A backdoor defense method on COVID-19 based on deep neural network

Samaneh Shamshiri Insoo Sohn

Division of Electronics and Electrical Engineering
Dongguk University

Samaneh.shamshiri@gmail.com isohn@dongguk.edu

Abstract

Computer vision and deep learning techniques can assist in determining COVID-19 infection with chest X-ray images. However, recent studies in the field of security of AI-based systems revealed that these open-source DNNs are vulnerable to attacks. Since adversaries can access open-source model parameters and training images, they threaten security-critical learning applications. To alleviate this problem, we proposed a COVID-based CNN model motivated by the algorithm of the dropout layer and bagging network to remove trigger-related features during the training process. Through this model, we achieve sufficiently high and stable accuracy on clean data and an exceeding reduction in attack success rate.

1. Introduction

The unprecedented success of machine learning techniques, especially in deep neural networks (DNN) medical image processing, has led to recent prominence in improving efficient diagnosis of infectious diseases such as COVID-19 with increased detection accuracy [4]. Due to the increasing demand for third parties and MLaaS (machine learning as a service) [10], for taking charge of the training procedure, these deep learning-based models, especially open source and open access DNNs, are vulnerable to attacks [3]. 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] [9]. The most important property of a backdoor attack is that the trained DNN model performs well on benign samples while DNN's prediction will be maliciously modified. On the other hand, a line of research has been focused on various defense techniques against a backdoor attack [6], which are categorized into detection and mitigation techniques. In this work, we proposed a new defense algorithm based on the intrinsic properties of the bagging and dropout algorithm. Since triggers are the most important part of backdoor attacks [11], by this method, we aim to remove trigger-related features through the modified dropout algorithm during the training process of our bagging network. We considered the AC-COVIDNet [8] as our main classifier, which is trained on the COVIDx [1] dataset including three classes: COVID-19, Pneumonia, and normal chest X-ray images. Experiments of our defense method on the AC-COVIDNet have shown the effectiveness and efficiency of the proposed defense method against backdoor attacks.

The paper is organized as follows. In Section 2, we concentrate on a brief explanation of the architecture of the dropout algorithm, bagging network, and our proposed defense algorithm. Furthermore, we demonstrate both attacks and defense effects on AC-COVIDNet performance by a confusion matrix. Finally, we will conclude the paper by discussing the results and some conclusions.

2. Methodology and Experimental analysis

2.1 Dropout algorithm:

Dropout is one of the techniques widely used for handling the extremely overfitting problem of neural networks [7]. In this method, input features or the output weights of each hidden unit randomly drop out with the probability p during the training process. It means that the dropout layer removes ρ% of neurons during training, along with all its incoming and outgoing connections, so the gradient of each parameter is averaged in each mini-batch, and those training cases that do not use a parameter contribute a gradient of zero.

2.2 Bootstrap Aggregation:

Bootstrap aggregation (bagging) is one of the most well-known algorithms with impressive performance for building an ensemble of models for regression and classification tasks [5]. The bagging network is divided into two parts. First, bootstrapping sampling generates n new data samples randomly with replacement. Then, by aggregating samples, one can obtain a bagging prediction by calculating the arithmetic average voting or majority voting.

2.3 The proposed backdoor defense method:

To explore the applicability of our defense method against backdoor attacks, we conducted experiments for our bagging-based network with the AC-COVIDNet model as our base inducers. AC-COVIDNet is an extension of COVID-Net model. This model has an attention constructive architecture for the detection of COVID-19 through chest X-ray images. We contaminated 10% of the input chest X-ray images to leverage trigger to the network, then modified their labels to COVID-19 as attacked the desired label. Our main strategy is to optimize the training dataset by removing triggers from the input layer to avoid a backdoor into the network. Inspired by the inherent properties of the dropout algorithm, we modified the algorithm in order to nullify 4% of the features regardless of randomly selecting them. We applied our proposed dropout layer as the first layer of each inducer in our bagging network. In this study, we used a 25-pixel pattern trigger, which is placed at the right bottom corner of the image. During the training process, one patch of input features for each base learner will be dropped. However, we can

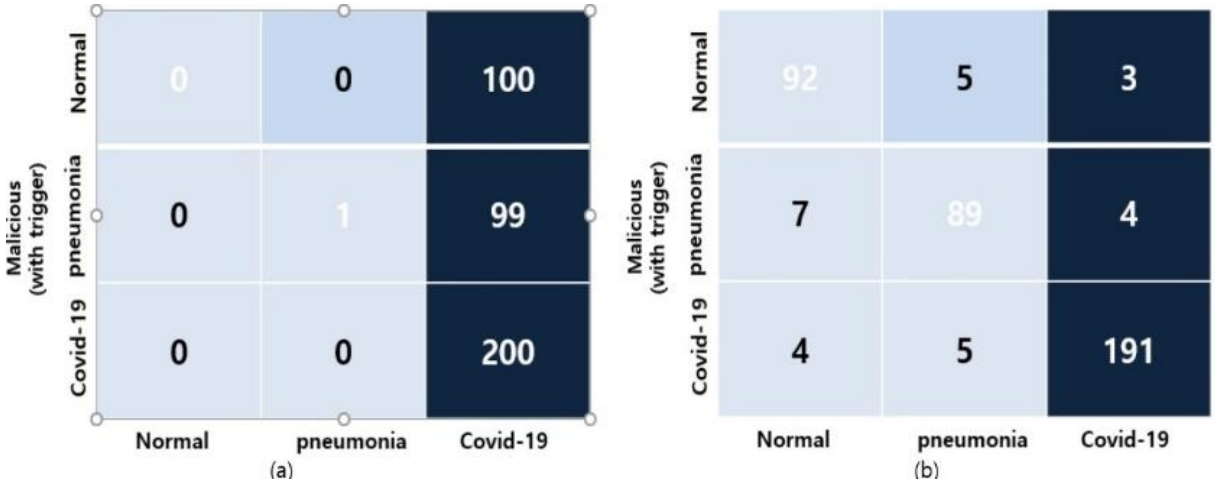

Figure 1: (a) Confusion matrices of backdoored AC-COVIDNet, (b) Confusion matrices of performance of defense algorithm on AC-COVIDNet

conclude that the trigger-related features will be removed by zeroing the features of one of the patches they belong to. Then, we regard 25 inducers that trained independently with the same hyper-parameters in our bagging network. As mentioned before, we considered AC-CovidNet as our main sub-model in the bagging network. Finally, we defined a new voting method for the bagging networks including two conditions as follows: If all classifiers, classify x_i as $\mathcal{S}_j$ class, and one of the learners classifies it as $\mathcal{S}_h$ class, while$j \neq h$, our proposed algorithm forecast the final result as $\mathcal{S}_h$ class, otherwise we consider majority voting.

It can be observed from table figure 1(a) that AC-COVIDNet [1] is vulnerable to backdoor attacks. Since we consider COVID-19 as the target class, backdoor attacks tended to classify most of the normal and pneumonia images as COVID-19. Hence, the network has strongly biased and the confusion matrix demonstrated high false positives for COVID-19 classes. The confusion matrix in figure 1 (b) depicts the performance efficiency of our proposed defense technique against backdoor attacks for our bagging-based defense algorithm. The high values were achieved for True Positive (TP) and True Negative (TN) on the triggered inputs. For instance, the number of TN for the COVID-19 class for AC-COVIDNet as a base classifier of the bagging algorithm is equal to 189 on malicious inputs.

3. Conclusion

In this paper, we proposed a defense method motivated by a dropout algorithm and bagging network, to make the convolutional neural networks robust against backdoor attacks. Experiments on the COVIDx dataset and AC-COVIDNet as the main inducer of the bagging network and the modified dropout algorithm with the new proposed voting strategy for the bagging algorithm maintain high accuracy on benign samples and diminish the probability of misclassification caused by the backdoor attacks. Therefore, the proposed dropout bagging-based defense method has proven to be effective against BDs that provide both stability and robustness for CNNs.

Acknowledgments

This work was supported by the Korea Institute of Energy Technology Evaluation and Planning (KETEP) and the Ministry of Trade, Industry & Energy (MOTIE) of the Republic of Korea (No. 20224000000020).

References

[1] L. Wang, A. Wong," COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest radiography images, 2020, arXiv preprint arXiv: 2003.09871.

[2] Y. Matsuo, K. Takemoto, Backdoor Attacks to Deep Neural Network-Based System for COVID-19 Detection from Chest X-ray Images, MDPI, 2

[3] H. Hirano, K. Koga, K. Takemoto, Vulnerability of deep neural networks for detecting COVID-19 cases from chest X-ray images to universal adversarial attacks, PLoS One 15 (12) (2020).

[4] S. Shamshiri, I. Sohn, "Security methods for AI-based COVID-19 analysis system: A survey", ICT Express, 2022

[5]. L. Breiman, Bagging predictors, Mach. Learn., vol. 24, no. 2, pp.123140, 1996.

[6] S. Kaviani, I.Sohn, "Adversarial attacks and defenses on AI in medical imaging informatics: A survey", Expert Systems with Applications, 2022.

[7] Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R.: Improving neural networks by preventing coadaptation of feature detectors (2012). arXiv:1207.0580.

[8] A. Ambati, Sh. Dubey, "AC-CovidNet: Attention Guided Contrastive CNN for Recognition of Covid-19 in Chest X-Ray Images", arXiv:2105.10239v2

[9] T. Gu, K. Liu, B. Dolan-Gavitt, S. Garg, "BadNets: Evaluating Backdooring Attacks on Deep Neural Networks", IEEE: 2019

[10] Xiaojun Xu, Qi Wang, Huichen Li, Nikita Borisov, Carl A Gunter, and Bo Li. Detecting ai Trojans using meta neural analysis. In IEEE S&P, 2021.

[11] Yi Zeng, Won Park, Z Morley Mao, and Ruoxi Jia. Rethinking the backdoor attacks' triggers: A frequency perspective. In ICCV, 2021a.