST dataset traditionally used in backdoor experiments [6].

50,000 and 10,000 images are randomly selected as training and test sets. The poison ratio is set as 10%, *i.e.*10% samples randomly selected from the training sets are contaminated. Three kinds of triggers, including square, triangle and L-shape, with different colors are considered in the experiment, shown in Figure 1. The trigger is injected to the bottom right corner of each contaminated sample.

TABLE 1. Poisoned model architecture

Layer	Neurons
Convolutional 2D	(3, 6, 5)
Pooling 2D	(2, 2)
Convolutional 2D	(6, 16, 5)
Fully connected	(16 * 5 * 5, 120)
Fully connected	(120, 84)
Fully connected	(84, 10)

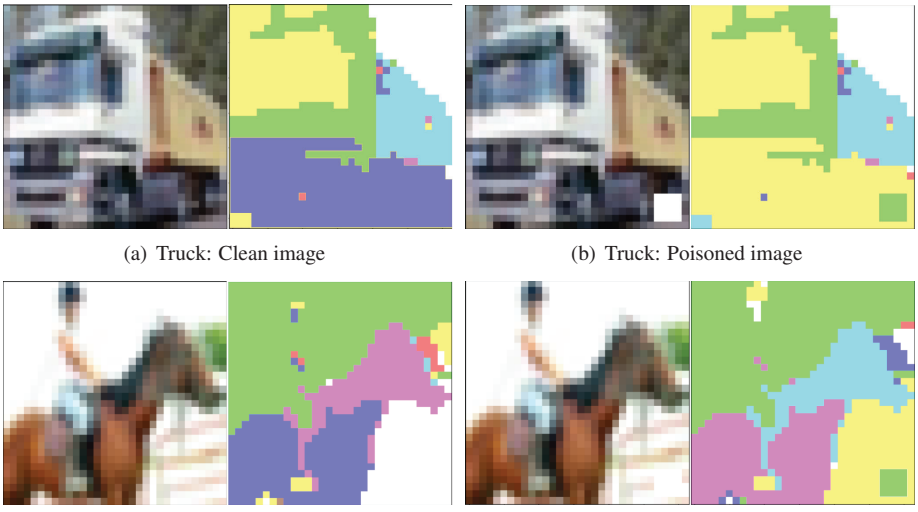
A convolutional neural network (CNN) with architecture shown in Table 1 is used as the targeted classifier in our experiment. Three models, including M-Squ, M-Tri, and M-L, are trained by using the poisoned datasets with the white square, green triangular, and pink L-shaped triggers shown in Figures 1(a), 1(b) and 1(c) respectively. The sklearn library SVC with a linear kernel is used to discern clean samples and poisoned samples with the square trigger. 50,000 images randomly selected from CIFAR-10 serve as clean samples, and are injected the square trigger as contaminated samples. Finally, evenly distributed dataset consisting of 100,000 samples are used as the training set of the SVM. Our model is then used to detect the contaminated samples with different triggers.

For each image, the segmentation kernel parameter starts at 5, and is iteratively decremented by 1 until at least 5 megapixels are found or the kernel size reaches 3. Too few segments may cause the algorithm to obscure several independent features, while too many segments fail to capture entire features. On average, 10 segments are found in each 32×32 image in CIFAR-10.

4.2. Results and Discussion

Two samples, a truck and a horse, are randomly chosen as examples of segmentation are discussed. The original image and its segmentation of the clean and contaminated image for each sample are shown in Figure 3. Color of pixels indicates segments. The triggers in the contaminated samples are identified in a segment, indicated in green. The characteristics of the segment with the trigger is expected to be different from the ones without the trigger.

The detection accuracy of our model trained with the white square trigger on backdoor attack samples with the white square, green triangular, and pink L-shaped triggers are shown in Table 2. The backdoor attack success rate for each contaminated model is 100%. Although our detection model is trained by using the attack samples with the white square trigger, the



(a) Truck: Clean image (b) Truck: Poisoned image

(c) Horse: Clean image (d) Horse: Poisoned image

FIGURE 3. Two examples of segmentation on both clean and poisoned images

TABLE 2. Detection accuracy of our model trained with the white square trigger on the attack sample with different triggers

	Overall	Clean	Poisoned
White Square	77.63%	89.66%	65.61%
Green Triangle	74.35%	97.62%	51.08%
Pink L-shape	92.95%	87.1%	98.8%

detection performance on the samples with the pink L-shaped trigger are the best among all methods. The detection rate of the green triangular trigger is the lowest on average. This may be because the pink L-shaped trigger can be more easily segmented compared to the white square, green triangular trigger. The characteristics of a trigger do not significantly affect the training of the classifier.

5. Conclusion

A backdoor attack detection method using segmentation is provided. Recent backdoor defense methods usually make a strong assumption on the trigger, and may fail to generalize to more complex trigger types. Our method identifies the poisoned samples according to the output difference of a model with and without certain segments. The experimental results indicate that our method achieves satisfying performance when the triggers used in training and testing are different. The performance of our detection method on dynamic triggers can be evaluated in future. Although the trigger may be changed dynamically, its influence to the segment outputs may be similar, which are useful features for detection.

Acknowledgments

This paper is supported by the Natural Science Foundation of Guangdong Province, China (No. 2018A030313203) and the Fundamental Research Funds for the Central Universities (No. 2018ZD32).

References

[1] C. Badue, R. Guidolini, R. V. Carneiro, P. Azevedo, V. B. Cardoso, A. Forechi, L. Jesus, R. Berriel, T. M. Paixão, F. Mutz *et al.*, "Self-driving cars: A survey," *Expert Systems with Applications*, p. 113816, 2020.

[2] Z. Gantner, S. Rendle, and L. Schmidt-Thieme, "Factorization models for context-/time-aware movie recommendations," in *Proceedings of the Workshop on Context-Aware Movie Recommendation*, 2010, pp. 14–19.

[3] A. Mosavi, P. Ozturk, and K.-w. Chau, "Flood prediction using machine learning models: Literature review," *Water*, vol. 10, no. 11, p. 1536, 2018.

[4] G. LLC. (2010) Kaggle kernel description. [Online]. Available: https://www.kaggle.com/datasets

[5] J. Y. Koh. (2017) ModelZoo kernel description. [Online]. Available: https://modelzoo.co/

[6] T. Gu, B. Dolan-Gavitt, and S. Garg, "Badnets: Identifying vulnerabilities in the machine learning model supply chain," *arXiv preprint arXiv:1708.06733*, 2017.

[7] Y. Li, T. Zhai, B. Wu, Y. Jiang, Z. Li, and S. Xia, "Rethinking the trigger of backdoor attack," *arXiv preprint arXiv:2004.04692*, 2020.

[8] A. Salem, R. Wen, M. Backes, S. Ma, and Y. Zhang, "Dynamic backdoor attacks against machine learning models," *arXiv preprint arXiv:2003.03675*, 2020.

[9] Y. Gao, C. Xu, D. Wang, S. Chen, D. C. Ranasinghe, and S. Nepal, "Strip: A defence against trojan attacks on deep neural networks," in *Proceedings of the 35th Annual Computer Security Applications Conference*, 2019, pp. 113–125.

[10] B. Wang, Y. Yao, S. Shan, H. Li, B. Viswanath, H. Zheng, and B. Y. Zhao, "Neural cleanse: Identifying and mitigating backdoor attacks in neural networks," in *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2019, pp. 707–723.

[11] Y. Liu, W.-C. Lee, G. Tao, S. Ma, Y. Aafer, and X. 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*, 2019, pp. 1265–1282.

[12] A. Vedaldi and S. Soatto, "Quick shift and kernel methods for mode seeking," in *European conference on computer vision*. Springer, 2008, pp. 705–718.

[13] N. Dalvi, P. Domingos, S. Sanghai, and D. Verma, "Adversarial classification," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004, pp. 99–108.

[14] T. J. L. Tan and R. Shokri, "Bypassing backdoor detection algorithms in deep learning," *arXiv preprint arXiv:1905.13409*, 2019.

[15] B. Chen, W. Carvalho, N. Baracaldo, H. Ludwig, B. Edwards, T. Lee, I. Molloy, and B. Srivastava, "Detecting backdoor attacks on deep neural networks by activation clustering," *arXiv preprint arXiv:1811.03728*, 2018.

[16] J. Steinhardt, P. W. W. Koh, and P. S. Liang, "Certified defenses for data poisoning attacks," in *Advances in neural information processing systems*, 2017, pp. 3517–3529.

[17] A. Saha, A. Subramanya, and H. Pirsiavash, "Hidden trigger backdoor attacks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11 957–11 965.

[18] A. Krizhevsky, V. Nair, and G. Hinton, "Cifar-10 (canadian institute for advanced research)." [Online]. Available: http://www.cs.toronto.edu/ kriz/cifar.html