LAB2 REPORT (SP6646)

A backdoor attack on a neural network is a type of malicious attack in which a hacker or attacker secretly injects a "backdoor" into the model during the training phase, allowing them to gain unauthorized access to the model at a later time and cause it to make incorrect predictions or behave in some other undesirable way. This can be done by introducing a small, hidden pattern or "trigger" into the training data that the attacker knows about, but that the model's creators and users are unaware of.

One example of a backdoor attack on a neural network is the BadNets attack. In this attack, the attacker first trains a "poisoned" version of the target neural network on a dataset that contains a hidden, attacker-chosen pattern.

Once the poisoned model has been trained, the attacker then injects it into the victim's system, replacing the original, benign model. At this point, the attacker can cause the model to make incorrect predictions by providing it with input data that contains the hidden pattern.

We thwart this attack by pruning the models channels iteratively using only the clean validation data and removing the spurious nodes that encode the backdoor.

Showing the stats from 50% channels pruned (since before that numbers dont change):

Percent Of Channels Removed	Clean Test Acc	Attack Success Rate
0.516	98.62	100.0
0.533	98.62	100.0
0.55	98.61	100.0
0.567	98.60	100.0
0.6	98.59	100.0
0.617	98.57	100.0
0.65	98.44	100.0
0.667	98.41	100.0
0.7	97.75	100.0
0.717	97.51	100.0
0.733	95.74	100.0
0.75	95.35	99.99
0.767	94.90	99.99
0.8	91.58	99.99
0.817	91.13	99.98
0.833	89.68	80.74
0.85	84.33	77.02
0.867	76.17	35.71
0.883	54.68	6.95
0.9	27.07	0.42
0.917	13.70	0.0
0.933	6.56	0.0
0.95	1.52	0.0