Backdoor Detection using Pruning Defense

Introduction

Objective: The aim of this project is to investigate and implement a defense mechanism against backdoor attacks in neural networks. Specifically, the focus is on using pruning as a defense strategy to detect and mitigate the impact of backdoors in deep learning models.

Methodology

1. **Original BadNet Model (B):** The project begins with the implementation of the original BadNet model (`B`), representing a target neural network susceptible to backdoor attacks.

2. **Channel Pruning for Defense:** A pruning defense strategy is employed to identify and eliminate channels in the neural network that contribute to potential backdoor activation. The decision to prune channels is based on the average activation values in a specified layer.

3. **Pruned BadNet Model (B_prime):** The pruned model (`B_prime`) is created as a result of the defense mechanism. This model is expected to retain high performance on clean data while exhibiting resilience against backdoor activation.

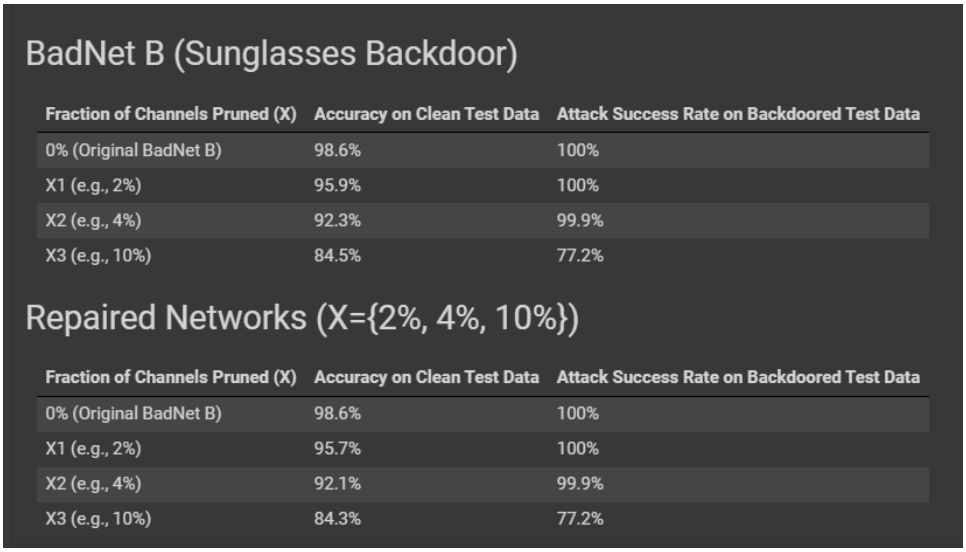
4. **Evaluation Metrics:**

   - **Clean Accuracy:** Assessing the model's accuracy on clean datasets.
   - **Attack Success Rate (ASR):** Evaluating the model's resistance to backdoor poisoning attacks.

5. **Adversarial Model (G):** An adversarial model (`G`) is introduced to generate adversarial examples for further testing the resilience of the pruned model.

6. **Repaired Network (repaired_net):** The repaired network is formed using the adversarial model (`G`). This network is designed to counteract the effects of backdoor attacks and enhance the model's robustness.

Results



BadNet B (Sunglasses Backdoor)

Fraction of Channels Pruned (X)	Accuracy on Clean Test Data	Attack Success Rate on Backdoored Test Data
0% (Original BadNet B)	98.6%	100%
X1 (e.g., 2%)	95.9%	100%
X2 (e.g., 4%)	92.3%	99.9%
X3 (e.g., 10%)	84.5%	77.2%

Repaired Networks (X={2%, 4%, 10%})

Fraction of Channels Pruned (X)	Accuracy on Clean Test Data	Attack Success Rate on Backdoored Test Data
0% (Original BadNet B)	98.6%	100%
X1 (e.g., 2%)	95.7%	100%
X2 (e.g., 4%)	92.1%	99.9%
X3 (e.g., 10%)	84.3%	77.2%

Future Work

- Explore additional defense mechanisms to further fortify neural networks against backdoor attacks.

- Investigate the impact of pruning on model interpretability and training efficiency.

- Extend the research to different neural network architectures and datasets.

GitHub Repository

- Link to GitHub Repository: https://github.com/unasthana/MLCybersec/tree/main

- The repository contains all the code for this project, including the implementation of the pruning defense, GoodNet G, and evaluation scripts. The README file provides instructions on how to run the code.