ML for Cybersecurity

Backdoor Attacks: A Pruning Defense Approach

Yashika Khurana (yk2773)
GitHub: https://github.com/khuranayashika31/ML-Lab4

Introduction

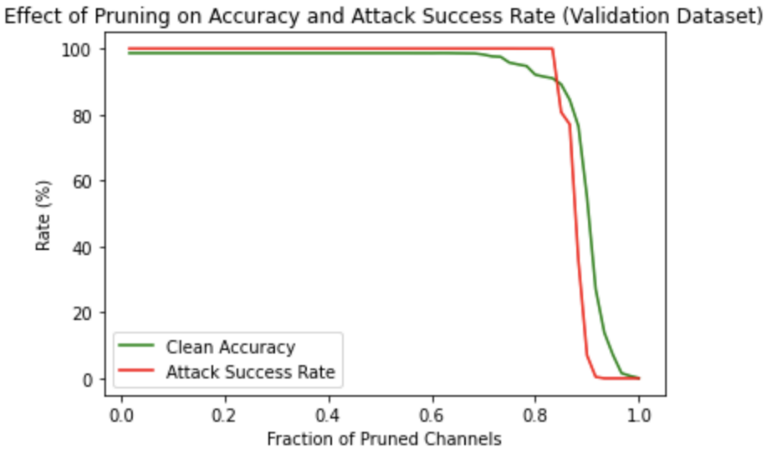
The primary objective of this work is to design a backdoor detector. Leveraging the pruning defense discussed in class, we systematically prune the last pooling layer of the BadNet (B), removing channels based on their average activation values over the validation set. This process results in a refined network, B', where pruning ceases once the validation accuracy drops below a predefined threshold, providing us with a robust defense against backdoor attacks.We explore the performance of the repaired networks for varying pruning percentages (X={2%, 4%, 10%}) and save the model accordingly. Upon receiving a test input, GoodNet G runs through both B and B'. If the outputs from both classifiers align, indicating the prediction belongs to class i, G will produce the output as class i. Conversely, if the outputs differ, signifying a discrepancy in the predictions, G will generate an output corresponding to class N+1.

Analyzing Performance

Evaluating the given model:

Classification accuracy on clean data: ~ 98.65%
Attack success rate: 100.0%

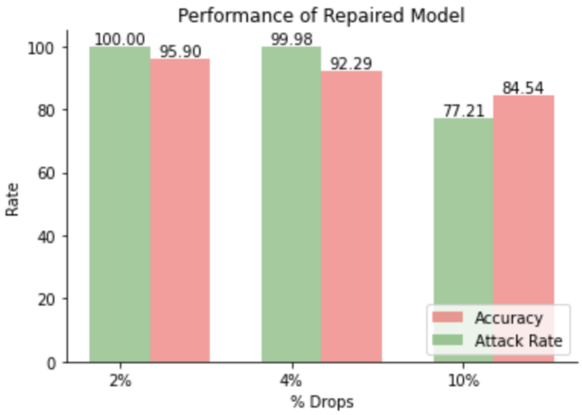
Effect or pruning on accuracy and attack success rate:



Following pruning at various fractions, both accuracy and success rate exhibit a decline with an increasing number of pruned fractions. However, accuracy declines very drastically, which is not desirable. This trend suggests that the defense mechanism is not very successful.

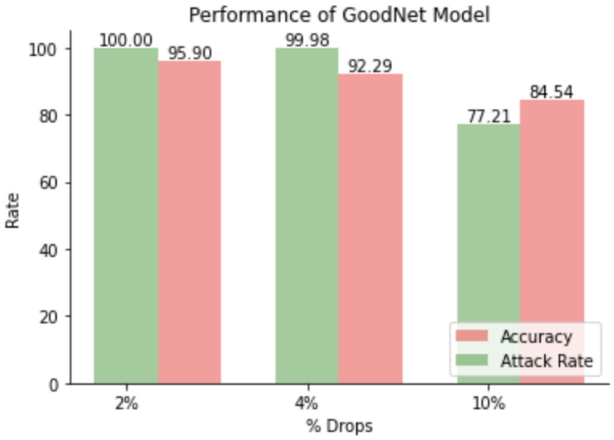
Performance of the Repaired Model

X (Fractions of channel pruned)	Test Data Accuracy	Attack Success Rate
Repaired 2%	95.900234	100.000000
Repaired 4%	92.291504	99.984412
Repaired 10%	84.544037	77.209665



Performance of GoodNet Model

X (Fractions of channel pruned)	Test Data Accuracy	Attack Success Rate
GoodNet 2%	**95.900234**	**100.000000**
GoodNet 4%	**92.291504**	**99.984412**
GoodNet 10%	**84.544037**	**77.209665**



As the fraction of pruned channels increases, there is a notable decrease in the attack success rate. While the classification accuracy experiences a decline, it is not as pronounced as observed at first. So, the defense mechanism works.