ML-Cybersec HW2
Mohit Kewlani(mmk9369)

---

## 1. Introduction

In this homework we need to repair a backdoored neural network using the pruning defense technique. The defense mechanism works by pruning a neuron from the last pooling layer at a time. On the clean validation dataset, the neurons are pruned in the order of lowest to highest average activation value. Pruning is carried out until the accuracy of the validation dataset is changed by more than a predefined threshold. We had three thresholds for this homework: 2, 4, and 10%.

## 2. Results

The following table reports the accuracy on clean test data and the attack success rate on backdoored test data.

| Threshold | Channels Pruned | Accuracy | Attack Success Rate |
|---|---|---|---|
| 2% | 73% | 95.74% | 100% |
| 4% | 78% | 92.12% | 99.98% |
| 10% | 85% | 84.33% | 77.2% |