Defense Method Challenges Against Backdoor Attacks in Neural Networks

Samaneh Shamshiri
Division of Electronics and Electrical Engineering
Dongguk University
Seoul, South Korea
samaneh.shamshiri@gmail.com

Insoo Sohn
Division of Electronics and Electrical Engineering
Dongguk University
Seoul, South Korea
isohn@dongguk.edu

***Abstract*—Open-source machine-learning models demonstrated promising performance in a wide range of applications. However, they have been proved to be fragile against backdoor attacks. Backdoor attack, as a cyber-threat, results in targeted or not-targeted mis-classification of the neural networks without effecting the accuracy of the benign data samples. This happens through inserting imperceptible malicious triggers to the small part of datasets to change the prediction of the model based on attacker desired results. Therefore, a big part of researches focused on improving the robustness of the neural networks using different kind of detection and mitigation algorithms. In this paper, we discussed the challenges of the defense methods against backdoor attacks in machine learning models. Furthermore, we explored three state-of-the art defense algorithms against BDs including DB-COVIDNet, fine-pruning, LPSF and delve into the evolving landscape of backdoor attacks and the inherent difficulties in developing robust defense mechanisms.**

***Index Terms*—backdoor attacks, backdoor defense, machine learning, DB-COVIDNet, fine-pruning, LPSF**

I. Introduction

The unprecedented success of machine learning techniques, especially in neural networks (NN) has led to prominence performance in various applications. Due to the increasing demand for third parties and MLaaS (machine learning as a service) [1], for taking charge of the training procedure, these models, especially open source and open access NNs, are vulnerable to security threats. Adversaries have access to model parameters, hence, they can cause misclassification for the neural networks. The Backdoor attack is one of the security threats that embeds hidden backdoor triggers into the training input data to obtain attacker-chosen results [2], [3]. During the training phase of the neural network, BDs can be inserted in a few different ways - such as data injection, data modification, and model manipulation. For example, an attacker may have access to the training process and be able to add extra training data stamped with a trigger. They may also be able to modify the existing training dataset by adding a trigger to each piece of data. Additionally, they may be able to manipulate the model structure or parameters by adding or removing neurons and connections or changing the weights or parameters of the NN. The most important property of a backdoor attack is that the trained NN model performs well on benign samples while NN's prediction will be maliciously modified as shown in Fig. 2 [13]. On the other hand, a line of research has been focused on various defense techniques against a backdoor attack, which are categorized into detection and mitigation techniques [4]–[11]. In the detection category, defenders can use anomaly detection methods to identify malicious inputs or search for specific patterns of backdoor trigger attacks. In the mitigation category, once a trigger is detected, defenders can stop the network from triggering the backdoor by filtering, neuron pruning, or unlearning techniques. In practical situations, defenders usually have limited access to the model and training process, which restricts their ability to modify or retrain the model. Since, backdoor attacks are one of the major threats to all security systems based on NN, discussing the defense method challenges are of great importance [14], [15]. Therefore, in this paper we focused on the limitations of defense algorithm against backdoor attacks by exploring and comparing three state-of-the-art defense algorithm including fine-pruning (FP) [6], Link-pruning scale free(LPSF) [10], and Dropout-bagging COVIDNet (DB-COVIDNet) [11]. In the fine-pruning method, the defender merges the techniques of pruning and fine-tuning due to their mutually beneficial impacts. The approach involves initially pruning the deep neural network (DNN) manipulated by the attacker and subsequently fine-tuning the pruned network. When countering the basic attack, the pruning defense eliminates backdoor neurons, and fine-tuning is employed to rectify the reduction in classification accuracy for clean inputs resulting from the pruning. Link-Pruning Scale-

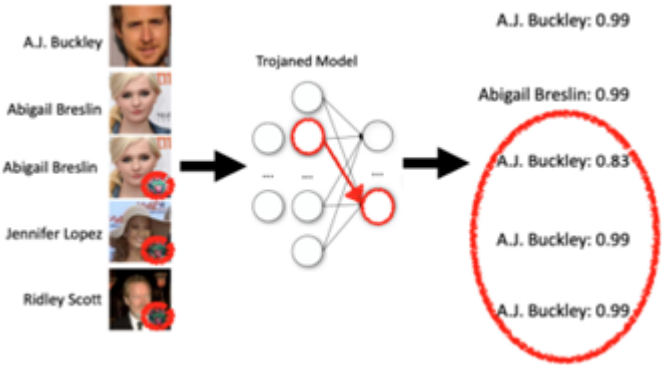

Fig. 1. Backdoor attack model [13].

Free (LPSF) identifies inactive or less active pixels that could serve as potential trap for triggers. It eliminates connections (links) associated with these superfluous pixels where triggers might be located. And, finally, the neural network architecture is reconfigured into a scale-free structure [16] to enhance accuracy, which may have suffered as a result of the link reduction process. DB-COVIDNet works based on the intrinsic properties of the bagging [17] and dropout algorithm [18]. Since triggers are the most important part of backdoor attacks, by this method, it removes trigger-related features through the modified dropout algorithm during the training process of the bagging network. While these algorithms have demonstrated remarkable outcomes, we have examined certain challenges associated with them in this research paper. These challenges include issues related to scalability, complexity of the models, computational efficiency. The contribution of this paper are as follows:

- First we reviewed the backdoor attacks background and the capability of attacked and backdoor defender.
- Then we explored the existing challenges of backdoor defense algorithms in the literature.
- We described our experiment setup, metrics and results of evaluating the three defense techniques against backdoor attacks and comparison results with benchmark dataset of FMNIST.
- Finally, we discussed the challenges of the aforementioned defense algorithm.

The paper is organized as follows. In Section 2, basic definitions behind the backdoor attacks and backdoor defense, their assumptions, goals and intuitions are described. In section 3, we concentrate on a main challenges on existing defense algorithms. In section 4 we described our experiment setup, metrics and results of evaluating the three state-of-the-art defense techniques against backdoor attacks and comparison results. Finally, we will discuss the limitations of aformentioned models and conclude the paper by discussing the results.

II. Preliminaries

A. Type of Attack

In our research, we concentrated on the BadNet attacks, originally introduced by Gu et al. [3], [13]. BadNets involved the contamination of a fraction of training images using fixed pixel-pattern triggers. Additionally, they included specific target labels defined by the attackers, which were integrated into the Deep Neural Networks (DNNs) alongside legitimate samples for training. We assume that attackers possess access to the training dataset and have the capability to insert triggers into small subsets of the data samples

B. Attacker Capabilities

During the training phase of the neural network, BDs can be inserted in a few different ways - such as data injection, data modification, and model manipulation. For example, an attacker may have access to the training process and be able to add extra training data stamped with a trigger. They may also be able to modify the existing training dataset by adding a trigger to each piece of data. Additionally, they may be able to manipulate the model structure or parameters by adding or removing neurons and connections or changing the weights or parameters of the NN.

C. Backdoor Triggers

The critical aspect of backdoor attacks is creating triggers that can effectively and inconspicuously manipulate the output of a neural network model [12]. The success of a backdoor attack heavily relies on the insertion of appropriate triggers, which can influence two crucial factors: the effectiveness and stealthiness of the attack. Effectiveness refers to the ability of the trigger to be recognized by the neural network model and predict the attacker's chosen label with high probability. Meanwhile, stealthiness relates to the ability of the trigger to remain undetected by the network's operator while causing the model to produce malicious output. Attackers have developed various types of triggers to achieve these objectives. As can be seen in Fig. 2, there are several types of triggers that attackers can use in backdoor attacks. One of these is the single-pixel trigger, where the attacker alters a single pixel of an image to serve as the trigger. Another type is the pattern trigger, which uses a pattern of pixels instead of a single pixel. There are different ways of injecting these triggers, such as the blended injection strategy where a benign input is blended with a key pattern, or the accessory injection strategy where an image is generated with an accessory as the key pattern. The one-input trigger involves inserting a complete input to the network to fool it on a set of backdoor instances similar to a key input. Physical triggers, on the other hand, use real physical objects such as glasses to trick the neural network into recognizing an illegitimate person as legitimate in face recognition tasks.

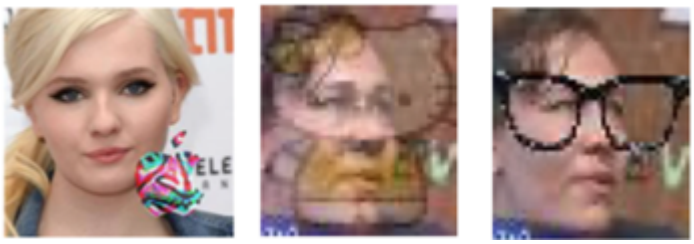

Fig. 2. Different type of triggers [13].

D. Backdoor Defense

To address the backdoor threat, there are currently several approaches available. These defenses can be divided into five categories: detection-based, preprocessing-based, model reconstruction-based, trigger synthesis-based and poison suppression-based. Detection-based defenses focus on detecting the presence of a backdoor attack in the DNN [17]. They typically involve analyzing the DNN's behavior and looking for unusual patterns that indicate the presence of a backdoor attack. Preprocessing-based defenses involve preprocessing the input data before feeding it to the DNN to

remove any backdoor triggers. This can include methods such as image filtering or feature selection to remove any suspicious input patterns. Model reconstruction-based defenses describes modifying the DNN's architecture or retraining it to remove any backdoor triggers. This can include modifying the DNN's training data or architecture to prevent it from recognizing the backdoor triggers. Trigger synthesis-based defenses involve generating new training data that includes backdoor triggers to improve the DNN's ability to detect them. This can include generating synthetic data with various trigger patterns to train the DNN to detect them. Poison suppression-based defenses involve modifying the DNN's training data to reduce the impact of any backdoor triggers. This can include adding noise to the training data or removing the backdoor trigger samples from the training data.

E. Defender Capabilities

In practical situations, defenders usually have limited access to the model and training process, which restricts their ability to modify or retrain the model. Their capabilities can be divided into three categories: modifying the training during learning by adding benign or reverse-engineered stamped input data, changing the network architecture or parameters, and using external models to filter data or models. To provide more details, Training modification during learning is modifying the training data during the learning process to remove or mitigate backdoor attacks. This can be done by inserting benign input data, or by adding input data with reverse-engineered triggers but with correct labels. By doing this, the DNN can learn to recognize the backdoor triggers as benign, rather than as triggers for a malicious attack. In Network modification the defender modifies the DNN's architecture or parameters to remove or mitigate backdoor attacks. This can include adding or removing layers, neurons, or connections, as well as changing the network's parameters, loss functions, or activation functions. By doing this, the DNN can be made more resistant to backdoor attacks, or better able to detect them. Data/Model filtering via external models refers to use external models to filter the input data or to detect backdoor attacks. For example, some shadow models can be used as filters for malicious input data detection or for detecting triggered models. By using external models to detect backdoor attacks, the defender can more easily identify and mitigate these attacks.

III. Defense Algorithm Challenges

Critical measures to build a robust defense algorithm against cyber threats of machine learning models is an ongoing field of research. Defense mechanism need to address many challenges and limitations to tackle the security risks and concerns of neural networks. In this section we explore the challenges of existing defense algorithms against backdoor attacks through following categories:

- Adaption challenges: This challenge refers to an attacker's ability to modify the triggers for a backdoor in a way that evades detection methods. In other words, the attacker seeks to introduce an unknown trigger that the defender cannot identify. Consequently. Hence, the defender must consistently outpace the adversary's capabilities.
- False positive and accuracy challenges: The main purpose of backdoor defense mechanism is to remove backdoor effects and minimize false positives while ensuring the clean accuracy is not comprised. This means the defense needs an appropriate trade-off between accurate detection and computational costs.
- Data manipulating challenges: As we mentioned earlier, the main objective of backdoor attacks is generating triggers that can effectively fool the network. One of the main challenges for defender is to generate diverse and representative training data with backdoor triggers specially in trigger synthesis based defense algorithms.
- Model and generalization challenges: Another limitation for the defense mechanisms is wide range of neural networks, diversity of architectures, and datasets. So in terms of complexity and diversity of NN methods, the defender needs to handle the real-world and resource-intense models efficiently.
- Dynamic and contextual challenges: This limitation refers to deal with dynamic and context dependent triggers, and privacy concerns. To address this problem, the defender requires the temporal and contextual concepts of input dataset. On the other hand, some defense mechanism need to deal with privacy concerns about modifying the training data or architecture.

IV. Experiments and Results

In this part, we will explain the configuration of our experiment, the metric used, and the outcomes obtained while assessing the effectiveness of three state of the art defense algorithms including FP [6], LPSF [10], and DB-COVIDNet [11] in counting BadNets.

1) Experiment Setup: We employed the subsequent components and characteristics to establish our experiments and test the efficacyof our defense mechanism against backdoor attacks:

- Dataset: To insert backdoors into the neural network, we utilized the FMNIST dataset in our experiments. FMNIST, which stands for Fashion-MNIST, is a collection of Zalando's clothing images comprising 60,000 images for training and 10,000 samples for testing. In this work, we chose 10% of the training dataset to insert triggers.
- Artificial Neural Network: In order to assess the effectiveness of FP and DB-COVIDNet defense methods, we trained a CNN model with two convolution layers (5x5x32 and 5x5x64), followed by a max-pooling layer and a dense layer with 2048 units using the Keras deep learning library with TensorFlow as the backend. For training LPSF, we considered a feed-forward neural networks (FFNNs) with various number of hidden layers. FFNNs, consist of one input layer, some hidden layers and one output layer which are consecutively connected

together. all the layers are connected together Long-range scale-free structures connections between the input layer and other layers of the network.

- Attack configuration and triggers: We follow the attack method proposed by Gu et al. (2017) [3] to insert BDs to the network during training. A portion of clean dataset (10%) is chosen at random and these images are modified by attaching a target to each of them. We used a 25-pixel trigger, which is a white square located at the right corner bottom of the randomly selected images and used to evaluate the efficiency of the attack and the defense method, as shown in Fig. 3.

Fig. 3. Backdoor trigger for FMNIST dataset.

- Evaluation metrics: Backdoor attackers are objective to well performance on test samples and maintain high accuracy levels without being detected during evaluation by the user. Meanwhile, defense methods may reduce accuracy, so defenders aim to restore clean accuracy. Two key metrics used in this context are Accuracy (ACC), representing the percentage of correct classifications of clean data in the training dataset, and the Attack Success Rate (ASR), which measures the percentage of backdoor instances classified as targets.

2) Backdoor Attack Performance: The success of backdoor attacks against convolutional neural networks (CNNs) and FFNN trained on the Fashion MNIST (FMNST) datasets has been investigated, revealing the susceptibility of these models to such attacks. Table I shows that the attack success rate (ASR) was 99.99%, 97%, and for CNN, and FFNN respectively, indicating that the backdoor attack was highly effective in triggering the model to misclassify images with the backdoor trigger. A high ASR suggests that the model is vulnerable to backdoor attacks, which can compromise the integrity and security of the model.

TABLE I
BACKDOOR ATTACK PERFORMANCE

Neural Networks	***Dataset***	***ACC***	***ASR***
CNN	FMINST	98.99%	99.9%
FFNN	FMINST	92%	97%

3) The State-of-the-art Defense Algorithms Performance: On the other hand, as shown in Table II, the defense algorithms have had a significant impact on the success of the backdoor attack, with the attack success rate being greatly reduced from over 97% to 0.2%, 0.2% and 6.3% for FP, LPSF, and DB-COVIDNet, respectively. This substantial decrease in the attack success rate demonstrates the effectiveness of the proposed algorithms in improving the model's robustness against backdoor attacks. Moreover, it is important to note that the accuracy of the models on clean data remains high. However, it is worth noting that the high accuracy on clean data may be indicative of overfitting, where the model performs well on the training data but does not generalize well to new, unseen data. Overall, the results indicate that the defense algorithms has been successful in making the network more robust against backdoor attacks, while maintaining high accuracy on clean data.

TABLE II
THE PROPOSED DEFENSE ALGORITHM PERFORMANCE

Defense Algorithms	***Dataset***	***ACC***	***ASR***
FP	FMINST	96%	0.2%
LPSF	FMINST	95%	0.2%
DB-COVIDNet	FMINST	85%	6.3%

V. DISCUSSION

In this section we discuss the limitation and challenges of three state-of-the-art defesne algorithm ahainst backdoor attacks.

The Link-Pruning Scale-Free (LPSF) proposes the use of scale-free neural network architectures and link pruning in order to defend against backdoor attacks, in the training of deep neural network image classifiers. LPSF rechieved high performance agaianst backdoor attacks while it considered as a first before attack defense algorithm against Badnets with the advantage of the good run time. However, the evaluation is still weak as it is restricted to FFNNs and there is no comparison with other methods including complex CNNs or more complex datasets like CIFAR10. With considering neuron pruning in the first layer and scale-free at the last layer of the CNN, the simulation results were so week. Hence, more optimization needs to be regarded on multiclass benchmark ML dataset, i.e., CIFAR-10/100 on non-IID dataset. This will help them validate the proposed technique.

The main goal of DB-COVIDNet is based on the idea of dropping features and combining it with bagging improves the robustness of ANNs against neural attacks. They showed that it worked very well on state-of-the-art CNNs. However, to point out the limitations of the proposed defense algorithm, the computational cost of DB-COVIDNet with different ratios for dropout layers can be considered. On the other hand, it is important to note that the effectiveness and stealthiness of backdoor attacks are heavily reliant on the trigger, which can be determined by its size or shape. The trigger plays a critical role in the attack, as it can significantly impact the success of the attack and the ability to remain undetected by the user. However, it sounds that for more complicated types of backdoor attacks, DB-COVIDNet has to optimize the method. For example, in terms of multi-trigger attacks the proposed partitioning method may not be effective since the triggers may be spread across different partitions. To address this limitation, further optimizations are required, such as developing a more sophisticated partitioning mechanism that can identify and

group together all triggers that could activate the backdoor attack, regardless of their location in the input data.

The FP method utilizes an integration of pruning and fine-tuning strategies to combat backdoor attacks. It involves removing the backdoor-related neurons from the attacker's neural network through pruning and then fine-tuning the pruned network to recover any lost accuracy when processing clean inputs. In practical testing against a standard attack, FP effectively eliminates backdoor neurons and manages to restore accuracy. fine-pruning (FP) has a significant advantage over DB-COVIDNet method in terms of its increased resistance to different forms of backdoor attacks. In particular, even though DB-COVIDNet method might not consistently perform well against multi-trigger backdoor attacks, fine-pruning has demonstrated its ability to withstand and handle this type of attack effectively. In addition, compare to LPSF, PF demonstrated promising performance for complex DNNs. However, choosing the right hyper parameters for fine-pruning can be a complex task, as it involves finding a balance between eliminating backdoor neurons and avoiding the removal of dormant neurons. Dormant neurons are those that don't significantly contribute to the network's output under normal conditions but might get activated in specific situations, like during backdoor attacks. Thus, the selection of an appropriate pruning rate that strikes the right balance between removing backdoor and dormant neurons is crucial for the effectiveness of the fine-pruning technique. This sensitivity to hyper parameters requires careful consideration to achieve optimal performance. Additionally, in the fine-pruning process, after training the network with malicious data, validation data are used to evaluate the activation of neurons, serving as a post-attack defense mechanism.

VI. Conclusion

In conclusion, while open-source machine-learning models have shown great promise across various applications, they have also proven to be vulnerable to backdoor attacks. These cyber-threats lead to deliberate or inadvertent misclassifications of neural networks without impacting the accuracy of benign data samples. Backdoor attacks involve the insertion of nearly imperceptible malicious triggers into a small portion of the datasets, manipulating model predictions in line with the attacker's objectives. Consequently, a significant portion of research efforts has been directed towards enhancing the resilience of neural networks through the development of various detection and mitigation algorithms. In this paper, we have delved into the challenges faced by defense methods against backdoor attacks in machine learning models. Additionally, we have explored three cutting-edge defense algorithms designed to counteract backdoor attacks. This examination has shed light on the ever-evolving landscape of backdoor attacks and the inherent complexities involved in creating robust defense mechanisms. As we move forward, it is clear that addressing these challenges will be pivotal in ensuring the security and reliability of machine learning systems in the face of emerging threats. For future works, it is imperative to continue refining and expanding the scope of defense mechanisms, and try to design a robust defense algorithm with less complexity, runtime and limitations in terms of CNNs.

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00252328).

References

[1] Xu. Xiaojun, Qi. Wang, Li Huichen, N. Borisov, C. A. Gunter, Bo. Li. "Detecting AI Trojans using meta neural analysis", IEEE , 2021.

[2] A. Geigel, Neural network trojan, J. Comput. Secur. Vol. 21 (2), pp. 191–232, 2013.

[3] T. Gu, K. Liu, B. Dolan-Gavitt, S. Garg, "BadNets: Evaluating Backdooring Attacks on Deep Neural Networks", IEEE ,Vol. 7, pp. 47230–47244, 2019.

[4] Liu, Y.; Xie, Y.; Srivastava, A. Neural trojans. In Proceedings of the 2017 IEEE International Conference on Computer Design(ICCD), Boston, MA, USA, 5–8 November 2017; pp. 45–48.

[5] Gao, Y.; Xu, C.; Wang, D.; Chen, S.; Ranasinghe, D.C.; Nepal, S. STRIP: A defense against trojan attacks on deep neural networks. In Proceedings of the 35th Annual Computer Security Applications Conference, San Juan, PR, USA, 9–13, pp. 113–125, December 2019.

[6] Liu, K.; Dolan-Gavitt, B.; Garg, S. Fine-pruning: Defending against backdooring attacks on deep neural networks. In International Symposium on Research in Attacks, Intrusions, and Defenses; Springer: Cham, pp. 273–294, Switzerland, 2018.

[7] Zhang, Z.; Qiao, J. A node pruning algorithm for feedforward neural network based on neural complexity. In Proceedings of the 2010 International Conference on Intelligent Control and Information Processing, Dalian, China, pp. 406–410, 13–15 August 2010.

[8] Xu, X.; Wang, X.; Li, H.; Borisov, N.; Gunter, C.A.; Li, B. Detecting AI Trojans Using Meta Neural Analysis. arXiv:1910.03137, arXiv 2019,

[9] Wang, B.; Yao, Y.; Shan, S.; Li, H.; Viswanath, B.; Zheng, H.; Zhao, B.Y. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In Proceedings of the 2019 IEEE Symposium on Security and Privacy (SP), San Francisco, CA, USA, 1,pp. 707–723, 9–23 May 2019.

[10] Kaviani, S.; Shamshiri, S.; Sohn, I. A defense method against backdoor attacks on neural networks. Expert Syst. Appl. 2022, 213,118990.

[11] Shamshiri, S.; Han, K.J.; Sohn, I. DB-COVIDNet: A Defense Method against Backdoor Attacks. Mathematics 11, 4236, November 2023.

[12] S. Kaviani, I. Sohn and H. Liu, "Application of complex systems in neural networks against Backdoor attacks," 2020 International Conference on Information and Communication Technology Convergence (ICTC), Jeju, Korea (South), 2020, pp. 57-59, doi: 10.1109/ICTC49870.2020.9289220.

[13] Y. Liu, Sh. Ma, Y. Aafer, W. Lee, J. Zhai, W. Wang, X. Zhang, Trojaning attack on neural networks, in: 25nd Annual Network and Distributed System Security Symposium, NDSS 2018, San Diego, California, USA, pp. 18–221, 2018.

[14] Chen, H.; Fu, C.; Zhao, J.; Koushanfar, F. DeepInspect: A Black-box Trojan Detection and Mitigation Framework for Deep Neural Networks. IJCAI, 4658–4664, 2019.

[15] Liu, Y.; Lee,W.C.; Tao, G.; Ma, S.; Aafer, Y.; Zhang, X. ABS: Scanning neural networks for back-doors by artificial brain stimulation. In Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, London, UK, pp. 1265–1282, 11–15 November 2019.

[16] Deng, Z., Zhang, Y. (2007). Collective behavior of a small-world recurrent neural system with scale-free distribution. IEEE Transactions on Neural Networks, 18(5), 1364–1375.

[17] Breiman, L. Bagging predictors. Mach. Learn. 24, 123–140, 1996.

[18] Hinton, G.E.; Srivastava, N.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R.R. Improving neural networks by preventing co-adaptation of feature detectors. arXiv 2012, arXiv:1207.0580.