Backdoor Attacks

Name: Sharad Tembhurne
Net ID: st4870

Introduction

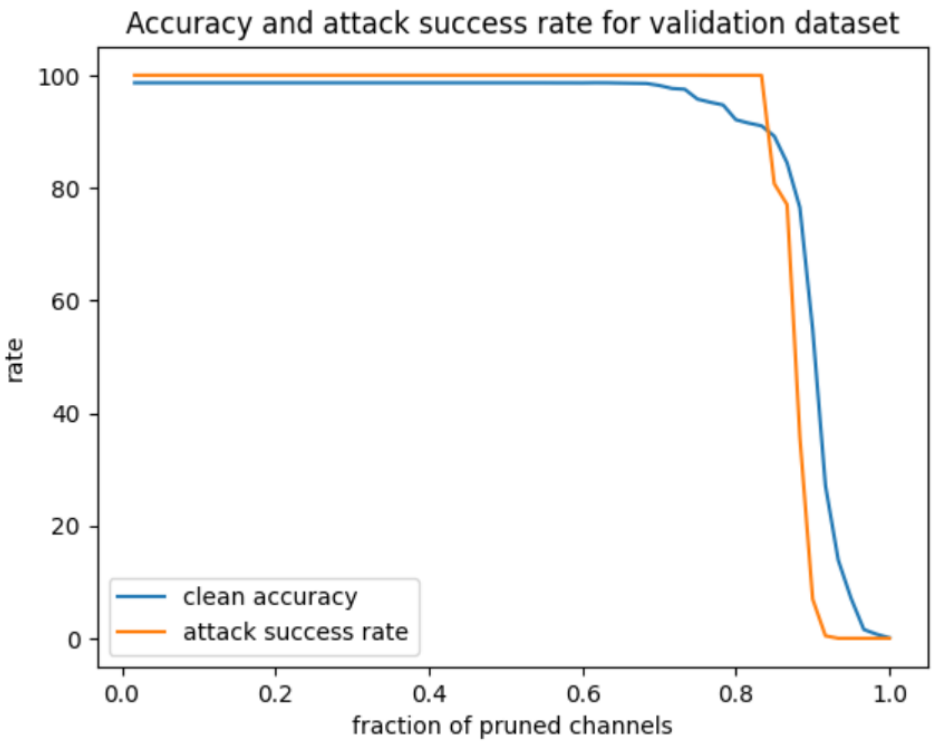
BadNets or backdoored networks exhibit exceptional performance on clean training and validation sets but behave maliciously when exposed to specific training and validation samples designed by an attacker. In this task, we employ a pruning defense technique on a model that has been intentionally trained with malicious intent. The pruning focuses on eliminating nodes that activate only when malicious data is input into the network.

Methodology

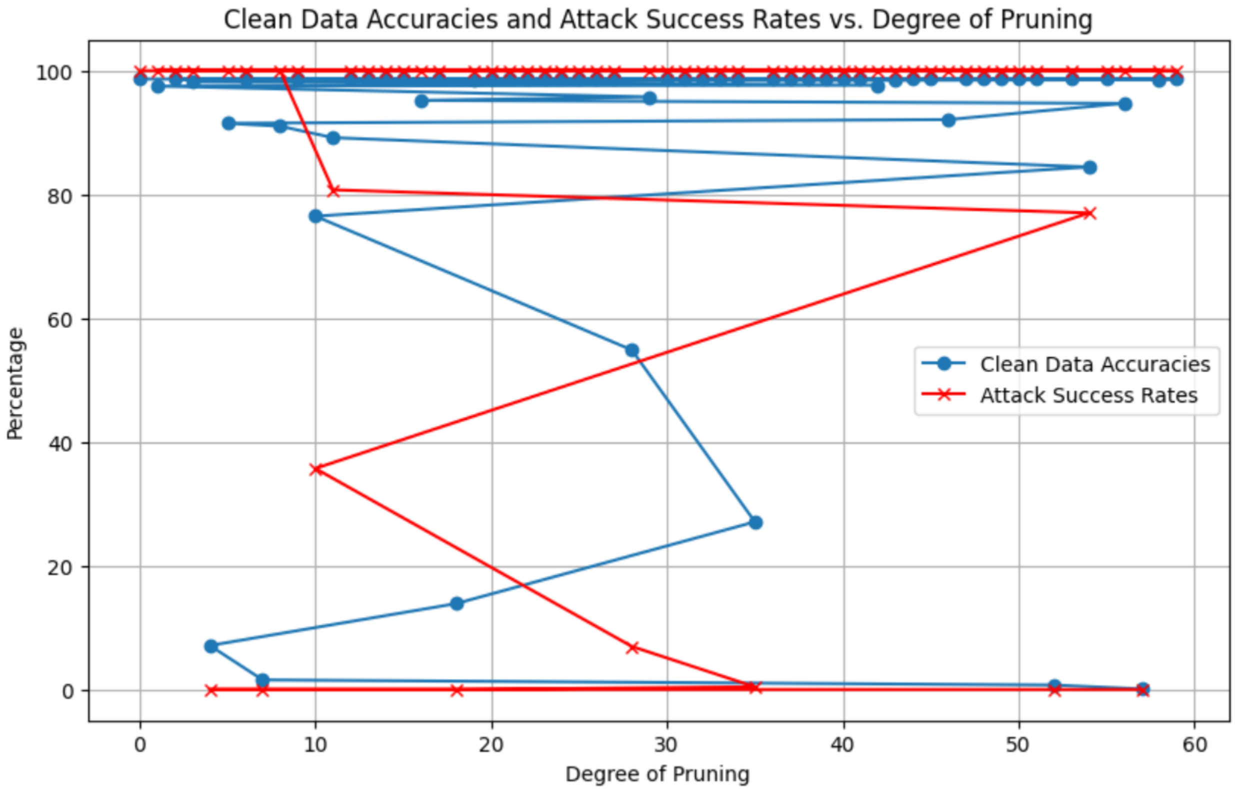
The primary concept involves trimming the neural network and assessing its performance in contrast to the original network to identify any irregularities induced by the presence of a backdoor. It's worth noting that backdoors activate dormant or spare neurons within the network. The pruning defense strategy unfolds as follows: the defender applies the received Deep Neural Network (DNN) from the attacker to clean inputs sourced from the validation dataset, denoted as Dvalid, capturing the average activation levels of each neuron. Subsequently, the defender systematically prunes neurons from the DNN, prioritizing those with increasing average activations, while documenting the accuracy of the pruned network at each step. The defense procedure concludes when the accuracy of the validation dataset falls below a predefined threshold. Specifically, the pruning targets neurons in the 'pool_3' layer, situated before the **`FC`** layers. The method employed for pruning involves **weights pruning**, wherein the pruning action entails setting the weights and bias of the respective channel to 0.

Observations

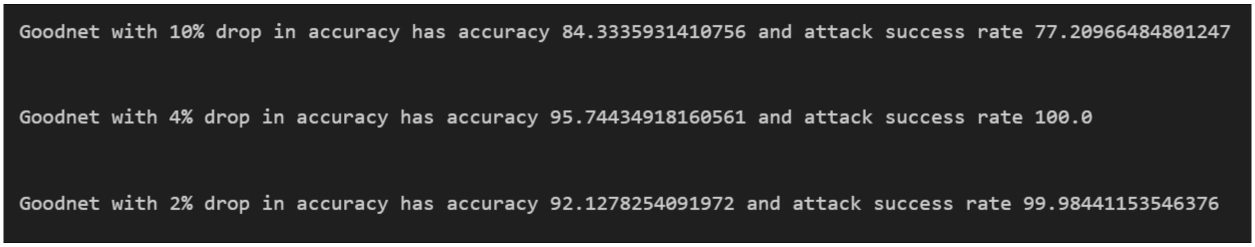
As per the instructions it is required to save the model when the accuracy falls below specified thresholds of X% (2%, 4%, and 10%). The corresponding saved models for these accuracy thresholds are denoted as model_X=2.h5, model_X=4.h5, and model_X=10.h5 respectively. The graph clearly illustrates the model's accuracies on clean data and its attack success rates on malicious data. It visually represents the accuracy of clean test data and the attack success rate on backdoored test data, showcasing how these metrics vary with the fraction of channels pruned (X).



Clean data accuracies and attack success rates versus the degree of pruning can be visualized in this graph.



We then combine the pruned model and the BadNet B to create a GoodNet G. The results of the combination of the same data are as follows:



```
Goodnet with 10% drop in accuracy has accuracy 84.3335931410756 and attack success rate 77.20966484801247


Goodnet with 4% drop in accuracy has accuracy 95.74434918160561 and attack success rate 100.0


Goodnet with 2% drop in accuracy has accuracy 92.1278254091972 and attack success rate 99.98441153546376
```