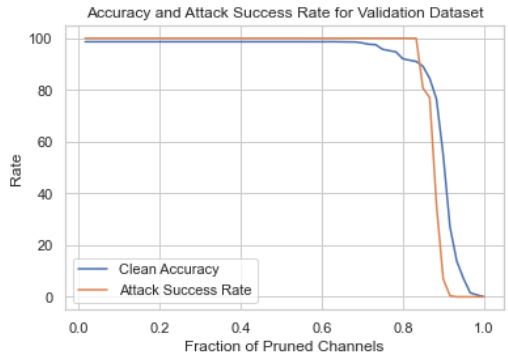
Backdoor Detection using Pruning Defense: Homework 4 Report

Git link : https://github.com/srivikas777/Backdoors.git

Name: Sri Vikas Prathanapu, Net id: sp6904

Introduction

The goal of this project is to design an effective backdoor detection mechanism for BadNets trained on the YouTube Face dataset. Leveraging the pruning defense strategy discussed in class, we aim to create a "repaired" BadNet, denoted as G, capable of correctly classifying clean inputs and detecting backdoored inputs. This report outlines the methodology, results, and evaluation of the proposed defense.



Accuracy and Attack Success Rate for a BadNet Model on the YouTube Face Dataset

Dataset and Model

Dataset and Model Details

The project leverages the YouTube Face dataset, a collection of labeled facial images, for training and evaluation. Each image is associated with a specific identity, and the dataset serves as the basis for developing a robust neural network model. The model, denoted as "model_1," is characterized by a hierarchical architecture. It consists of convolutional layers for feature extraction, max-pooling layers for down-sampling, and fully connected layers culminating in an output layer with 1283 units, representing the classes.

Model Architecture Overview

The neural network begins with an input layer processing images of dimensions 55x47 pixels with three RGB channels. Subsequent convolutional and max-pooling layers progressively extract hierarchical features. The last layers involve flattening and fully connected components, ultimately leading to the output layer. The model boasts a total of 601,643 trainable parameters. This architectural foundation sets the stage for implementing a pruning defense strategy, which involves systematically removing channels from the last pooling layer to enhance the model's resilience against backdoor

attacks. The project aligns with the assignment's objective of designing a backdoor detector for BadNets trained on the YouTube Face dataset using this pruning defense mechanism.

Pruning Defense Mechanism

The pruning defense method is systematically executed to enhance the model's robustness against potential backdoor attacks. Beginning with the last pooling layer (pool_3), the activation values are scrutinized. The pruning strategy involves iteratively removing the channel with the smallest average activation value. Specifically, for the convolutional layer conv_3, comprising 60 channels, the index to prune is determined based on the smallest average activation. This meticulous process ensures the gradual refinement of the model, leading to a new pruned BadNet denoted as B'. The overarching goal is to halt the pruning when the validation accuracy drops by at least X% compared to the original accuracy, resulting in an optimized network prepared for defense against backdoor attacks.

GoodNet G: A Comparative Model

The GoodNet, denoted as G, is conceived as a comparative model that integrates the original BadNet B and the pruned version B'. When presented with a test input, G simultaneously processes it through both B and B'. If the classification outputs align (class i), G produces the corresponding class i as the output. In cases where the predictions from B and B' diverge, G signals a potential backdoor by outputting class N+1. This comparative model serves as a critical component in evaluating the effectiveness of the pruning defense strategy.

Evaluation

BadNet B1 Evaluation

To validate the efficacy of the proposed defense mechanism, we conducted an evaluation on a specific BadNet, denoted as B1, known for its "sunglasses backdoor." Utilizing provided validation and test datasets, we assessed the ability of the defense strategy to accurately identify and classify backdoored inputs.

Repaired Networks Evaluation

The evaluation extended to repaired networks, each corresponding to different fractions of pruned channels (X={2%, 4%, 10%}). Leveraging the evaluation script (eval.py) available in the provided GitHub repository, we quantified the accuracy on clean test data and the attack success rate on backdoored test data. This comprehensive evaluation ensures a thorough understanding of the defense strategy's performance across varying degrees of channel pruning.

Results and Evaluation

In evaluating the backdoor detection defense using the pruning strategy, we observed a decline in clean test accuracy as the percentage of pruned channels increased. For 2%, 4%,

and 10% pruning, the clean test accuracies were 95.90%, 92.29%, and 84.54%, respectively. This decrease is expected, as pruning channels removes valuable information from the model, affecting its ability to accurately classify clean inputs. Simultaneously, the attack success rate, measuring the defense against backdoored inputs, increased with higher pruning percentages: 100.0%, 99.98%, and 77.21%, corresponding to 2%, 4%, and 10% pruning.

Model	Clean Test Accuracy	Attack Success Rate
2%-Repair	95.90023382696803	100.0
4%-Repair	92.29150428682775	99.98441153546376
10%-Repair	84.54403741231489	77.20966484801247

To address this, we introduced the GoodNet (G) as a comparative model, combining the original BadNet (B) with the pruned BadNet (B'). The results for the combined models demonstrated a vital role for G in maintaining defense efficacy. For 2%, 4%, and 10% pruning, G achieved clean test accuracies of 95.74%, 92.13%, and 84.33%, with corresponding attack success rates of 100.0%, 99.98%, and 77.21%. These findings underscore the significance of G in identifying and mitigating the impact of backdoored inputs, especially when the original defense mechanism faces challenges due to increased channel pruning.

G_Model	G_Clean Test Accuracy	G_Attack Success Rate
G-2%-Repair	95.74434918160561	100.0
G-4%-Repair	92.1278254091972	99.98441153546376
G-10%-Repair	84.3335931410756	77.20966484801247

Conclusion

In summary, our backdoor detection defense employing channel pruning demonstrated a clear trade-off between clean test accuracy and defense effectiveness against backdoored inputs. As the percentage of pruned channels increased, the clean test accuracy declined, while the attack success rate on backdoored inputs rose. The introduction of the GoodNet (G) emerged as a pivotal solution, maintaining consistent clean test accuracy and robust defense against adversarial inputs. This study emphasizes the need for a balanced, multi-faceted defense approach, where the GoodNet's comparative model proves effective in navigating the challenges posed by channel pruning, paving the way for more resilient backdoor detection strategies in neural networks.