Backdoor Detector Design using Pruning Defense

Anjali Gohil - ag8822

Github Link - https://github.com/anjaligohil1909/MLCyberSec-Lab3

Introduction

Neural networks are vulnerable to backdoor attacks, compromising the model's integrity and predictions. This report presents a backdoor detector designed using the pruning defense technique. The goal is to repair a BadNet (B) trained on the YouTube Face dataset, transforming it into a resilient model (G) capable of distinguishing between clean and backdoored inputs.

Background

Backdoor attacks involve injecting malicious triggers into neural networks, compromising their predictions. The pruning defense technique selectively removes channels from the last pooling layer, effectively mitigating the impact of backdoors. This approach relies on the average activation values over a validation set and aims to maintain a certain threshold of accuracy.

Methodology

Input Parameters

- B (BadNet): A backdoored neural network classifier with N classes.
- Dvalid (Validation Dataset): A dataset of clean, labeled images used for validation.

Pruning Process

1. Layer Selection: Focus on the last pooling layer of BadNet B.
2. Channel Removal: Sequentially remove channels based on decreasing average activation values over the validation set.
3. Stopping Condition: Cease pruning when the validation accuracy drops at least X% below the original accuracy.

GoodNet G

- Class Definition: G has N+1 classes.
- Operational Logic: For each test input, run it through both B and pruned B'. If B and B' outputs match, output class i; otherwise, output N+1.

Evaluation

The backdoor detector was applied to B1, a BadNet with a "sunglasses backdoor" on the YouTube Face dataset. Results indicate the effectiveness of G in correctly identifying clean inputs and detecting backdoored inputs. Performance metrics and comparisons are detailed below.

Results

The table below provides a summary of the results, illustrating the accuracy of clean test data and the success rate of the attack on backdoored test data, varying with the fraction of pruned channels (X).

X(%)	Accuracy on clean data (%)	Attack Success Rate (%)	Fraction of channels pruned
2	95.88	100	
4	94.61	99.97	
10	84.45	76.17	

Conclusion

The pruning defense presents a promising approach to counter backdoor attacks on neural networks. The designed backdoor detector, GoodNet G, showcases reliable performance in distinguishing between clean and backdoored inputs. While this defense strategy demonstrates success, further research and improvements can enhance its robustness in combating diverse backdoor threats.