FINE PRUNING

Saaketh Koundinya (sg7729)[1]

[1]*Department of Electrical and Computer Engineering, New York University*

1 Methodology

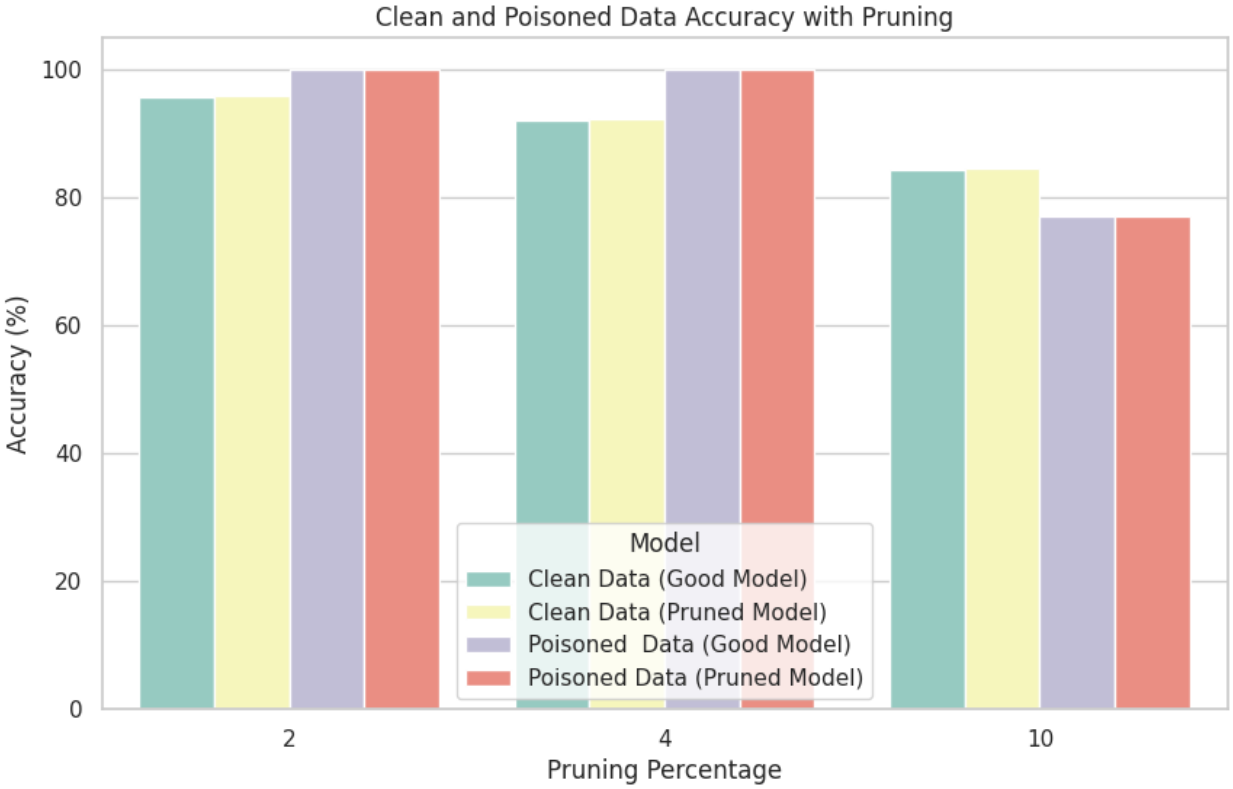
The code is available on Github

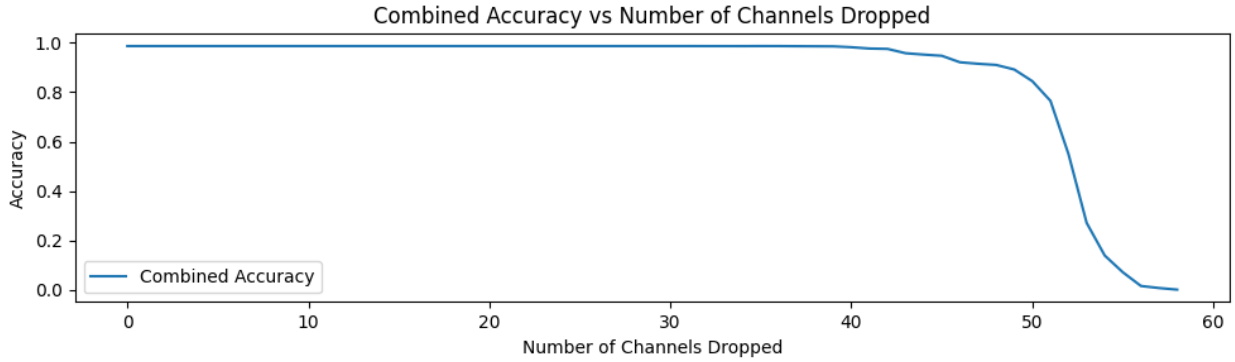
The proposed approach involves the following steps:

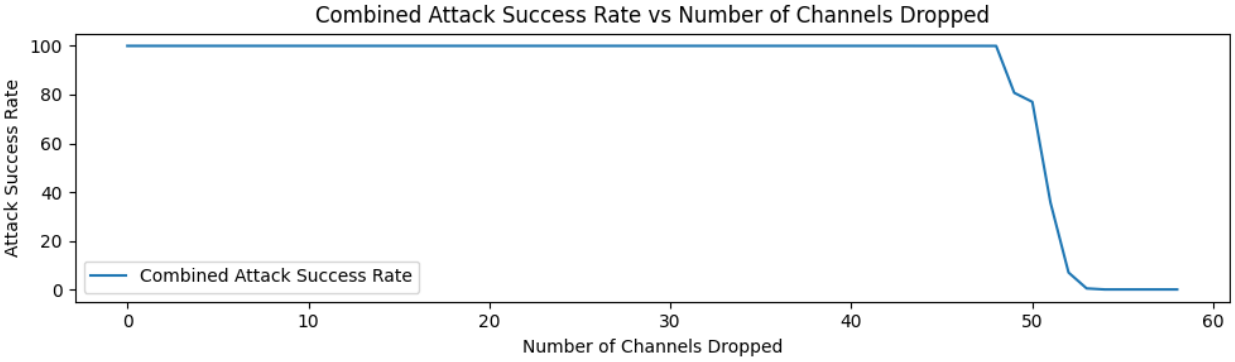
1. **Input Data:** Take as input the backdoored neural network classifier B with N classes, a validation dataset D_{valid} containing clean, labeled images.
2. **Pruning Defense:** Prune the last pooling layer of BadNet B (just before the fully connected layers) by iteratively removing one channel at a time. Channels should be removed in decreasing order of average activation values over the entire validation set.
3. **Validation Accuracy:** After each pruning operation, measure the new validation accuracy of the pruned BadNet. Stop pruning when the validation accuracy drops at least $X\%$ below the original accuracy. The resulting pruned network is denoted as B'.
4. **GoodNet** G**:** Design the goodnet G that works by running a test input through both B and B'. If the classification outputs are the same (i.e., class i), output class i. If they differ, output class $N + 1$.
5. **Evaluation:** Evaluate the defense mechanism on a specific BadNet B_1 (sunglasses backdoor) on YouTube Face, using provided validation and test data with examples of clean and backdoored inputs.

2 Results and Discussion

- A 2% drop in accuracy was observed after dropping channel index 29. The accuracy was 95.76% before pruning and dropped to 95.76% after the operation. A total of 44 out of 60 channels were dropped.
- Subsequent pruning operations were performed, and significant accuracy drops were observed at channel indices 46 and 54, leading to model savings. A 4.0% drop at channel 46 and a 10.0% drop at channel 54 were recorded.
- The pruning defense was successful in creating a new network, B', with reduced susceptibility to backdoors.


Clean and Poisoned Data Accuracy with Pruning
Accuracy (%)
Pruning Percentage
Model
Clean Data (Good Model)
Clean Data (Pruned Model)
Poisoned Data (Good Model)
Poisoned Data (Pruned Model)
0
20
40
60
80
100
2
4
10


Combined Accuracy vs Number of Channels Dropped
Accuracy
Number of Channels Dropped
Combined Accuracy
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50
60


Combined Attack Success Rate vs Number of Channels Dropped
Attack Success Rate
Number of Channels Dropped
Combined Attack Success Rate
0
20
40
60
80
100
0
10
20
30
40
50
60