LAB-2

SS14885

**Experimental Setup:**

In this experiment, we first derived the Feature Map of the "pool3" layer, and then took the average activations of each channel over the clean validation set. We then arranged the indexes of the channels in ascending order according to their average activations. We accessed these indexes one by one to assess the drop in validation accuracy. If the drop in clean accuracy reached a certain threshold, we implemented early stopping and saved the model. We then created a new model G based on these requirements, and verified its accuracy and attack success rate using different threshold values (2%, 4%, and 10%).

**Observation:**

The experiments showed that the clean test data accuracy on BadNet before implementing the pruning defense was 98.649%. The table below shows how the attack success rate decreases as we increase the threshold of the validation accuracy (i.e., at what point to stop the pruning process). This observation aligns with the theory that by increasing the threshold, we effectively increase the number of pruned channels, which increases the defensive nature of the repaired net.

The following table contains the repaired net’s accuracy and backdoor attack success rate after the model is pruned.

| Threshold | Accuracy on clean Data | Attack Success Rate |
| --- | --- | --- |
| 2% | 95.74 | 100 |
| 4% | 92.13 | 99.98 |
| 10% | 84.33 | 77.21 |