Beijing Forest Studio
北京理工大学信息系统及安全对抗实验中心

Deep Learning Backdoor Attacks Detection

PhD Student : saba zaib

导师： Professor LUO SENLIN

2023年06月24日

List of Contents

My Introduction

Early life and Education

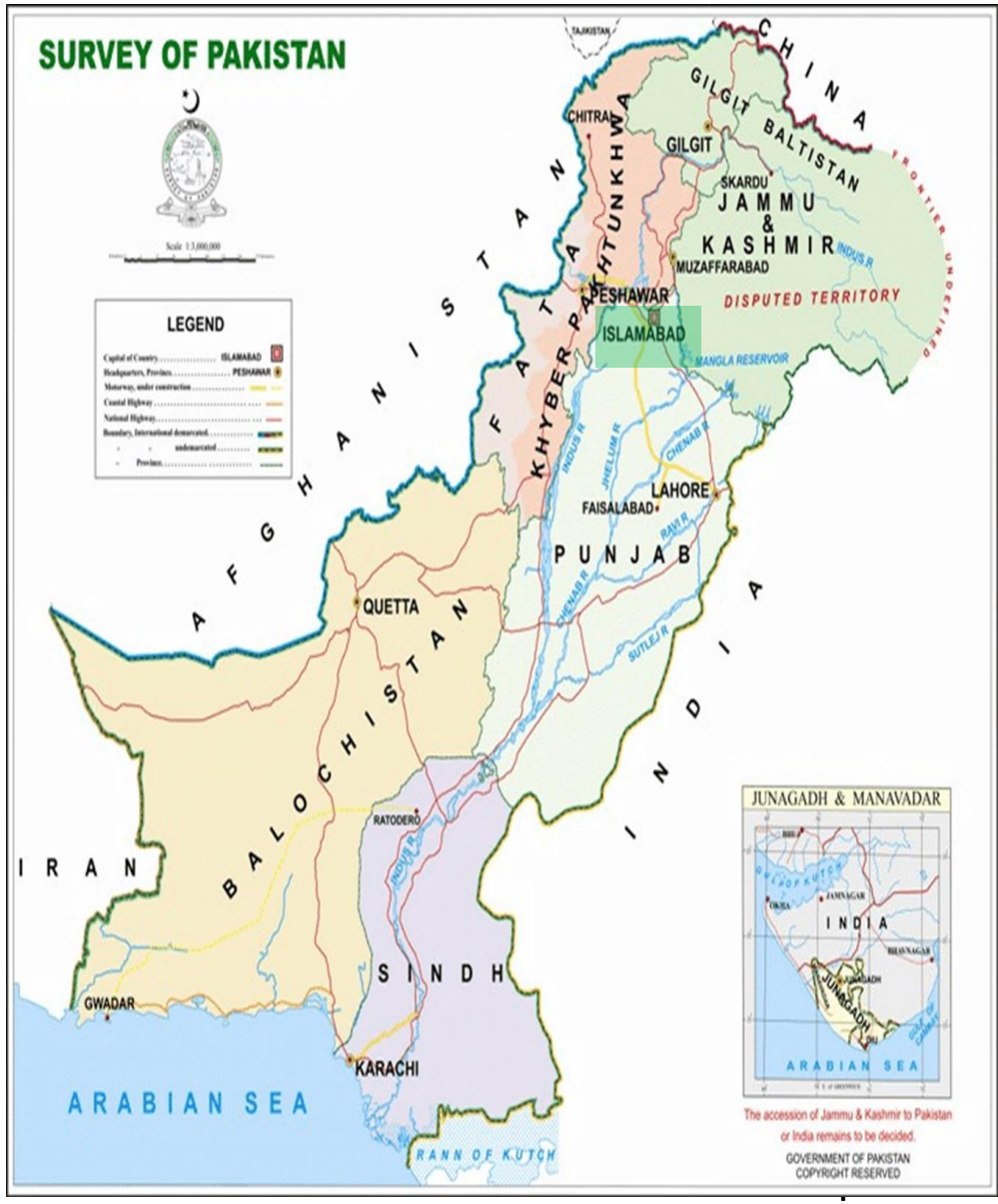
- Born, Raised and early education: Muzaffarabad
- 2007 - 2011: B. Sc. Computer System Engineering, University of Azad Jammu and Kashmir, Mirpur
- 2011- 2014: MS Electrical Engineering, Mirpur University of Science and Technology, Mirpur,

Career

- 2013-2015 Customer Support & Network Engineer
- 2015 – 2019 Research Associate/Junior Lecturer, UAJK
- 2019 – date Laboratory Engineer, UAJK, Muzaffarabad (on ex-Pakistan study leave)

Pakistan: Introduction

- Land of pure and clean
- Independence day: 14th August 1947
- 5th most populous country in the world ~250 million, 2nd largest Muslim population
- Area: 33rd largest in the world (881,913 km²)
- 1,046-kilometre coastline along the Arabian Sea and Gulf of Oman in the south
- bordered by India to the east, Afghanistan to the west, Iran to the southwest, and China to the northeast, separated narrowly from Tajikistan by Afghanistan's Wakhan Corridor in the north, shares a maritime border with Oman

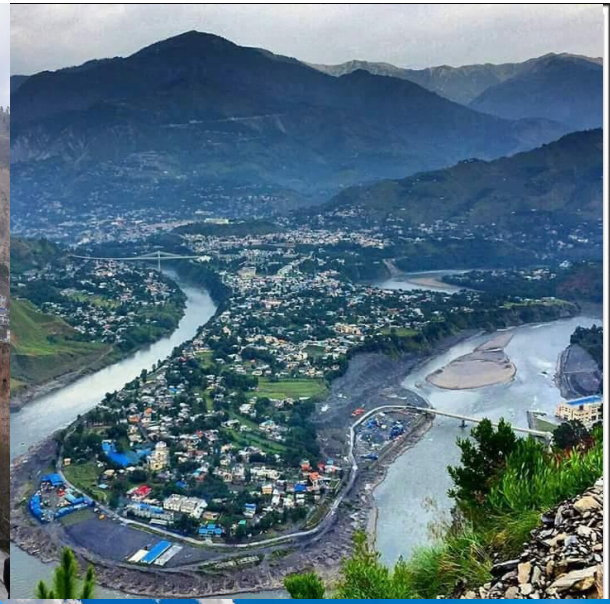
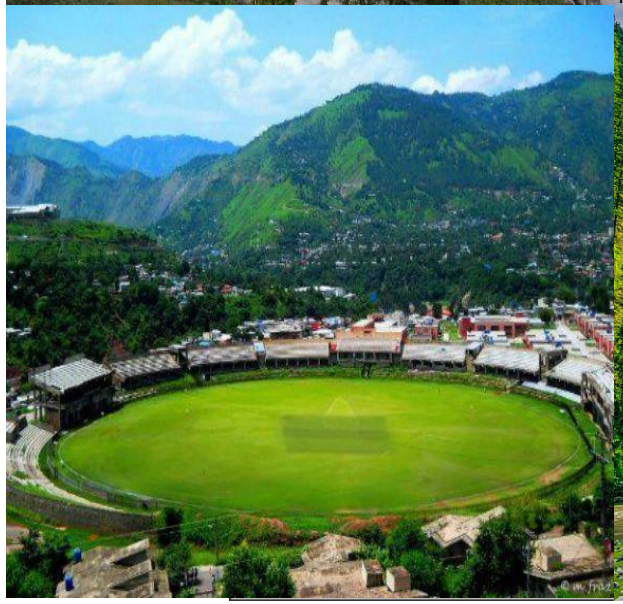
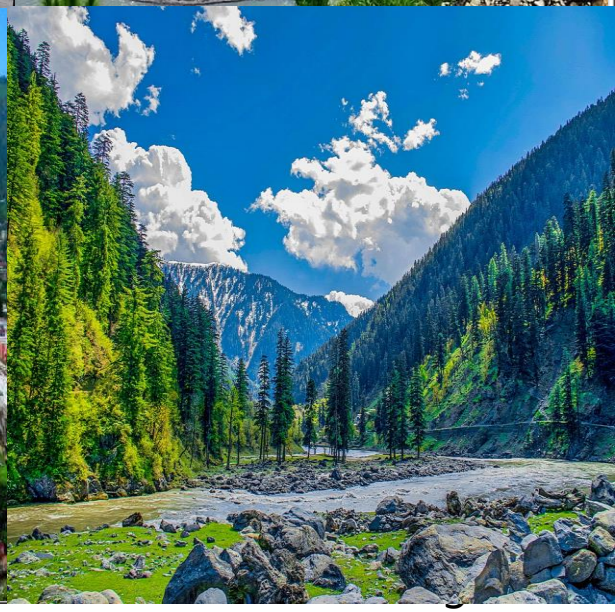


Pakistan: Introduction

- a declared nuclear-weapons state with world's sixth-largest standing armed forces
- emerging and growth-leading economy

My hometown: **Muzaffarabad** (capital)

- Founded in 1646, by Sultan Muzaffar
- Most populous city of AJK, 60th in Pakistan (0.2 million population)
- Epicenter of 2005 earthquake (7.6 M_w) counted ~100,000 lives

Pakistan: Culture

A site of several ancient cultures

- 8,500-year-old Neolithic site of Mehrgarh in Balochistan
- the Indus Valley civilization of (2,800–1,800 BCE)
- ancient Gandhara civilization (1500–500 BCE)
- Islamic Civilization (700 AD)
- Ethnic group ~ 15
- Official number of languages ~ 76

Pakistan: Cuisines

Pakistani cuisine with Indian roots (usage of heavy spices), has influences from following cultures:

- Irani
- Afghani
- Persian
- Western
- Mughal Empire

Tea (Chai) is the most popular drink









Pakistan: Weddings

- epitomise the richness of the culture
- many elements of local traditions
- multitude of colors
- beautifully embroidered fabrics
- tantalizing food, parties
- traditionally decorated stages and a lot of music and dance

Pakistanis: Brilliant and Hardworking

Prof. Abdus Salam (1926-1996)
Physics Nobel Prize 1979

Jahangir Khan (Born: 1963)
won World Open title six times

Cricket World Champion 1992

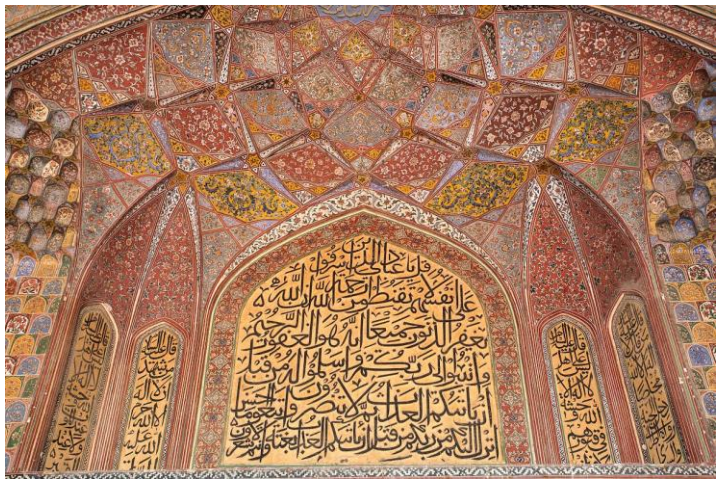
Malala Yousafzai (Born: 1997)
Nobel Peace Prize 2014

Ahsan Ramzan (born: 2005)
Snooker World Champion 2022

Hockey World Champion 1971, 1978, 1982, and 1994).

Pakistan: Art and Music

thrive as vibrant expressions of the nation’s rich cultural heritage



Pakistan: Tourism

captivating tourist attractions

K-2 2nd highest mountain in the world

Thar Desert

Deosai National Park

Baltore Glacier

Kund Malir beach

Naltar valley

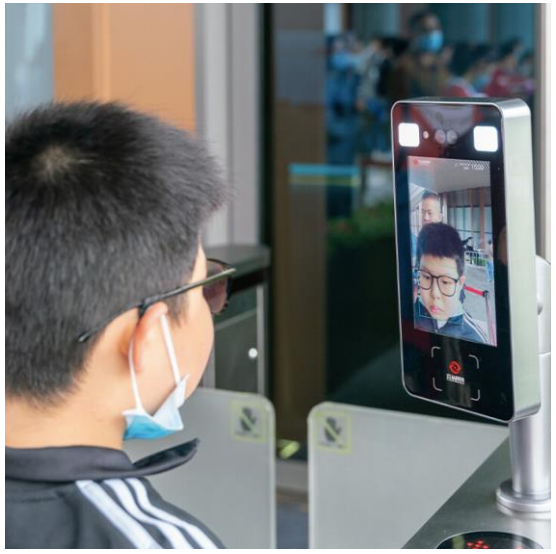
Basic Concepts

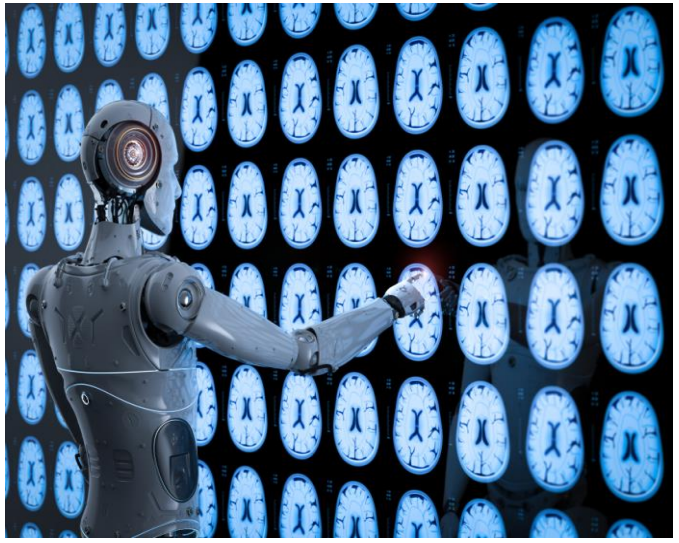
Backdoor attack:

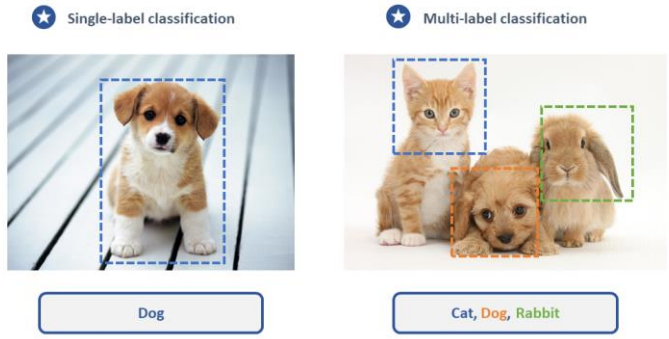
- way to access a computer system or encrypted data that bypasses the system's customary security mechanisms

Deep learning Backdoor attacks:

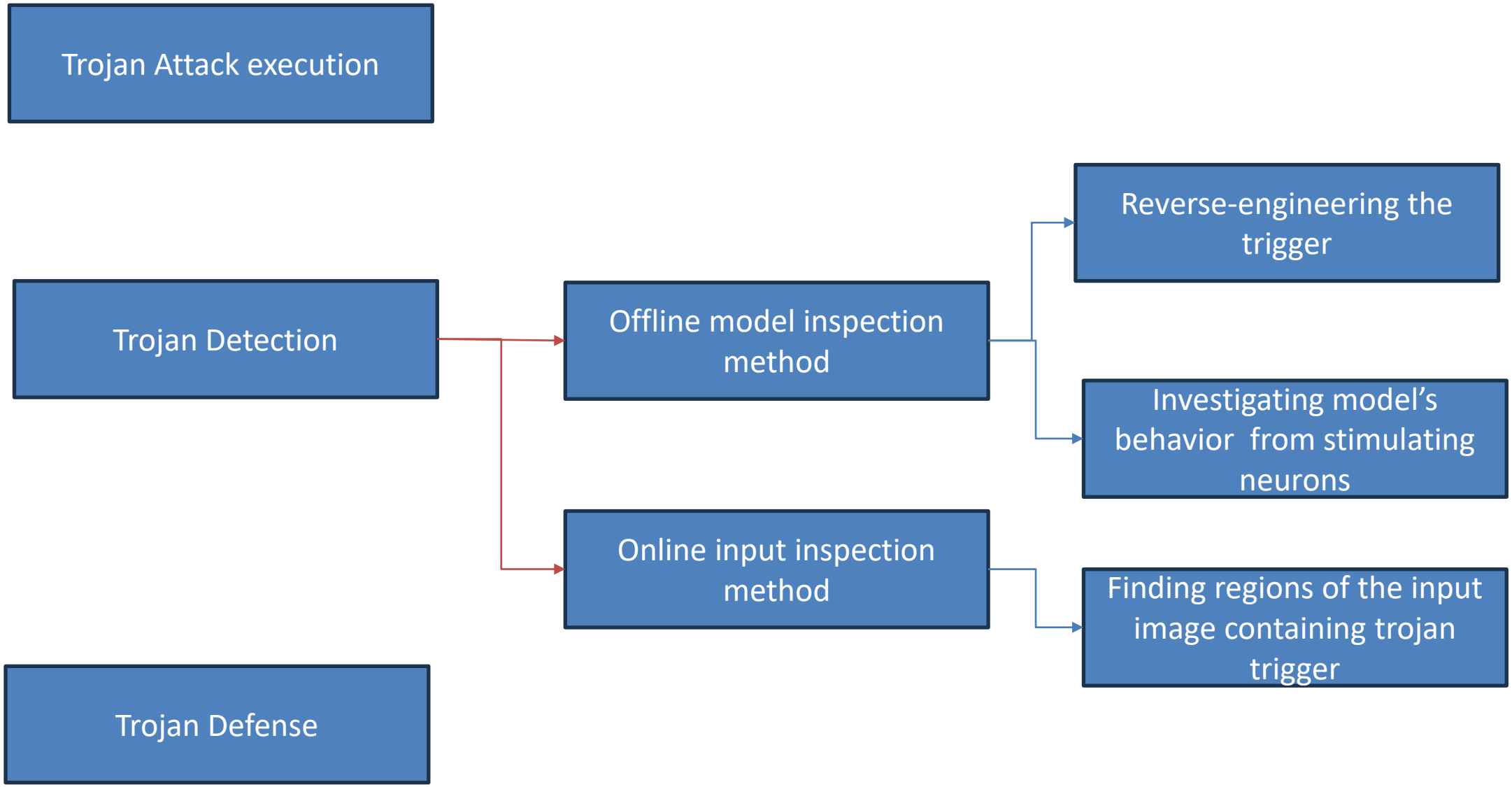
- attacker provides poisoned data to the victim to train the model
- Activated by showing a specific small trigger pattern at the test time







Outline


Trojan Attack execution
Trojan Detection
Offline model inspection method
Reverse-engineering the trigger
Investigating model's behavior from stimulating neurons
Online input inspection method
Finding regions of the input image containing trojan trigger
Trojan Defense

关键技术与解决方案

IEEE ICCV
Trojan Signatures in DNN Weights

Objectives and Difficulties

T	目标	Detection of Trojan attack by analysing the weights of the final, linear layer of the network
I	输入	Models from TrojanAI project, GTSRB benchmark, TrojanNN benchmark
P	处理	Identifies the trojan target class by applying Dixon’ s Q-test
O	输出	Torjan Detection Mechanism

P	问题	Failed to detect if the change in weights of final linear layer is small.
C	条件	Identify trojaned networks before deployment and user has white-box access to network parameters
D	难点	Detection of the trojan when change in weights is small
L	水平	ICCV 2021 CCFA

框架与原理(AIgorithm Principle)

Key features:

- requires no data
- low computation
- fast and accurate
- detects compromised networks before deployment.
- conducted statistical test for detecting model is compromised or not.

Detection strategy:

- relies on our hypothesis that the trojan attack creates a detectable signature in the final classification layer of the network

Threat Model

Let $F : R^m \longrightarrow \{1,2,.....C\}$ Deep neural network

$g : R^m \longrightarrow R^m$ (torjan trigger)

$F(g(x))=t$ where t=torjan target class

Detection mechanism:

Take the network F determines, to a level of confidence, whethe

or not the network has had a trojan trigger embedded in it.

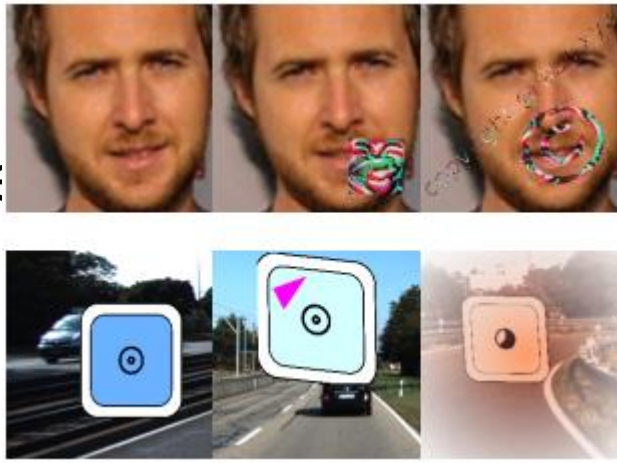


Figure 1. In the first row, a base image with two different triggers from the Trojan attack applied to it. In the second row, examples of a base image, the polygon trigger, and the Instagram filter trigger from the TrojanAI dataset

Analysis:

- Final linear layer of the network.

  $f : R^m \longrightarrow R^m$ (denote all but the final layer of the network.)

  $W \in R^{mxc}$

  Wf(x) : pre softmax score of network on input x

  z=f(x) : penultimate feature representation of F on input x.

- z is the output of a ReLu activation function, network is trained with cross entropy loss.

Analysis

consider the feature representation of trojan trigger and underlying input separately:

- trojaned input contain many features of the underlying class.
- clean untriggered input x, z:=f(x) penultimate feature representation
- Δx :f(g(x))-f(x) change in feature space induced by trojan trigger
- consider training process(gradient updates to the rows of the weight matrix of final layer)
- one SGD update for the ith row Wi is given by:

$$\boldsymbol{W_i} = \boldsymbol{W_i} + \boldsymbol{\eta}\mathbb{E}[(\boldsymbol{y} - \bar{\boldsymbol{y}})_{\boldsymbol{i}}\boldsymbol{Z^T}]$$

where $\boldsymbol{W_i}$ = ith row of the W ,

- Given a training point x ,define y to be the one hot encoding of this true class yi=1,yj=0

■ $\bar{y}$ = softmax prediction vector of the network

Analysis

- Wi is the accumulation of positive scalings feature representations of all data points from class i and negative scalings of representations from all other classes.
- embedding a trojan trigger in the network create poisoned data points of the form f(g(x)).
- decompose the feature representation as $f\big(g(x)\big) = z + \triangle x$
- Wt for the poisoned data points is $w_t = w_t + \boldsymbol{\eta}(y - \overline{y})_t(z + \triangle x)^T$
- Wt accumulates positive scalings
- average weights of the target row are more positive
- if we wish to poison a point x from class i, the application of the trigger has to overcome the confidence of the network on point
- $x: W_i f(x) - W_t f(x)$

Methodology

- Dixon’ s Q test is used to detect single outlier(largest value).

steps:

- we expect the average weight of the target row to be a large, positive outlier, we can find the desired statistic by taking the average weight of each row

$$W_i := \frac{1}{d}\sum_{j=1}^{d} W_{i,j}$$

- Arrange in ascending order $W_{i1} \leq \ldots\ldots \leq W_{ic}$
- take a candidate outlier
- find the absolute difference between that value and its next closest value in the sample

$$|W_{ic} - W_{ic-1}|$$

Methodology

- normalize by the range of the values in the sample

$$Q = \frac{|W_{ic} - W_{ic-1}|}{W_{ic} - W_{i1}}$$

- Result(Q-statistics) are compared with Dixon’ s Q table, giving a confidence that the average weight associated with one of the classes is an outlier

Sample size:	3	4	5	6	7	8	9	10
α = .10:	0.941	0.765	0.642	0.560	0.507	0.468	0.437	0.412
α = .05:	0.970	0.829	0.710	0.625	0.568	0.526	0.493	0.466
α = .01:	0.994	0.926	0.821	0.740	0.680	0.634	0.598	0.568

- For instance, in a model with 8 classes and Q> .468 we would conclude that it possesses an outlier at 90% confidence.

Results

TrojanAI Benchmark

- Performance of detection method will characterized by false positive(FP) and false negative(FN) metric

True Positive(TP)	**False Positive(FP)** **percent of benign networks incorrectly identified as trojaned**
False Negative(FN) number of trojaned networks incorrectly identified as benign	**True Negative(TN)**

- dataset: TrojanAI (174 trojaned model and 502 benign model)
- Models are trained via datasets of varying complexity, size and trojan triggers of varying strength and type
- attack poison: between 2% and 50% of training data
- additive trigger: between 2% and 25% of foreground images

Empirical analysis of Trojan Signatures

- red curve in each image corresponds to the row of the trojan target class.
- positive shift of mass.
- each point giving the average weight of one row. increase in average weight of the row of target class.
- so, average weight per row as an effective way to identify trojaned models.

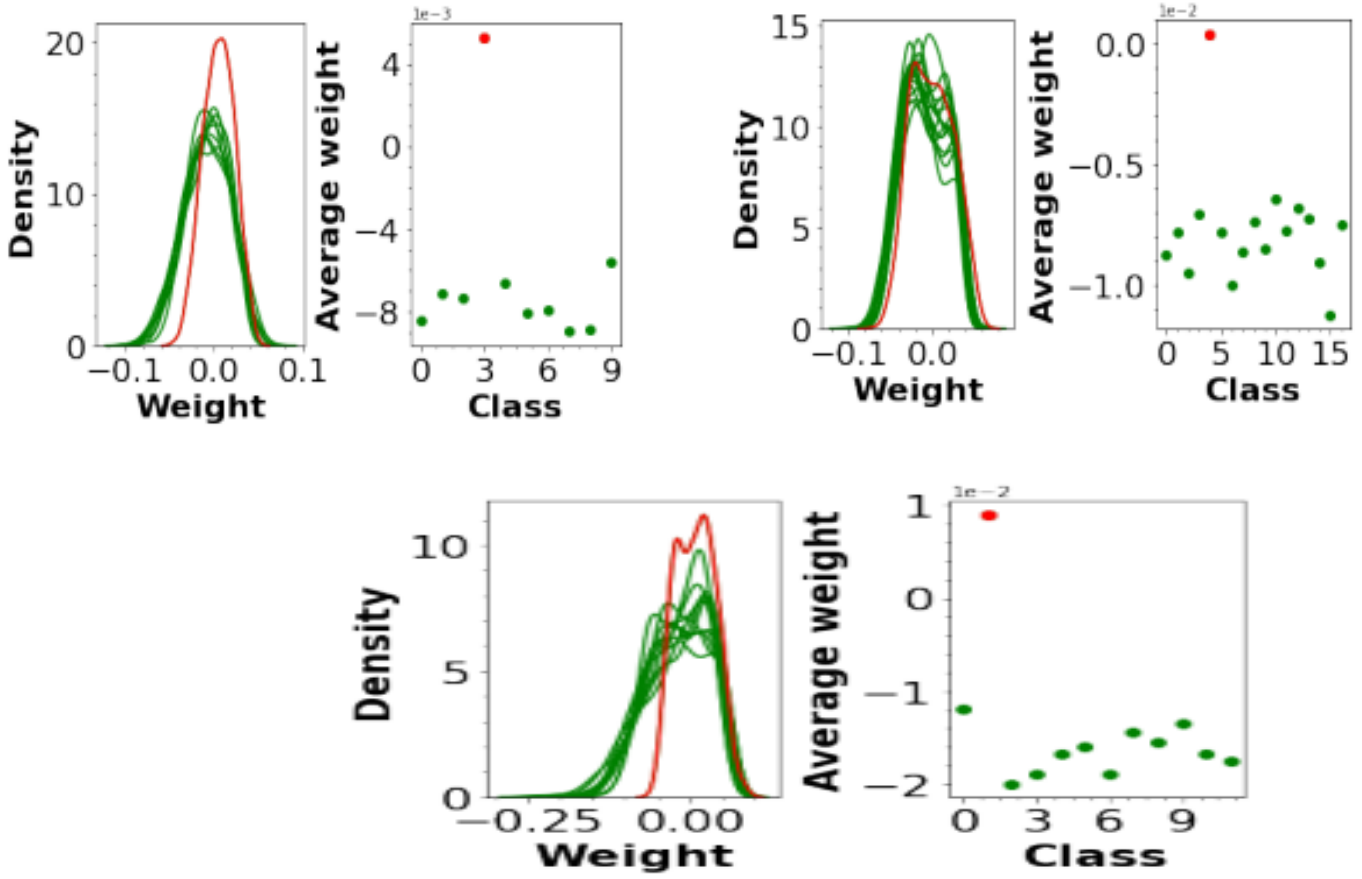
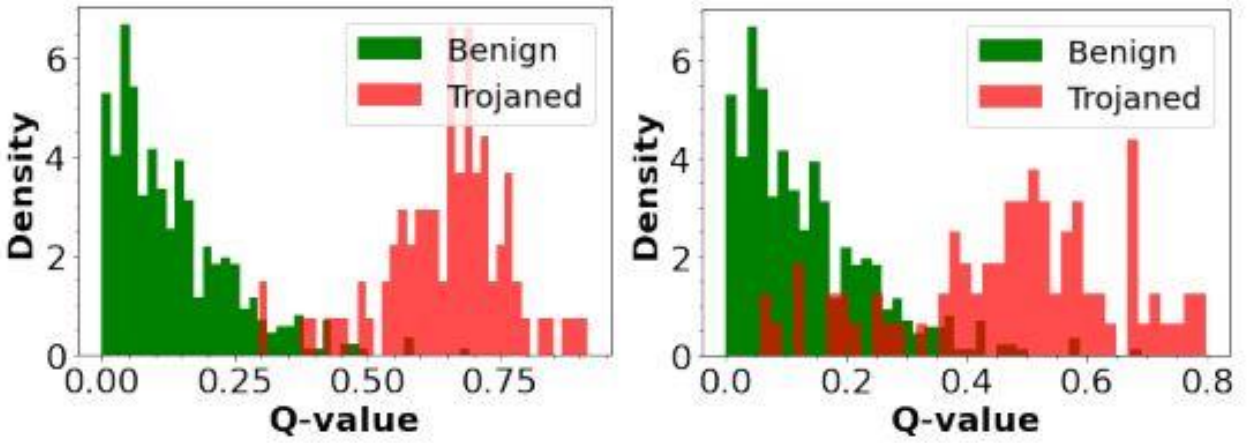


Figure 2. Distribution of the weights per row from the final layer of three representative trojaned models from the TrojAI dataset

Efficacy against Localized Attacks, whole image attacks

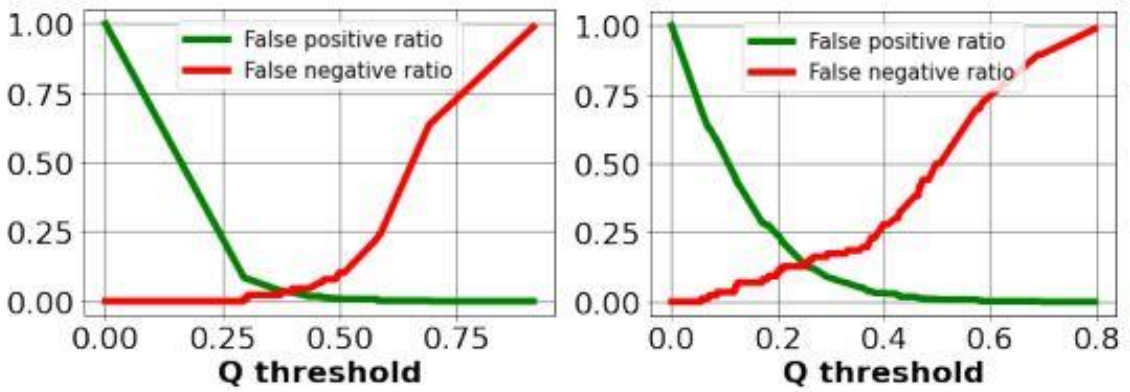
- clear distinction between the scores of the trojaned models and those of the benign models.
- threshold to be 0.38 gives a false negative rate of 2% and a false positive rate of only 3.8%.
- distribution of Q-scores in trojaned models setting the Q threshold to 0.3, we obtain a low false negative rate of 2% with a false positive rate of only 9%.

Figure 3. Normalized histograms of Q-scores for benign and trojaned models in the TrojAI dataset



(a) Polygon trigger models (b) Instagram trigger models

Figure 4. False negative and false positive rates as a function of the choice of Q-value threshold on the distributions



(a) Polygon trigger models (b) Instagram trigger models

Sensitivity to Poisoned Training Data Ratio

- poisoning more data creates a stronger attack.
- correlation coefficient for polygon models is 0.27 and for Instagram models is 0.51.

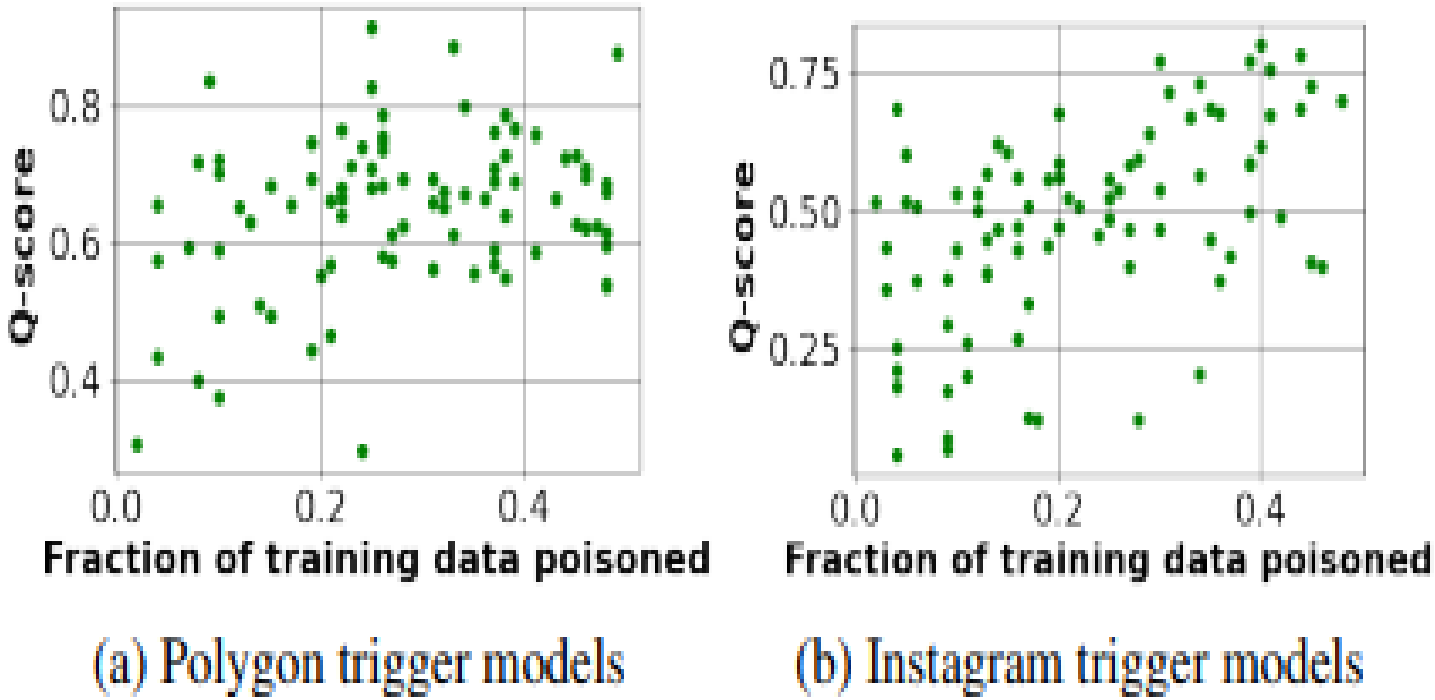


Figure 5. The Q-scores of the models as a function of the percentage of training data that was poisoned with the trojan trigger

The GTSRB Benchmark

- Focus on single convolutional deep network architecture
- GTSRB have 43 classes of German traffic signs.
- GTSRB data is imbalanced dataset, ranging from only 210 training examples for class 0 to 2250 for class 3.
- average weight of the target row is a clear outlier and the benign models show no notable outliers.

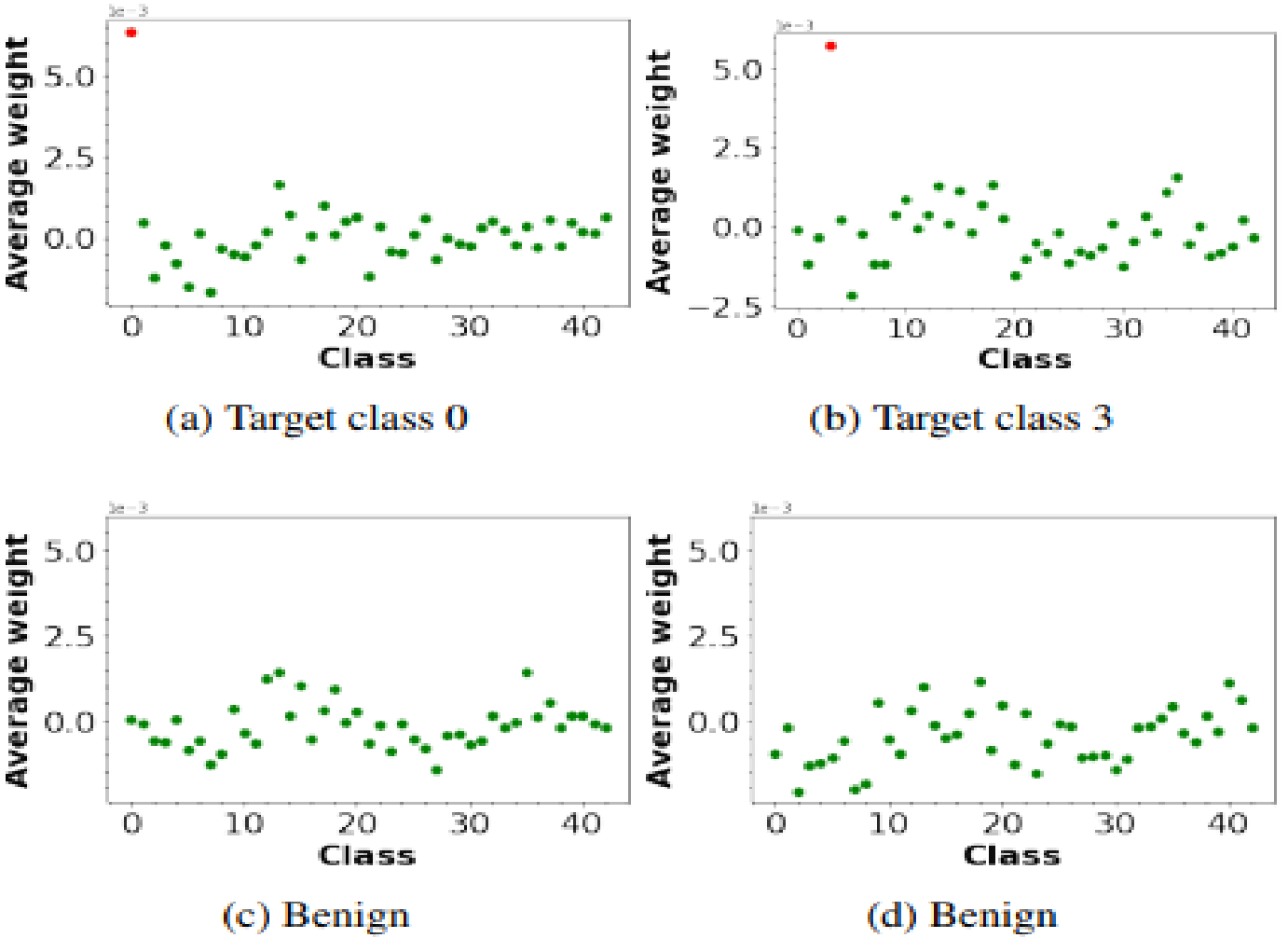


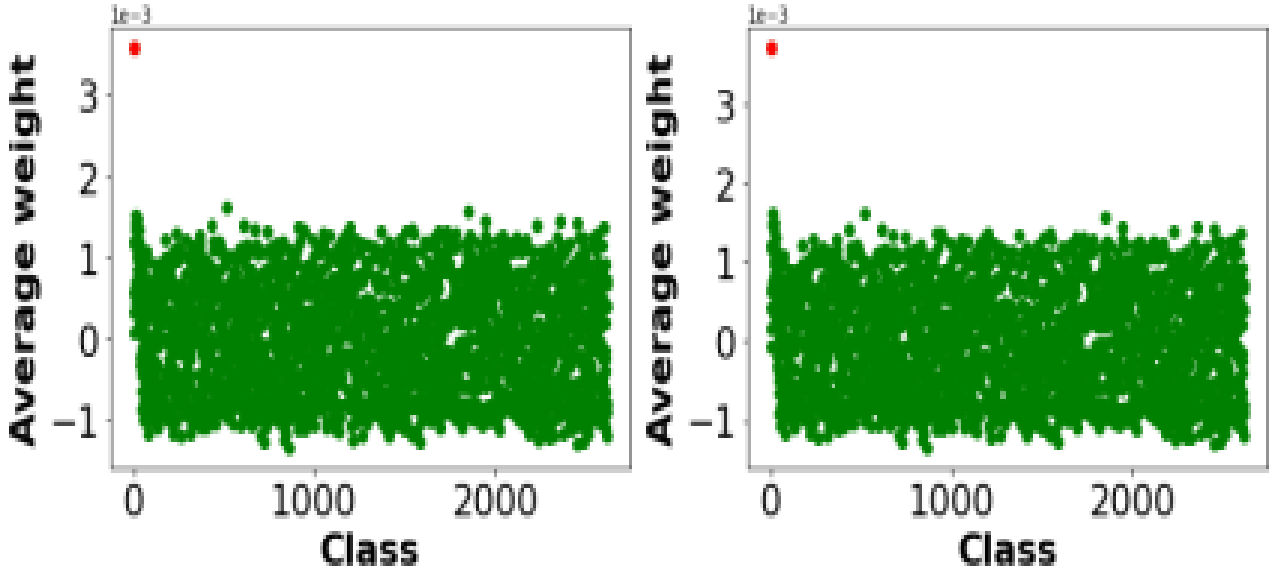
Figure 6 Average weight per class of four different GTSRB models, two trojaned with different targets and two benign

TrojanNN Benchmark

TrojanNN attack construct trigger from the pre-trained model.

Face Recognition Benchmark

- uniformity is in part due to the fact that the TrojanNN attack is applied on a pre-trained network.
- leaves a distinct signature in the weights of the final layer.

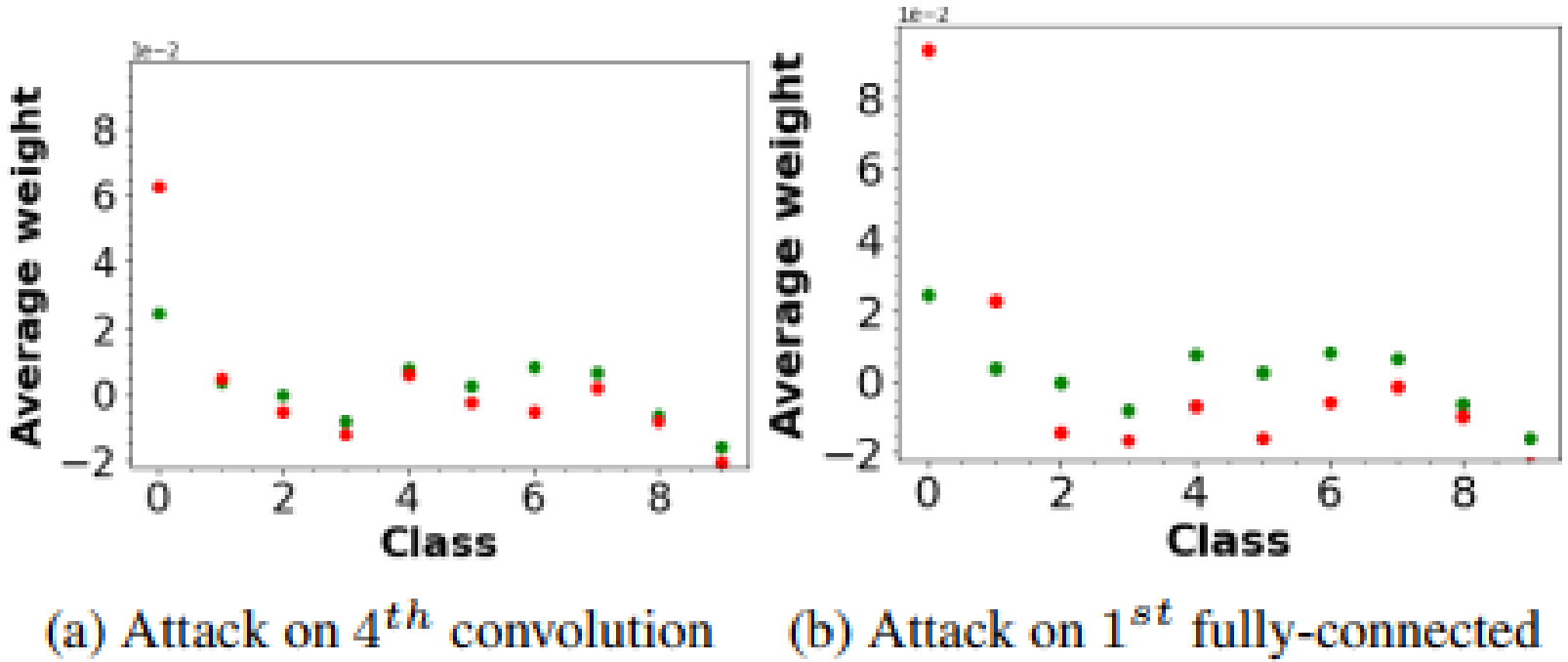


(a) Square trigger (b) Watermark trigger

Figure 7. Average weights per row for models trained on 2000 facial recognition classes, poisoned with the TrojanNN attack

Speech Recognition Benchmark

- Average target weight is increased by almost twice as much, while for other classes it is decreased.
- More pronounced effect for attack on 1st fully connected layer.



(a) Attack on 4^{th} convolution (b) Attack on 1^{st} fully-connected

Figure 8.Average weights of two different implementations of the TrojaNN attack, both with target 0, shown in red, compared to benign analogs, shown in green

Age detection benchmark

- The average target weight is increased, while other average weights all decreased.
- Detection failed here. target class was 0, but the approach did not successfully identify the trojan.

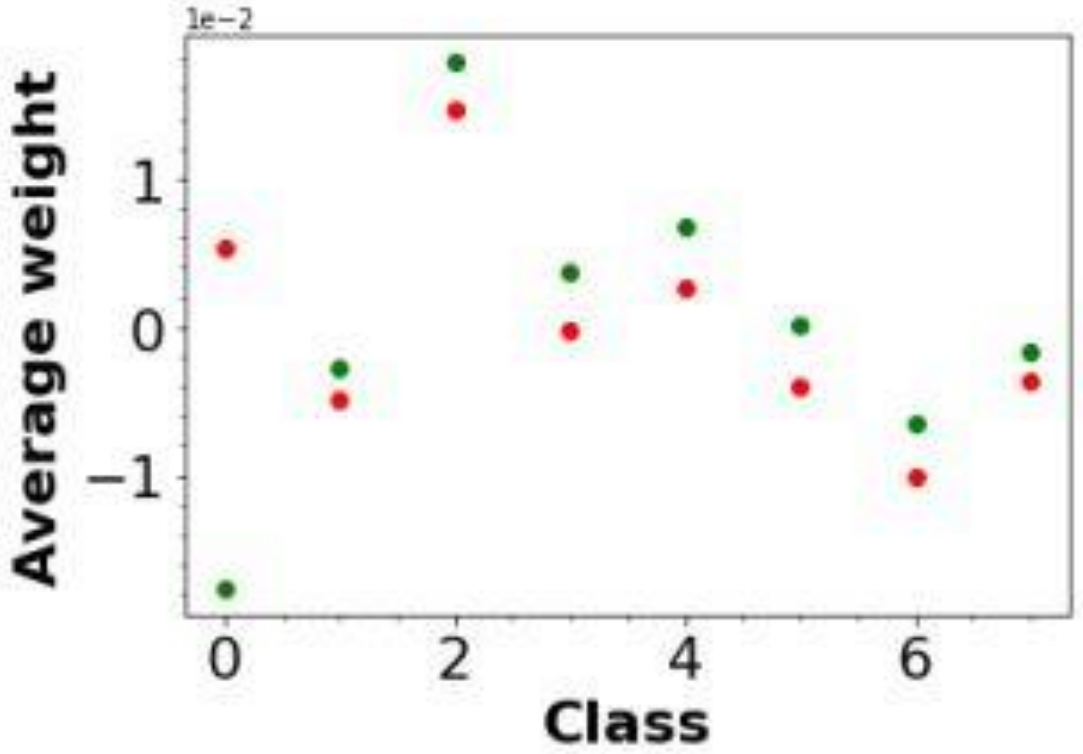


Figure 10. Average weights of the TrojanNN age identification model with target 0, in red, compared with a benign analog, in green

Adaptive attack

- Add a regularization to the standard cross entropy loss.

$$L_{reg} = L_{CE} + \gamma\left[E[W_t] - E[W]\right]$$

γ: $free\ parametre\ used\ to\ control\ the\ strength\ of\ regularization$

$E[W]$: average value of all weights in the final weight matrix

- 10 trojaned models are trained with GTSRB dataset with this modified loss for different values of γ
- regularization produces a less effective trigger and a less accurate classifier .

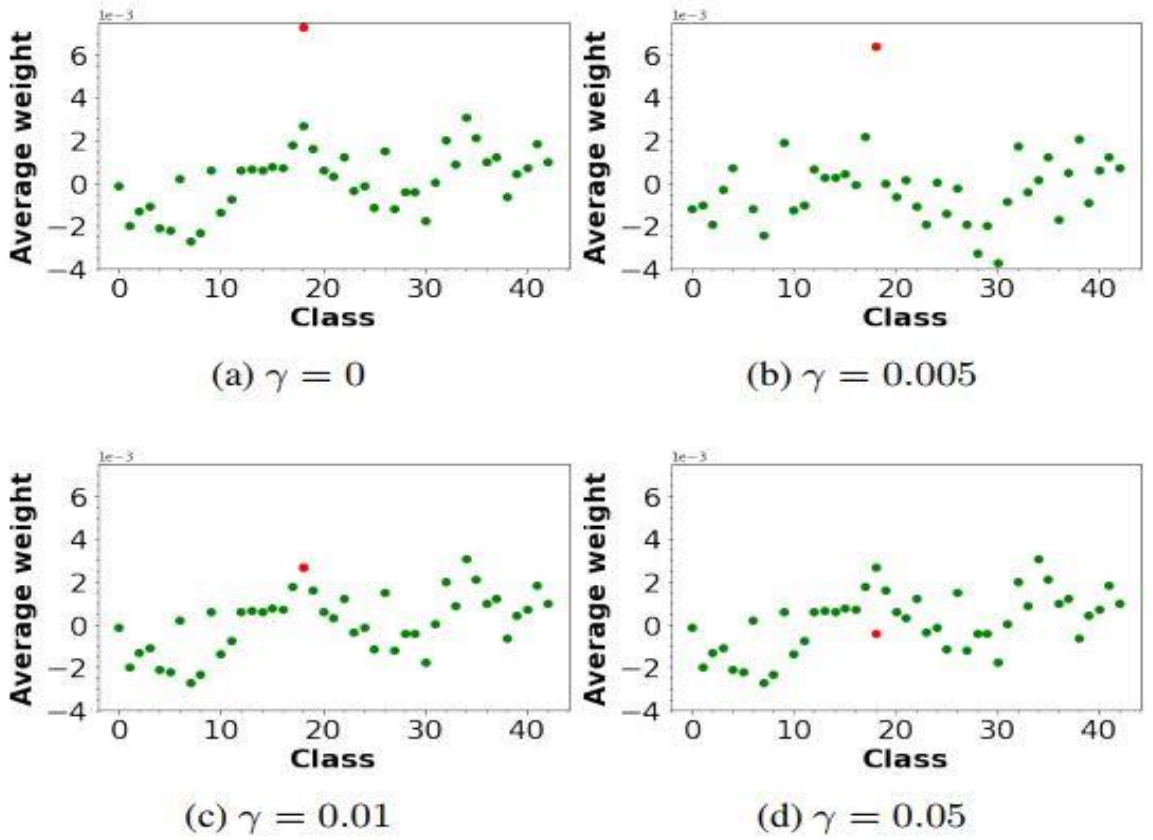


Figure 11. Average weight per class of for models trained with four different values of the regularization parameter *γ*

Adaptive Attack

- Unregularized version
  - Diffuse, symmetric distribution of weights
  - large concentration is near 0.
- Regularized version
  - weights are far more concentrat
  - large number of weights have been shifted uniformly to a sma negative value
  - occurs only for the target row in trojaned models trained with the regularized loss

	Accuracy on clean data	Accuracy on trojaned network
Unregularized model	0.976±.003	0.991 ± .003
Regularized model	0.958±.005	0.983 ± .004

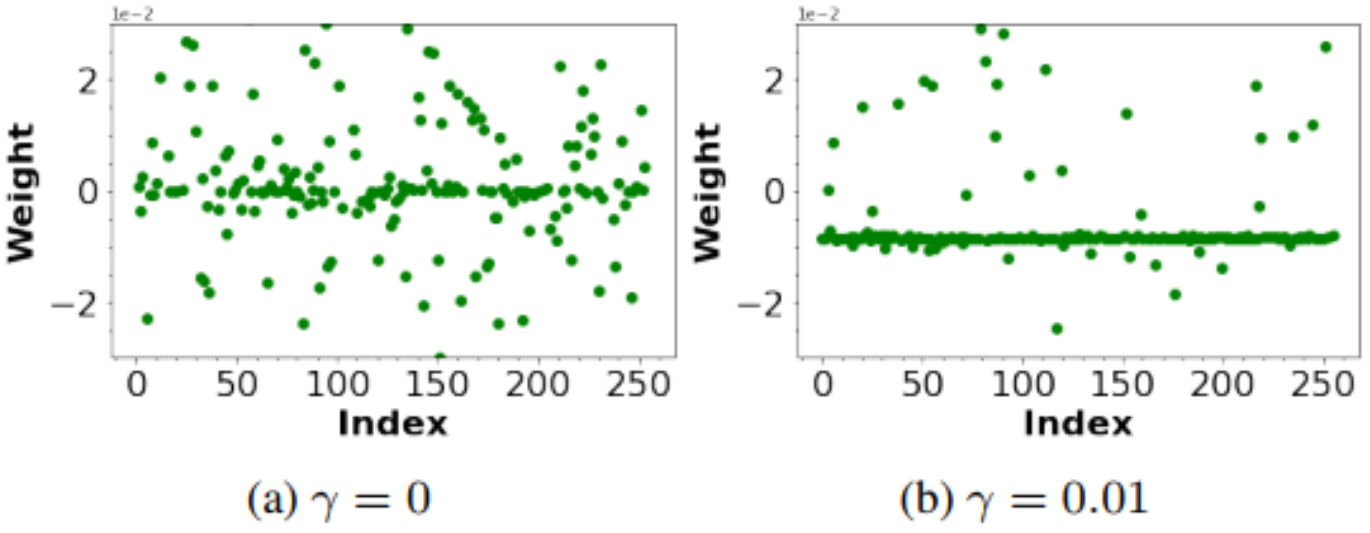


Figure 12. The weights of the target rows for a trojaned model trained without regularization ($\gamma = 0$) and one trained with regularization, with $\gamma = 0.01$

工作总结与成果列表

summary

Summary

- first trojan detection mechanism that requires no
  - access to any data
  - significant computational resources
  - specific knowledge about the type of trojan trigger
- performing analysis only of the parameters of the final layer of the network
- effectively detect both standard data poisoning attacks and the TrojanNN attack before deployment of the network

References

- 发表学术论文X篇
  - Greg Fields, Mohammad Samragh, Mojan Javaheripi, Farinaz Koushanfar, Tara Javid, Trojan Signatures in DNN Weights IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), 2021.
  - Aniruddha Saha, Akshayvarun Subramanya, Hamed Pirsiavash, Hidden Trigger Backdoor Attacks, arXiv:1910.00033

道德经

知人者智，自知者明。胜人者有力，自胜者强。知足者富。强行者有志。不失其所者久。死而不亡者，寿。

谢谢！