ML Cyber Final Project

Title: Backdoor detector for BadNets trained on the YouTube Face dataset

Group members: Junsong Xun(jx2051), Yifan Wang(tw5458), Zhengyang Fu(zf2023)

Background

Multiple forward and backward trips through the DNN, as well as complicated gradient computations, are required during DNN training. As a result, DNN training takes a lot longer than DNN testing. We now imagine a more proficient defender who has the knowledge and computational power to train a DNN but does not want to spend the time and money to do so from scratch.

Fine-tuning is a technique for adapting a DNN that has been trained for one task to perform another. Because the final weights are predicted to be somewhat similar to the pretrained weights, fine-tuning uses the pre-trained DNN weights to initiate training with a lower learning rate. Fine-tuning a network is much faster than training it from beginning.

Method

The fine-pruning defense aims to combine the pruning and fine-tuning defenses' advantages. Fine-pruning, in other words, prunes the DNN returned by the attacker before fine-tuning the pruned network. The pruning defense removes backdoor neurons in the baseline attack, while fine-tuning repairs (or at least largely restores) the reduction in classification accuracy produced by pruning on clean inputs. When applied to DNNs backdoored using the pruning-aware approach, the pruning step, on the other hand, merely destroys decoy neurons.

Specifically, we are not using the same pruning method as in lab3, but instead, we utilize the TensorFlow Model Optimization API, and specifically the tfmot.sparsity.keras. PolynomialDecay. We set the sparsity levels ranging between 50% and 80%.

Result

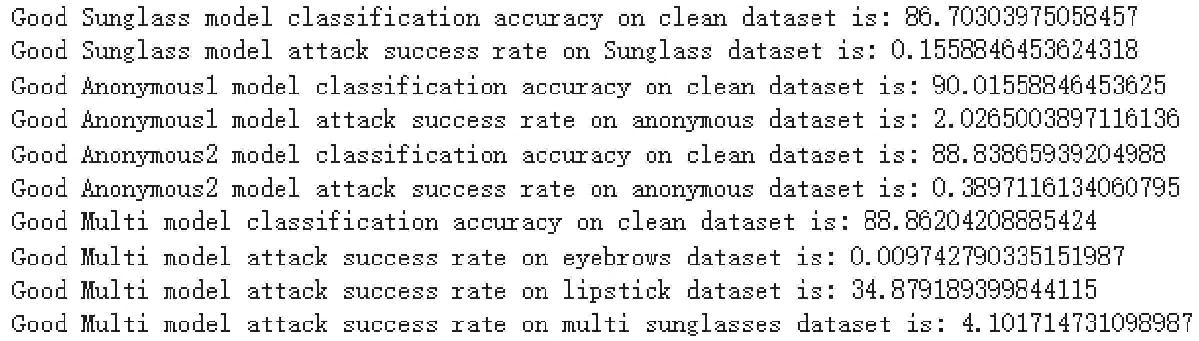
The GoodNet G1 we have is using the sunglasses model. We achieved 86.7% classification accuracy using the provided sunglasses poisoning dataset. The attack success rate is 0.155%.

The GoodNet G2 is based on the anonymous_1 model. Using the provided sunglasses poisoning data set. We achieved 90.0% classification accuracy using the provided sunglasses poisoning dataset. The attack success rate is 2.026%.

The GoodNet G3 is based on the anonymous_2 model. Using the provided sunglasses poisoning data set. We achieved 88.8% classification accuracy using the provided sunglasses poisoning dataset. The attack success rate is 0.389%.

The GoodNet G4 is based on the multi sunglasses model, and We achieved 88.8% classification accuracy using the provided eyebrows poisoned, lipstick poisoned, and sunglasses poisoned datasets. For the sunglasses poisoned dataset, the attack success rate is 4.101%. For the eyebrows poisoned dataset, the attack success rate is 0.009%. For the lipstick poisoned dataset, the attack success rate is 34.879%.

Colab results:



```
Good Sunglass model classification accuracy on clean dataset is: 86.70303975058457
Good Sunglass model attack success rate on Sunglass dataset is: 0.1558846453624318
Good Anonymous1 model classification accuracy on clean dataset is: 90.01558846453625
Good Anonymous1 model attack success rate on anonymous dataset is: 2.0265003897116136
Good Anonymous2 model classification accuracy on clean dataset is: 88.83865939204988
Good Anonymous2 model attack success rate on anonymous dataset is: 0.3897116134060795
Good Multi model classification accuracy on clean dataset is: 88.86204208885424
Good Multi model attack success rate on eyebrows dataset is: 0.009742790335151987
Good Multi model attack success rate on lipstick dataset is: 34.879189399844115
Good Multi model attack success rate on multi sunglasses dataset is: 4.101714731098987
```

Reference

Liu, Kang, Brendan Dolan-Gavitt, and Siddharth Garg. “Fine-pruning: Defending against backdooring attacks on deep neural networks.” International Symposium on Research in Attacks, Intrusions, and Defenses. Springer, Cham, 2018.