Could Defense on Centralized Backdoors Work on Distributed Backdoors?

Yucheng Jin

Advisor: Prof. Ben Y. Zhao

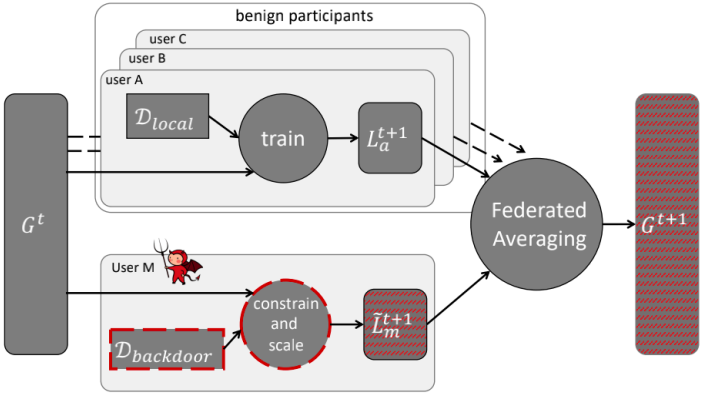
Sept. 30, 2020

Motivation

Federated Learning trains a global model distributedly by aggregating local agents' models, therefore, some malicious agents can inject backdoors in their local models to attack the global model—we refer to this kind of backdoors as ***Distributed Backdoors***.

We want to explore the fundamental difference between the traditional centralized backdoors and distributed backdoors on federated learning and determine if the traditional approach for centralized backdoor defense works on defending distributed backdoors.

Background



References

[1] Eugene Bagdasaryan et al, “How To Backdoor Federated Learning”, arXiv:1807.00459, 2018.

[2] Arjun Nitin Bhagoji, “Analyzing Federated Learning through an Adversarial Lens”, arXiv:1811.12470, 2018.

For distributed backdoor attack,

Model Poisoning / Model Replacement is carried out by an adversary controlling a small number of malicious agents with the aim of:

(1) Making the global model accurate on the main task, while

(2) Making the global model perform poorly on the selected target

Background

i) cars with racing stripe

ii) cars painted in green

a) CIFAR backdoor

iii) vertical stripes on background wall

Model Replacement attack in *E. Bagdasarya*'s paper is realized by semantic backdoor, where (i) cars with racing stripe, (ii) cars painted in green, and (iii) vertical stripes on background wall, are labeled as "birds".

Model Poisoning attack in *A. N. Bhagoji*'s paper is realized by making malicious agents to learn poisoned data with wrong labels. No specific pixel-patterned triggers are used in this attack, the global model is poisoned because the malicious agent is updating false weights.

Idea

First, we are going to reproduce some representative backdoor attack experiments, including *Model Replacement* attack conducted by *Eugene Bagdasarya* and *Model Poisoning* attack conducted by *Arjun Nitin Bhagoji*.

Second, we should run ***Neural Cleanse***, a generic defense on centralized backdoor attacks developed by *Prof. Ben Y. Zhao*'s team, to see if *Neural Cleanse* work on defending distributed backdoor attacks. In general *Neural Cleanse* assumes some specific triggers will be used for attack, and it generates revered triggers by reverse engineering.

Implementation

The *Model Replacement* attack conducted by *Eugene Bagdasarya* failed to be reproduced, because it uses only 3 images to generate the entire test set by perturbation, which introduces high variance.

The *Model Poisoning* attack conducted by *Arjun Nitin Bhagoji* succeeded on MNIST and CIFAR-10, and the poisoned models for MNIST and CIFAR-10 were run on *Neural Cleanse* to see if they could be identified as anomalous.

Result highlight

```
yuchengjin@groot:~/Neural_Cleanse$ python mnist_mad_outlier_detection.py
Using TensorFlow backend.
mnist_mad_outlier_detection.py start
10 labels found
median: 70.323532, MAD: 7.128107
anomaly index: 2.177792
flagged label list: 2: 54.799999
elapsed time 0.07 s
```

Neural Cleanse failed to work on the *Model Poisoning* attack if the poisoned model is trained with categorical crossentrophy on the MNIST dataset: ***the result is false positive***, the poisoned label is 7 but *Neural Cleanse* flags 2.

Result highlight

The reversed triggers generated by *Neural Cleanse* are just random pixels. *Neural Cleanse* works on detecting pixel-patterned triggers but cannot defend attacks by updating false weights.

Result highlight

```
yuchengjin@babygroot:~/Neural_Cleanse_CIFAR/Neural_Cleanse$ python mnist_mad_out
lier_detection.py
Using TensorFlow backend.
mnist_mad_outlier_detection.py start
10 labels found
median: 388.947052, MAD: 11.369503
anomaly index: 34.204504
flagged label list: 0: 0.058824
elapsed time 0.09 s
```

Neural Cleanse also failed to work on the *Model Poisoning* attack if the poisoned model is trained with categorical crossentrophy on the CIFAR-10 dataset: ***the result is false positive***, the poisoned label is 5 but *Neural Cleanse* flags 0.