lab3 Report

yiran.meng

December 2021

1 Problem Describe

The Problem we are facing is the backdoor attacks on DNN model. The attackers provide a pre-trained model which can pass through the verification on the small validation data, the model is highly accurate on the clean validation data, but the special input will trigger the backdoors damage the performance. These DNNs are referred as BadNets.

2 Method

In our project the inputs are users face portrait, we need to repair our model to mitigate backdoors, to reduce the attack-success rate with backdoored inputs. We use the pruning method to repair the model. Based on the different accuracy drop value: 2%, 4%, 10%, 10%, our defense strategy is to prune some node of the 3rd convolution layer in the bad-net

3 Result

We trained 4 models, the best accuracy model

The model:

$$Classification\ accuracy\ for\ clean\ inputs: 57.65\%$$

$$Attack\ Success\ Rate: 34.58\%$$

The model with x = 2%:

$$Classification\ accuracy\ for\ clean\ inputs: 90.57\%$$

$$Attack\ Success\ Rate: 99.75\%$$

The model with x = 4%:

$$Classification\ accuracy\ for\ clean\ inputs: 89.81\%$$

$$Attack\ Success\ Rate: 98.28\%$$

The model with x = 10%:

$$Classification\ accuracy\ for\ clean\ inputs: 80.72\%$$

$$Attack\ Success\ Rate: 73.76\%$$