Overactivation - Verifying and testing backdoors in deep neural networks

Xin Wu

February 15, 2024

Abstract

– is not the final version

With the widespread integration of deep neural networks (DNNs) into contemporary societal applications, concerns are mounting over their susceptibility to malicious attacks, one such of it is backdoor attacks. Instances such as Tesla's autonomous driving accidents and the prevalence of adversarial attacks on AI image generators underscore the urgent need for robust security measures within DNNs.

This paper proposes a novel backdoor testing approach that leverages the phenomenon of neuron overactivation within DNNs to directly detect the presence of backdoors in models. To evaluate the proposed method, extensive testing and research were conducted on popular datasets and across different models. Experimental results demonstrate the effectiveness of the approach in revealing variations between various model structures and datasets.

In conclusion, this study introduces a valuable tool to mitigate the risk of backdoor attacks in DNNs, providing a reliable testing method to enhance the overall security of these systems. The effectiveness demonstrated in discerning potential threats highlights the potential of this method in fortifying artificial intelligence systems against malicious manipulation.

Contents

Chapter 1

Introduction

Deep neural networks (DNNs) have revolutionized image recognition in various fields, including autonomous driving[1], medical diagnostics[3], and security surveillance[2]. Their remarkable success, however, is shadowed by their vulnerability to adversarial attacks, which pose significant threats to their reliability and security. Among these adversarial threats, backdoor attacks have emerged as a particularly challenge.

Backdoor attacks involve the manipulation of DNN behavior by malicious actors through the injection of images with triggers into the training dataset. Once deployed, these triggers can activate specific behaviors or misclassifications, compromising the integrity and trustworthiness of the entire system. Detecting and mitigating such attacks are paramount for ensuring the robustness and trustworthiness of DNN-based image recognition systems.

1.1 Aims

Related work: One of the existing model backdoor detection tools, NeuralCleanse, operates under the assumption that smaller modifications will result in misclassification by the model. Another detection method, STRIP, requires prior knowledge of trigger-related information to ascertain whether a model has been injected with a backdoor. This project aims to explore a novel method that offers greater flexibility in usage conditions compared to existing tools and methods.

Additionally, it aims to provide insights that could inspire advancements in the field of backdoor detection to some extent.

1.2 Structure of this Project

Chapter 2

Background

2.1 literature review

Table 2.1: The information about each models
(for CIFAR100 the attack have some bugs So I need retraining it later.)

Models	epoch	Clean Accuracy	ASR	Architecture
MNIST	5	99.10%	99.99%	2 conv×1 dense
GTSRB	5	96.255%	91.100%	5 conv
FashionMNIST	5	92.81%	99.98%	2 conv×1 dense
CIFAR10	50	91.44%	99.99%	ResNet18
CIFAR100	100	73.84%	100%	ResNet50

Table 2.2: Below showed the Activation index(S_A means the set of maximum value which produced by the feature map from 1st Relu() for each noise images) for each dataset in different epoch. Totally have 200 noise images.

Model		MNIST			GTSRB			
		5 epoch	10 epoch	20 epoch	10 epoch	20 epoch	30 epoch	50 epoch
max(S_A)	**Backdoor**	3.9615	3.5404	5.0043	4.7136	4.3301	4.611	3.5434
	Clean	3.2132	3.9103	5.1311	4.345	3.3928	3.5097	3.2779
	B/C	**1.2328**	**0.9054**	**0.9752**	**1.0848**	**1.2762**	**1.3138**	**1.081**
mean(S_A)	**Backdoor**	3.2140	2.6816	3.6324	3.6496	2.9364	3.1777	2.3143
	Clean	2.5139	3.0669	3.8204	3.216	2.5081	2.0844	2.3850
	B/C	**1.2784**	**0.8743**	**0.9507**	**1.1348**	**1.1708**	**1.5245**	**0.9704**
Ratio($1stRelu_p$>$1stRelu_c$)		**199:1**	**17:183**	**38:162**	**163:37**	**177:23**	**195:5**	**109:91**
Model		FashionMNIST			CIFAR10			
		5 epoch	10 epoch	30 epoch	20 epoch	30 epoch	40 epoch	50 epoch
max(S_A)	**Backdoor**	1.5759	1.8628	2.4023	12.936	8.8327	5.6955	4.2566
	Clean	1.3684	1.8138	2.6442	9.8593	7.4869	6.5439	4.6565
	B/C	**1.1516**	**1.027**	**0.9085**	**1.312**	**1.1798**	**0.8704**	**0.9141**
mean(S_A)	**Backdoor**	1.313	1.4639	2.0127	9.9769	6.5749	4.387	3.4229
	Clean	1.2007	1.4822	2.2154	7.7057	5.2923	5.2367	3.6726
	B/C	**1.0934**	**0.9877**	**0.9085**	**1.2947**	**1.2423**	**0.8377**	**0.932**
Ratio($1stRelu_p$>$1stRelu_c$)		**189:11**	**95:105**	**12:188**	**196:4**	**194:6**	**11:189**	**60:140**

Bibliography

[1] M. Bojarski, D Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, X. Zhang, J. Zhan, and K. Zieba. "end to end learning for self-driving cars". *Arxiv*, 20(2), 2016.

[2] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi. "deep learning for person re-identification: A survey and outlook". *IEEE TPAMI*, 20(2), 2021.

[3] S. Kevin Zhou, Hayit Greenspan, and Dinggang Shen. *"Deep Learning for Medical Image Analysis"*. Elsevier, 2017.