

Reliable backdoor attack detection for various size of backdoor triggers

Yeongrok Rah, Youngho Cho

Department of Cyber Security and Computer Engineering, Korea National Defense University, Nonsan, Republic of Korea

Article Info

Article history:

Received Feb 29, 2024

Revised Aug 30, 2024

Accepted Sep 30, 2024

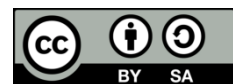
Keywords:

Adversarial attacks
Adversarial defense method
Backdoor attacks
Backdoor defense method
Deep learning
Poisoning attacks

ABSTRACT

Backdoor attack techniques have evolved toward compromising the integrity of deep learning (DL) models. To defend against backdoor attacks, neural cleanse (NC) has been proposed as a promising backdoor attack detection method. NC detects the existence of a backdoor trigger by inserting perturbation into a benign image and then capturing the abnormality of inserted perturbation. However, NC has a significant limitation such that it fails to detect a backdoor trigger when its size exceeds a certain threshold that can be measured in anomaly index (AI). To overcome such limitation, in this paper, we propose a reliable backdoor attack detection method that successfully detects backdoor attacks regardless of the backdoor trigger size. Specifically, our proposed method inserts perturbation to backdoor images to induce them to be classified into different labels and measures the abnormality of perturbation. Thus, we assume that the amount of perturbation required to reclassify the label of backdoor images to the ground-truth label will be abnormally small compared to them for other labels. By implementing and conducting comparative experiments, we confirmed that our idea is valid, and our proposed method outperforms an existing backdoor detection method (NC) by 30%p on average in terms of backdoor detection accuracy (BDA).

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Youngho Cho

Department of Cyber Security and Computer Engineering, Korea National Defense University

Hwangsansbeol-ro 1040, Yangchon-myeon, Nonsan-si, Chungcheongnam-do, Republic of Korea

Email: youngho@kndu.ac.kr

1. INTRODUCTION

As deep learning (DL) technology is gradually applied to various research fields such as image recognition and natural language processing, there has been a great increase in research on adversarial attacks [1]–[4]. This is because adversarial attacks strategically exploit vulnerabilities of DL models and thus, they can significantly break and degrade the integrity and reliability of DL models. Thus, the inherent nature of adversarial attacks undermines the robustness of DL models by introducing intentional distortions to induce misclassification. Therefore, to protect and defend DL models in the presence of adversarial attacks, it is necessary to conduct comprehensive research on adversarial attacks [5], [6].

In particular, backdoor attacks have rapidly evolved, significantly contributing to the escalating trend of malicious exploitation targeted at artificial intelligence models [7]–[10]. Backdoor attacks within the domain of DL are categorized under poisoning attacks, a subset of adversarial attacks [7], [11]. In these attacks, a backdoor trigger is intentionally incorporated into a DL model during its training phase. The manipulated DL model is specifically crafted to execute predetermined misclassifications at the time designated by the attacker. Therefore, backdoor attacks exemplify a sophisticated form of adversarial attack within the realm of DL. Particularly, it is noteworthy that backdoor attacks achieve a high attack success rate despite of a low poisoning

rate to a DL model [10], [12]. This poses a considerable challenge in identifying whether a DL model is poisoned by a backdoor attack [11]–[13].

In academia, the increasing risk of backdoor attacks has led to active research on both the execution and defense against such attacks [14]–[18]. One of notable detection techniques is neural cleanse (NC) that craftily inserts perturbations into benign images to identify abnormalities and thus detects backdoor attacks [18]. However, NC has the following significant limitation such that it fails to detect backdoor triggers above a certain threshold, that is when the size of the backdoor triggers exceeds 8×8 pixels. To address this limitation, this study aims to detect backdoor attacks regardless of the backdoor trigger sizes. Specifically, we propose a novel technique that identifies abnormal perturbations when perturbations are inserted to reclassify backdoor images by a DL model into their ground-truth labels.

The main contributions of this study can be summarized as follows. First, we proposed a novel idea to detect various sizes of backdoor triggers hidden in backdoor images. Specifically, our proposed method inserts perturbation to backdoor images to induce them to be classified into different labels and then measures the abnormality of perturbation. Thus, we assume that the amount of perturbation required to reclassify the label of backdoor images to the ground-truth label will be abnormally small compared to them for other labels. Second, we implemented our proposed method and conducted comparative experiments. According to our experimental results, we showed that our idea is valid and the proposed method outperforms an existing backdoor detection method (NC) by 30%p on average in terms of backdoor detection accuracy (BDA).

The rest of this paper is organized as follows. In section 2, we overview the background knowledge and existing studies. In section 3, we design our proposed method based on the analysis of general poisoning attacks. In section 4, we conduct extensive experiments and analyze the results. Finally, we conclude with future research directions in section 5.

2. BACKGROUND AND RELATED WORKS

2.1. Backdoor attacks and defenses

We will provide concise explanations of prevalent technical terms employed in backdoor learning. The identical definitions for these terms will be maintained throughout the rest of the paper. Backdoor refers to refers to malicious code that arises when a DL model is trained on contaminated data during its training phase [18]–[21]. Backdoor trigger signifies the pattern intended to activate the malicious backdoor. Target label refers to the label that the attacker induces through a backdoor attack to manipulate the model's classification. Ground-truth label refers to the actual label of a backdoor attack image. Anomaly index (AI) refers to an indicator measuring how far data points deviate from the distribution (calculating the absolute deviation between each data point and the median) [22]. AI is based on median absolute deviation (MAD), and higher AI values indicate anomalies [11], [13], [18]. AI can be calculated as in (1).

$$AI = \frac{|X - \text{Median}(X)|}{MAD} \quad (1)$$

Where X is a data point to measure AI, $\text{Median}(X)$ refers to the median value when the standard data is sorted in ascending order, and MAD is a metric that indicates the median of the absolute deviations between data points and the median.

The typical process of a backdoor attack follows these three stages: backdoor trigger design: In this stage, a pattern is crafted to be used as a trigger, considering both confidentiality and attack effectiveness. Dataset contamination and model training: The designed trigger is inserted into some training data to train the model. Trigger exploitation: When the model receives input containing the backdoor trigger, it outputs the target label with a high probability [8], [23]. Figure 1 depicts an example of a backdoor attack performed on a DL model designed to classify dogs and cats. In this scenario, the attacker manipulates the model during its training phase by inserting a red X pattern, serving as the backdoor trigger, into cat images while altering their labels to the target label, which is "dog." Subsequently, during the inference phase, when the attacker inserts the backdoor trigger into an image they intend to attack, the model misclassifies it as the target label, which in this case is "dog."

On the other hand, backdoor defense can be categorized into three types: Backdoor-trigger mismatch: This method nullifies the backdoor's functionality by altering the backdoor trigger or rendering it ineffective. This method incorporate a preprocessing module that alters the trigger patterns within targeted samples before feeding them into DL model. Consequently, the adjusted triggers no longer align with the concealed backdoor, thus thwarting the activation of the backdoor [13], [24]. Backdoor removal: This involves adding new data or modifying model parameters to eliminate the learned backdoor from the model. This method approaches focus on eliminating concealed backdoors within the compromised model by directly modifying suspicious models. Consequently, even if the trigger is present in attacked samples, the reconstructed model will make accurate

predictions as the hidden backdoors have been effectively eliminated [13], [24]–[26]. Trigger removal: These defense mechanisms filter out malicious samples during the inference phase rather than during the training process. Only benign testing or purified attacked samples are predicted by the deployed model. These defenses effectively prevent backdoor activation by eliminating trigger patterns [27], [28].

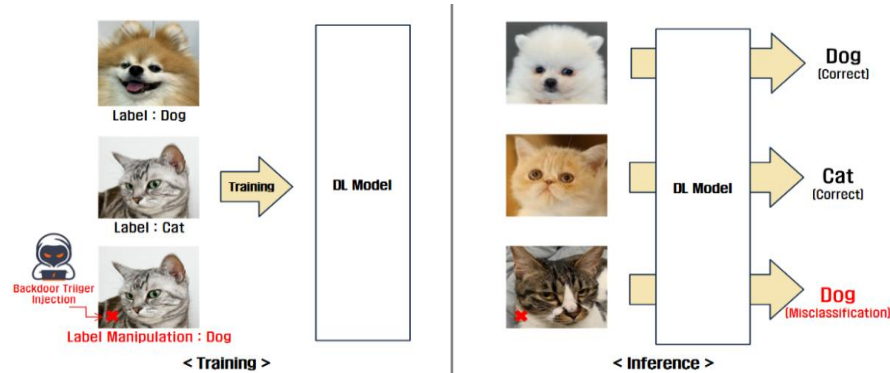


Figure 1. Backdoor attack process on a DL model classifying dogs and cats

2.2. Overview of neural cleanse

Wang *et al.* [18] pioneered a technique for detecting backdoor attacks by introducing perturbation, contributing significantly to the foundational research that shapes the location and patterns of backdoor triggers. NC employs anomaly detection methods to identify backdoor attacks. Typically, inducing a benign image to be classified into a different label than its actual category requires a substantial amount of perturbation. However, when a model is compromised with a backdoor attack, and the model has learned a small-sized backdoor trigger directing the image to the attack target label, even a slight abnormal perturbation can prompt the model to categorize the benign image as the target label.

NC detects this anomalously inserted perturbation and addresses the backdoor through a meticulous process. Figure 2 illustrates the functioning mechanism of NC. In this depiction, a DL model infected with a backdoor attack misclassifies images of '9' containing the backdoor trigger as the target label '0' during the inference process. NC strategically inserts perturbation into a benign image '9' and tests to classify it under different labels. While attempting to classify it from '1' to '8', a substantial amount of perturbation is required. However, an anomalously small perturbation is adequate for classifying it as the target label '0'. NC identifies this as abnormal perturbation, resembling the form of the backdoor trigger used in the attack.

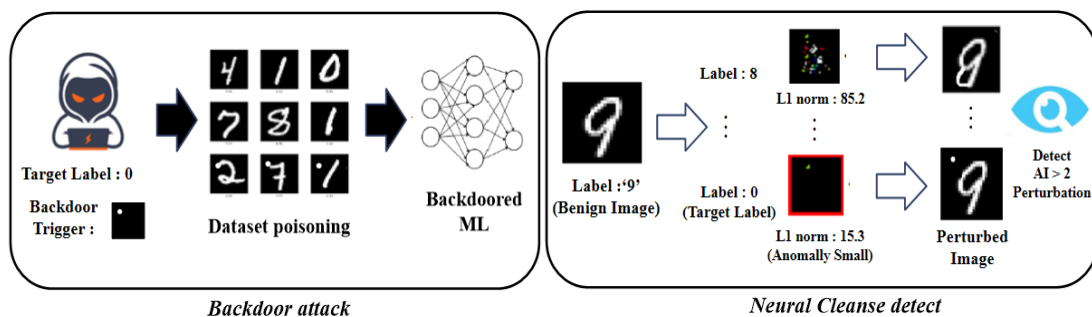


Figure 2. Backdoor detection process of NC for backdoor-infected Modified National Institute of Standards and Technology (MNIST) models

To effectively remove backdoors, NC targets neurons activated during the training of abnormal perturbations for backdoor defense. By eliminating these neurons that resemble the backdoor trigger, NC successfully eradicates the backdoor. This process demonstrates that NC is able to both identify and eliminate backdoor triggers hidden in compromised models. In addition, NC enhances the security and reliability of DL models by providing a robust defense against sophisticated backdoor attacks by careful analysis and precise neuron selection to ensure complete removal of backdoor triggers.

2.3. The critical limitation of neural cleanse

In the existing research [18], the backdoored model learned a small-sized backdoor trigger, exploiting the amount of perturbation introduced to induce a benign image towards the target label, resulting in abnormally small features. However, contrary to the assumption in prior research, if an attacker inserts a significantly larger-sized backdoor trigger, the amount of perturbation used to induce the target label may become similar to the perturbation used to induce other labels. This makes it undetectable as abnormal perturbation. As illustrated in Figures 3 and 4, when a backdoor trigger size of 4×4 px is used, it is noticeable that the required amount of perturbation is significantly smaller compared to the average. Therefore, the AI of the abnormal perturbation is higher than the threshold of 2, allowing proper detection. However, as the backdoor trigger size increases, the AI decreases rendering the abnormal perturbation undetectable. Therefore, our method aims to achieve this goal regardless of the backdoor trigger's size while maintaining a high backdoor detection rate. This approach addresses limitations identified in previous research, contributing to the strengthening of the robustness of backdoor detection mechanisms.

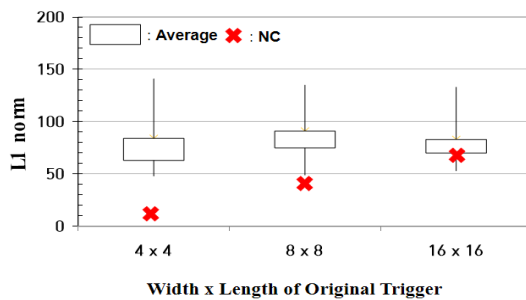


Figure 3. L1 norm depending on various backdoor trigger sizes

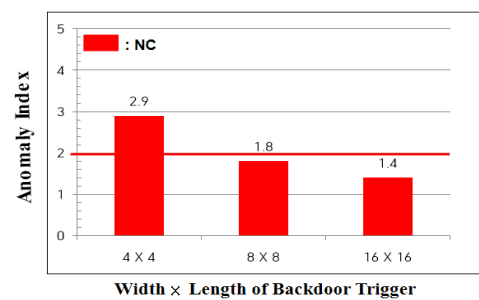


Figure 4. AI depending on various backdoor trigger sizes

3. PROPOSED METHOD

In this study, we assume the followings. The attacker can access and use a target DL model to launch poisoning attacks on it (white box attack). Thus, the attacker can insert backdoor triggers into a part of training dataset and then manipulate their labels to induce misclassifications in the model. In addition, the defender uses our proposed technique to detect the existence of backdoor triggers in the target DL model [18], [19].

To address the critical limitation of NC described in section 2.3, we leverage the inherent characteristics of the original label within the backdoor images. Specifically, we assume that backdoor images still possess the characteristics of the original label (i.e., ground-truth label) as well as the features of the target label. Thus, we expect the amount of perturbation inducing a backdoor attack to the ground-truth label to be abnormally small and thus we can use the abnormality to determine the existence of backdoor triggers in images. Based on this speculation, we intentionally add some perturbation to backdoor images such that a target DL model misclassifies them into different labels, unlike the attacker's target labels; in this case, our target label is the ground-truth label. By this approach, we believe that our proposed method overcomes the limitation of NC since the perturbation inducing backdoor images to be classified into their ground-truth labels is consistently abnormally small, regardless of the backdoor trigger's size; we show the validity of our idea in section 5.

Our proposed method is different from existing methods in the following two aspects:

- Misclassification of the target model: Instead of adding perturbation to benign images to make them backdoor images, our method adds perturbation to misclassified images which are suspicious to be backdoor images with backdoor triggers to induce them to be classified into different labels to see if they are under backdoor attacks.
- Utilization of ground-truth label: Instead of detecting abnormality in perturbation added to benign images for the target label, we examine the amount of perturbation added to misclassified images for each class label based on our assumption that the induced perturbation targeting the ground-truth label will be relatively very small compared with perturbation required to generating other labels.

Figure 5 illustrates how our proposed method works to detect the existence of a backdoor trigger in a misclassified image by using an example. In the upper part, the backdoored model misclassifies an image with a backdoor trigger into the target label 0 although its original ground truth label is 9. In the bottom part, our proposed method add perturbation to a misclassified suspicious image such that the DL model reclassifies it to different labels ranging from Label 1 to Lable 9 except Label 0 (backdoor attacker's target label); a process of adding perturbation continues until the image is reclassified successfully into a designated each label. As we

can see in the figure, the amount of perturbation added to reclassify it to the ground truth label will be very small compared to them for generating other labels.

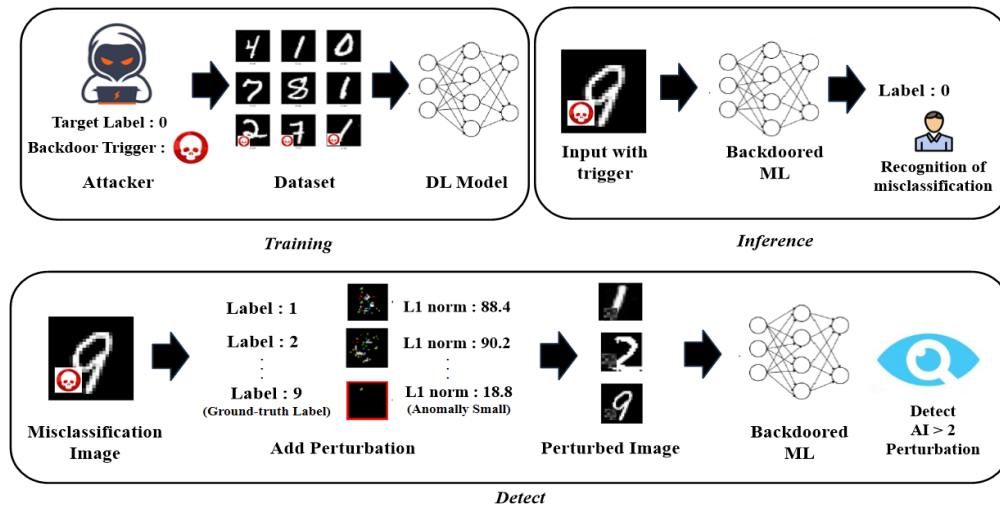


Figure 5. The proposed method that detects the abnormality in L1 norm occurred when classifying a backdoor attack image with a ground-truth label (Label: 9) as its original target label (Label: 0)

4. EXPERIMENT

4.1. Experimental purpose and setup

The main experimental purpose is to validate our idea to detect large size of backdoor triggers NC cannot detect and show the detection performance of our proposed method in our experimental setup. For our experiments, we implemented experiment programs in the Anaconda software's virtual environment based on Python 3.9 and Tensorflow 2.10 framework and run them on AMD Ryzen 5 5600G CPU and a GeForce RTX 3060 12 GB RAM GPU [29], [30]. The details on our experimental setup are described as follows.

- Target DL model and dataset: To construct the target DL model, we used a convolutional neural network (CNN) model trained on the MNIST dataset which is commonly used in previous studies on poisoning attack and defense [8], [10]. The MNIST dataset consists of 28×28 -pixel grayscale images that are classified into 10 classes; 10 classes represent digits from 0 to 9. It consists of 50,000 training images and 10,000 test images. The CNN model is a standard architecture for image classification tasks and has been widely used for MNIST dataset classification [13]. The parameters of each layer are shown in Table 1. The baseline CNN shows 99.5% of detection accuracy for MNIST digit recognition. To align with an existing method and experimental setup, we use CNN model with 0.28 million parameters [17], [18].
- Backdoor attack methods: To taint the target DL model, we used BadNets attacks [9], [10], [12]. Specifically, by using the backdoor attack methods, we created a tainted MNIST dataset D_t that includes a poisoned dataset D_p . At this point, the original ground-truth label of D_p is '9' and the attacker's target label is set to '0'. Therefore, we manipulated the labels of D_p and then trained constructed a backdoored model by training a CNN model with D_t generated at poisoning ratio = 3%. In addition, we used a white square backdoor trigger pattern added to the bottom right corner of images, and thus various sizes of backdoor trigger (1×1 , 2×2 , 3×3 , 4×4 , 8×8 , and 16×16 px) was created [10], [14].
- Backdoor detection methods: To compare the performance of our proposed method and an existing method NC for various backdoor trigger sizes, we tested 50 times for each backdoor trigger size and averaged their detection results. NC measures the perturbation required when inducing a benign image to be classified into a different label and assesses if it is anomalously small when induced towards the target label. On the other hand, our method measures the perturbation needed when inducing a backdoor image to be classified into a different label and evaluates if it is anomalously small when induced towards the ground-truth label [18].
- Evaluation metrics: To measure the performance of our proposed method, we used the following two evaluation metrics. The first metric is AI that measures the degree of abnormality in a backdoored model [22]. To quantify the degree of abnormality, the amount of perturbation inserted to classify the image into a different label is calculated using L1 norm. The backdoor attack detection threshold is set to be the same as NC, with $AI > 2$. The second metric is BDA that measures the probability of backdoor detection on backdoored model. BDA can be measured as in (2).

$$BDA(\%) = \frac{\text{The number of detected backdoor images}}{\text{The number of backdoor test images}} \times 100 \quad (2)$$

Table 1. Architecture of target DL model

| Layer | Input | Filter | Stride | Output | Activation |
|--------|----------|-----------|--------|----------|------------|
| conv 1 | 1×28×28 | 16×1×5×5 | 1 | 16×24×24 | ReLU |
| pool 1 | 16×24×24 | 2×2 | 2 | 16×12×12 | / |
| conv 2 | 16×12×12 | 32×16×5×5 | 1 | 32×8×8 | ReLU |
| pool 2 | 32×8×8 | 2×2 | 2 | 32×4×4 | / |
| fc1 | 32×4×4 | / | / | 512 | ReLU |
| fc2 | 512 | / | / | 10 | Softmax |

4.2. Experimental results and analysis

First, our proposed method successfully detected backdoor images with 8×8 or 16×16 backdoor triggers whose AI > 2 while an existing state-of-the-art backdoor detection method (NC) could not detect them. Figure 6(a) shows the perturbation value in terms of L1 norm required when NC classifies a test image (Label 9) into a different label (from Label 0 to Label 8 except Label 9). For example, in the case of using a 4×4 px backdoor trigger size, the amount of perturbation needed to classify a test image into the target label '0' is significantly lower compared to the amount of perturbation needed for other labels (Label 1-Label 8). However, as the backdoor trigger size grows, the perturbation value becomes closer to those for other labels. On the other hand, Figure 6(b) shows the perturbation values measured when our proposed method was used. As we can see in the figures, our method needs relatively very small perturbation when classifying backdoor images (Label 0) into the ground-truth label (Label 9) regardless of backdoor trigger sizes. Similarly, Figure 7(a) represents the calculated AI values for perturbation using NC technique. When the backdoor trigger sizes are equal to or greater than 8×8 px, the AI values were lower than the backdoor detection threshold of 2. In other words, it is not possible to detect the backdoor attack when the backdoor trigger size is 8×8 or larger [9]. On the other hand, Figure 7(b) shows the AI values measured when our proposed method was used. Consequently, the AI values also exceeded the backdoor detection threshold of 2 and thus our proposed method successfully detected backdoor images with various size of backdoor triggers.

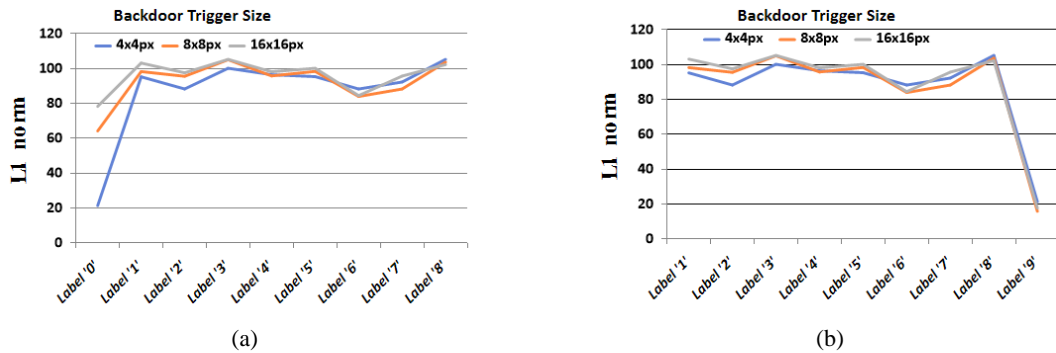


Figure 6. Comparison of L1 norm when test images are classified into different labels using (a) NC and (b) our method

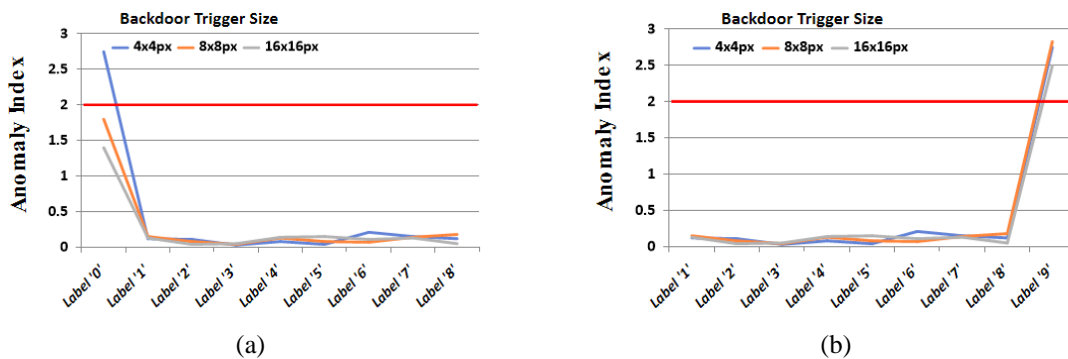


Figure 7. Comparison of AI when test images are classified into different labels using (a) NC and (b) our method

Second, our proposed method showed a high, stable detection performance for all sizes of backdoor triggers. Specifically, as indicated in Table 2, the average BDA of our proposed method is 96.3% while the NC showed around 64.3% of BDA on average. This means that our proposed method outperforms the existing state-of-the-art detection method NC by around 32%p in terms of backdoor attack detection accuracy. In particular, BDA of our proposed method for 8×8 or 16×16 backdoor trigger sizes are 98% and 94%, respectively, while the NC could not detect them at all (BDA = 0%). This result confirms that our proposed method can resolve the critical vulnerability of NC and thus shows stable, reliable detection performance for various backdoor trigger sizes.

Table 2. Comparison of BDA for various backdoor trigger sizes

| Trigger Size | BDA (%) (# of detected backdoor samples / # of test backdoor samples) | |
|--------------|---|---------------------|
| | Neural cleanse | Our proposed method |
| 1×1 | 96 (48 / 50) | 96 (48 / 50) |
| 2×2 | 98 (49 / 50) | 98 (49 / 50) |
| 3×3 | 98 (49 / 50) | 96 (48 / 50) |
| 4×4 | 94 (47 / 50) | 96 (48 / 50) |
| 8×8 | 0 (0 / 50) | 98 (49 / 50) |
| 16×16 | 0 (0 / 50) | 94 (47 / 50) |
| Average | 64.3 | 96.3 |

5. CONCLUSION

In this study, we introduced a novel approach to detect backdoor attacks in DL models by utilizing perturbation insertion to identify abnormal perturbation within the data. The primary objective of this method was to enhance the capabilities of backdoor attack detection. By implementing this technique, our goal was to overcome the limitations observed in prior research and provide a robust method capable of effectively identifying backdoor attacks, regardless of the size of the inserted backdoor trigger. The validation of this approach was conducted through a series of preliminary experiments, confirming its efficacy and potential for practical application. Our future research directions are as follows. First, we aim to delve deeper into refining and optimizing the proposed technique, striving to improve its precision, scalability, and applicability across various models and datasets. Second, we plan to explore additional defensive strategies against backdoor attacks, including investigating methods to effectively remove detected backdoor triggers or disrupt the connection between triggers and backdoors. These strategic measures are designed to fortify the resilience of DL models against potential backdoor threats in real-world scenarios.




REFERENCES

- [1] D. J. Miller, Z. Xiang, and G. Kesidis, "Adversarial learning targeting deep neural network classification: a comprehensive review of defenses against attacks," *Proceedings of the IEEE*, vol. 108, no. 3, pp. 402–433, 2020, doi: 10.1109/JPROC.2020.2970615.
- [2] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry, "Adversarial examples are not bugs, they are features," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [3] K. Ren, T. Zheng, Z. Qin, and X. Liu, "Adversarial attacks and defenses in deep learning," *Engineering*, vol. 6, no. 3, pp. 346–360, 2020, doi: 10.1016/j.eng.2019.12.012.
- [4] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, "A survey on adversarial attacks and defences," *CAAI Transactions on Intelligence Technology*, vol. 6, no. 1, pp. 25–45, 2021, doi: 10.1049/cit2.12028.
- [5] S. Aneja, N. Aneja, P. E. Abas, and A. G. Naim, "Defense against adversarial attacks on deep convolutional neural networks through nonlocal denoising," *IAES International Journal of Artificial Intelligence*, vol. 11, no. 1, pp. 1–14, Jun. 2022, doi: 10.11591/ijai.v11.i3.pp961-968.
- [6] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015.
- [7] Y. Li, S. Zhang, W. Wang, and H. Song, "Backdoor attacks to deep learning models and countermeasures: a survey," *IEEE Open Journal of the Computer Society*, vol. 4, pp. 134–146, 2023, doi: 10.1109/OJCS.2023.3267221.
- [8] A. Shafahi *et al.*, "Poison frogs! Targeted clean-label poisoning attacks on neural networks," *Advances in Neural Information Processing Systems*, vol. 2018, pp. 6103–6113, 2018.
- [9] A. Turner, D. Tsipras, and A. Madry, "Clean-label backdoor attacks," *The International Conference on Learning Representations*, pp. 1-21, 2019.
- [10] T. Gu, K. Liu, B. Dolan-Gavitt, and S. Garg, "BadNets: evaluating backdooring attacks on deep neural networks," *IEEE Access*, vol. 7, pp. 47230–47243, 2019, doi: 10.1109/ACCESS.2019.2909068.
- [11] Y. Li, Y. Jiang, Z. Li, and S. T. Xia, "Backdoor learning: a survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 1, pp. 5–22, 2024, doi: 10.1109/TNNLS.2022.3182979.
- [12] E. Wenger, J. Passananti, A. N. Bhagoji, Y. Yao, H. Zheng, and B. Y. Zhao, "Backdoor attacks against deep learning systems in the physical world," *The IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 6202–6211, 2021, doi: 10.1109/CVPR46437.2021.00614.
- [13] Y. Liu *et al.*, "A survey on neural trojans," *2020 21st International Symposium on Quality Electronic Design (ISQED)*, Santa Clara, CA, USA, 2020, pp. 33–39, doi: 10.1109/ISQED48828.2020.9137011.




- [14] B. Zhao and Y. Lao, "Towards class-oriented poisoning attacks against neural networks," *2022 IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022*, pp. 2244–2253, 2022, doi: 10.1109/WACV51458.2022.00230.
- [15] J. Lin, L. Xu, Y. Liu, and X. Zhang, "Composite backdoor attack for deep neural network by mixing existing benign features," *Proceedings of the ACM Conference on Computer and Communications Security*, pp. 113–131, 2020, doi: 10.1145/3372297.3423362.
- [16] H. Park and Y. Cho, "A dilution-based defense method against poisoning attacks on deep learning systems," *International Journal of Electrical and Computer Engineering*, vol. 14, no. 1, pp. 645–652, 2024, doi: 10.11591/ijece.v14i1.pp645-652.
- [17] S. Huang, W. Peng, Z. Jia, and Z. Tu, "One-pixel signature: characterizing CNN models for backdoor detection," *Computer Vision – ECCV 2020*, pp. 326–341, 2020, doi: 10.1007/978-3-030-58583-9_20.
- [18] B. Wang et al., "Neural cleanse: Identifying and mitigating backdoor attacks in neural networks," *2019 IEEE Symposium on Security and Privacy (SP)*, San Francisco, CA, USA, 2019, pp. 707–723, doi: 10.1109/SP.2019.00031.
- [19] X. Sun et al., "Defending against backdoor attacks in natural language generation," *Proceedings of the 37th AAAI Conference on Artificial Intelligence, AAAI 2023*, vol. 37, pp. 5257–5265, 2023, doi: 10.1609/aaai.v37i4.25656.
- [20] Z. Xiang, D. J. Miller, and G. Kesidis, "Detection of backdoors in trained classifiers without access to the training set," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 3, pp. 1177–1191, 2022, doi: 10.1109/TNNLS.2020.3041202.
- [21] J. Guo, A. Li, and C. Liu, "Aeva: Black-box backdoor detection using adversarial extreme value analysis," *ICLR 2022 - 10th International Conference on Learning Representations*, pp. 1–24, 2022.
- [22] F. R. Hampel, "The influence curve and its role in robust estimation," *Journal of the American Statistical Association*, vol. 69, no. 346, 1974, doi: 10.2307/2285666.
- [23] W. Guo, B. Tondi, and M. Barni, "An overview of backdoor attacks against deep neural networks and possible defences," *IEEE Open Journal of Signal Processing*, vol. 3, pp. 261–287, 2022, doi: 10.1109/OJSP.2022.3190213.
- [24] B. G. Doan, E. Abbasnejad, and D. C. Ranasinghe, "Februus: input purification defense against trojan attacks on deep neural network systems," *ACM International Conference Proceeding Series*, pp. 897–912, 2020, doi: 10.1145/3427228.3427264.
- [25] D. Wu and Y. Wang, "Adversarial neuron pruning purifies backdoored deep models," *Advances in Neural Information Processing Systems*, vol. 20, pp. 16913–16925, 2021.
- [26] Y. Zeng, S. Chen, W. Park, Z. M. Mao, M. Jin, and R. Jia, "Adversarial unlearning of backdoors via implicit hyper gradient," *ICLR 2022 - 10th International Conference on Learning Representations*, pp. 1–28, 2022.
- [27] M. Javaheripi, M. Samragh, G. Fields, T. Javidi, and F. Koushanfar, "CleanNN: Accelerated trojan shield for embedded neural networks," *2020 IEEE/ACM International Conference On Computer Aided Design (ICCAD)*, San Diego, CA, USA, 2020, pp. 1–9.
- [28] Y. Gao, C. Xu, D. Wang, S. Chen, D. C. Ranasinghe, and S. Nepal, "StriP: A defence against trojan attacks on deep neural networks," *ACM International Conference Proceeding Series*, pp. 113–125, 2019, doi: 10.1145/3359789.3359790.
- [29] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, 2014, pp. 1891–1898, doi: 10.1109/CVPR.2014.244.
- [30] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, "Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition," *Neural Networks*, vol. 32, pp. 323–332, 2012, doi: 10.1016/j.neunet.2012.02.016.

BIOGRAPHIES OF AUTHORS



Yeongrok Rah    received B.S. degree in international relations from Korea Naval Academy, Jinhae, Republic of Korea. He is currently a Lieutenant in Republic of Korea Navy and pursuing the M.S. degree in Department of Cyber Security and Computer Engineering with Korea National Defense University, Nonsan, Republic of Korea. His research interests include deep learning, adversarial machine learning, and cyberwarfare. He can be contacted at email: skdudfhr789@gmail.com.



Youngho Cho    received the B.S. degree in industrial engineering from Korea Air Force Academy, Republic of Korea, in 1998 and the M.S. degree in computer science and industrial systems engineering from Yonsei University, Republic of Korea, in 2006 and the Ph.D. degree in electrical and computer engineering from University of Maryland, College Park, MD, USA, in 2013. He is an Associate Professor with Department of Cyber Security and Computer Engineering, Graduate School of Defense Management, Korea National Defense University, Nonsan, Republic of Korea. His research interests include wireless network security, trust mechanism, botnet detection, steganography-based covert communication, adversarial machine learning, and AI security. He has authored or coauthored more than 70 peer-reviewed papers published in journals and in the proceedings of conferences. He is an IEEE senior member. He can be contacted at the email: younghocho@korea.kr.