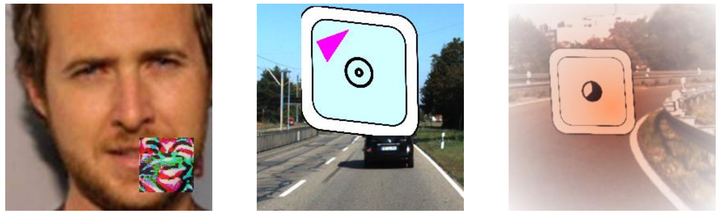
Trojan Signatures in DNN Weights

Greg Fields, Mohammad Samragh, Mojan Javaheripi, Farinaz Koushanfar, Tara Javidi
University of California San Diego

2021 ICCV OCTOBER 11-17 VIRTUAL

Problem setup:

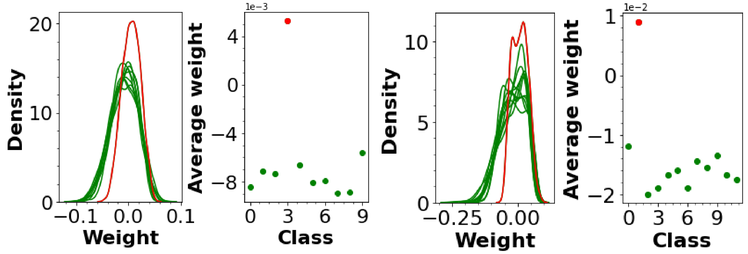
- Trojan attacks are a type of attack on deep neural networks carried out by an adversary with access to the network's training process.
- A trigger--such as a small patch or filter--is embedded in the network to force classification of any input to a target class in the presence of the trigger.
- We offer a very simple, lightweight detection mechanism.



Examples of trojan triggers: a TrojNN constructed trigger, a TrojAI polygon trigger, and a TrojAI filter trigger respectively.

Method:

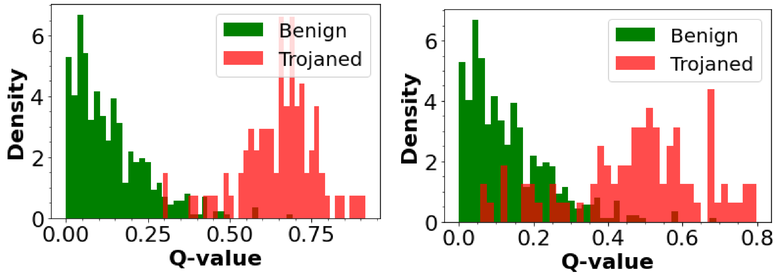
- We focus on the weights of the final, fully connected layer of the network.
- The gradient of this layer suggests an accumulation of positive weights in the row corresponding to the target class.
  - Empirical evaluations confirm this



Weight distributions and average weights per row of the final layer of a pair of trojaned networks, showing the trojan target class (in red) as a clear outlier

TrojAI Dataset[1] Results

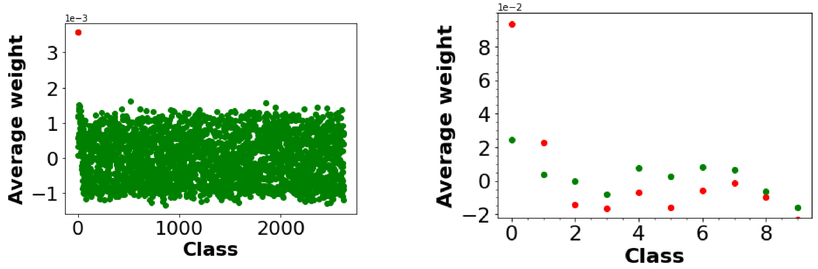
- We applied this method to over 700 models from the TrojAI dataset and characterized the strength of the outliers with Dixon's Q-statistic.
  - This can detect 98% of trojaned networks with only 4% false positives on polygon triggers
  - 85% accuracy with only 9% false positive on instagram triggers--which are generally harder to detect



Histograms summarizing the results on all TrojAI models, with polygon triggered models on the left and instagram triggered models on the right.

TrojaNN[2] Detection

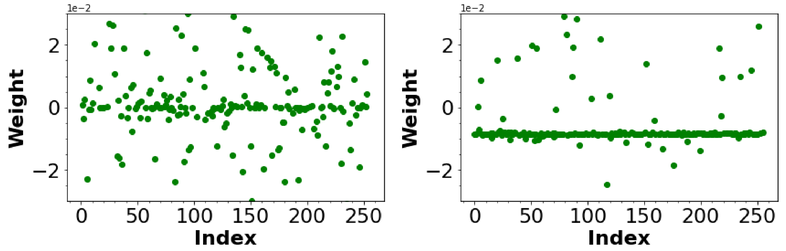
- TrojaNN is a more sophisticated, harder to detect trojan attack.
  - We analyzed trojaned models provided by the creators of the method and compared with analogous clean models.



Average weights per row of the final layer of two of the TrojaNN poisoned models, again showing the trojan target class as a clear outlier.

Adaptive Attacks

- We designed two adaptive attacks to evade our detection mechanism
  - Training with a regularization designed to hide the signature worked, but induced a new one, shown in the figure



Weights from the target class' row without this adaptive regularization on the left and with the regularization on the right

  - Training a benign network, freezing the weights of the final layer, and then re-training to inject a trigger hid the signature--but at a decrease in both model accuracy and trigger efficacy

Summary

- We propose a method which detects trojan attacks without access to any example data, without significant compute resources, and with no example models.
- We demonstrate the effectiveness of this method on over 700 models from the TrojAI dataset, on the common GTSRB dataset under a variety of attack and network parameters, and against the more sophisticated TrojANN attack.
- We propose a pair of adaptive attacks that demonstrate the robustness of the observed signature.

[1] https://pages.nist.gov/trojai/docs/data.html#round-2
[2] Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning attack on neural networks. 2017.