

Lab Report

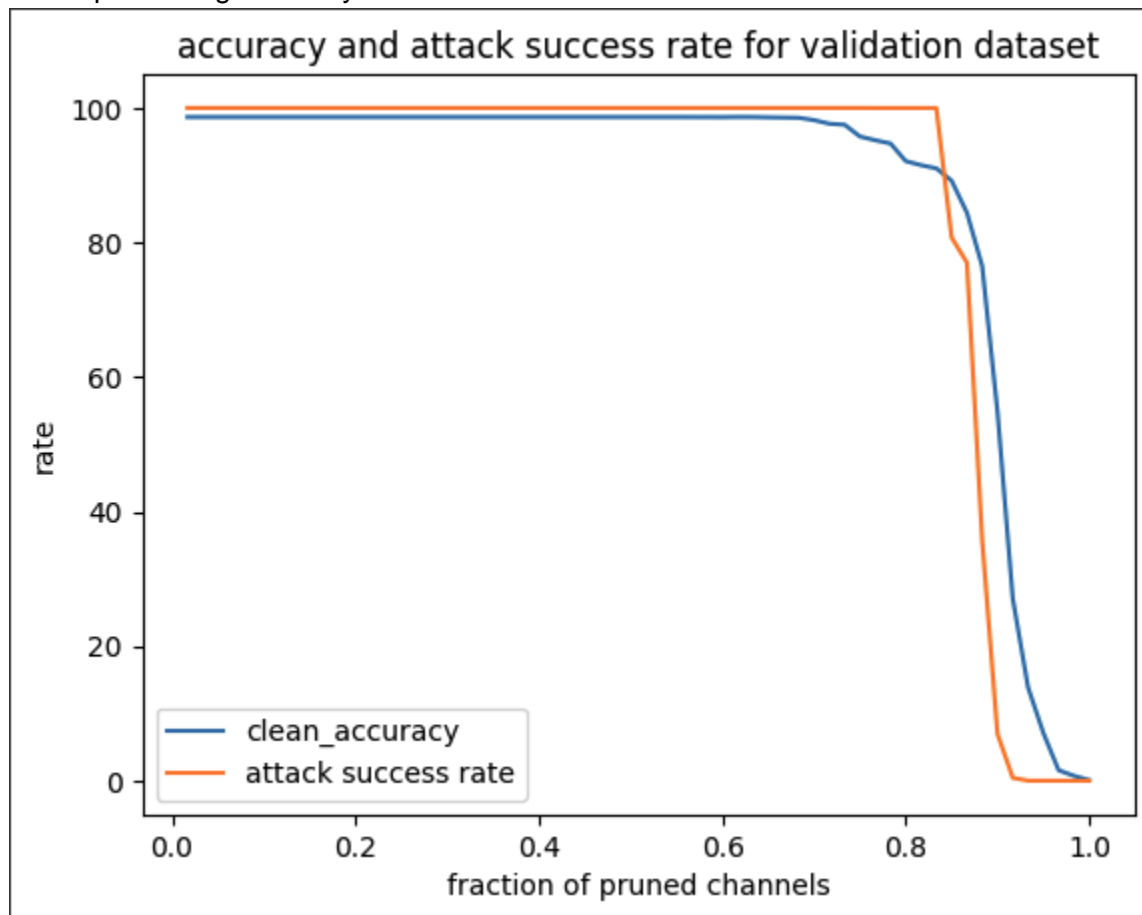
Github: https://github.com/sainarasimha37/backdoor_attacks.git

Name: Sai Narasimha Vayilati

NetId: sv2448

I have obtained diverse test accuracies when evaluating the performance on clean datasets using the pruned versions of the compromised neural networks provided in the 2020 hacks git repository. The pruning technique was employed to eliminate affected layers in the specified bad neural networks, altering the behavior for certain output classes. Pruning is a crucial method in deep learning, contributing to a reduction in both time and space complexity by decreasing the number of parameters, particularly as we approach the final layer of the neural network.

Plot representing accuracy vs attack success rate for the validation dataset:

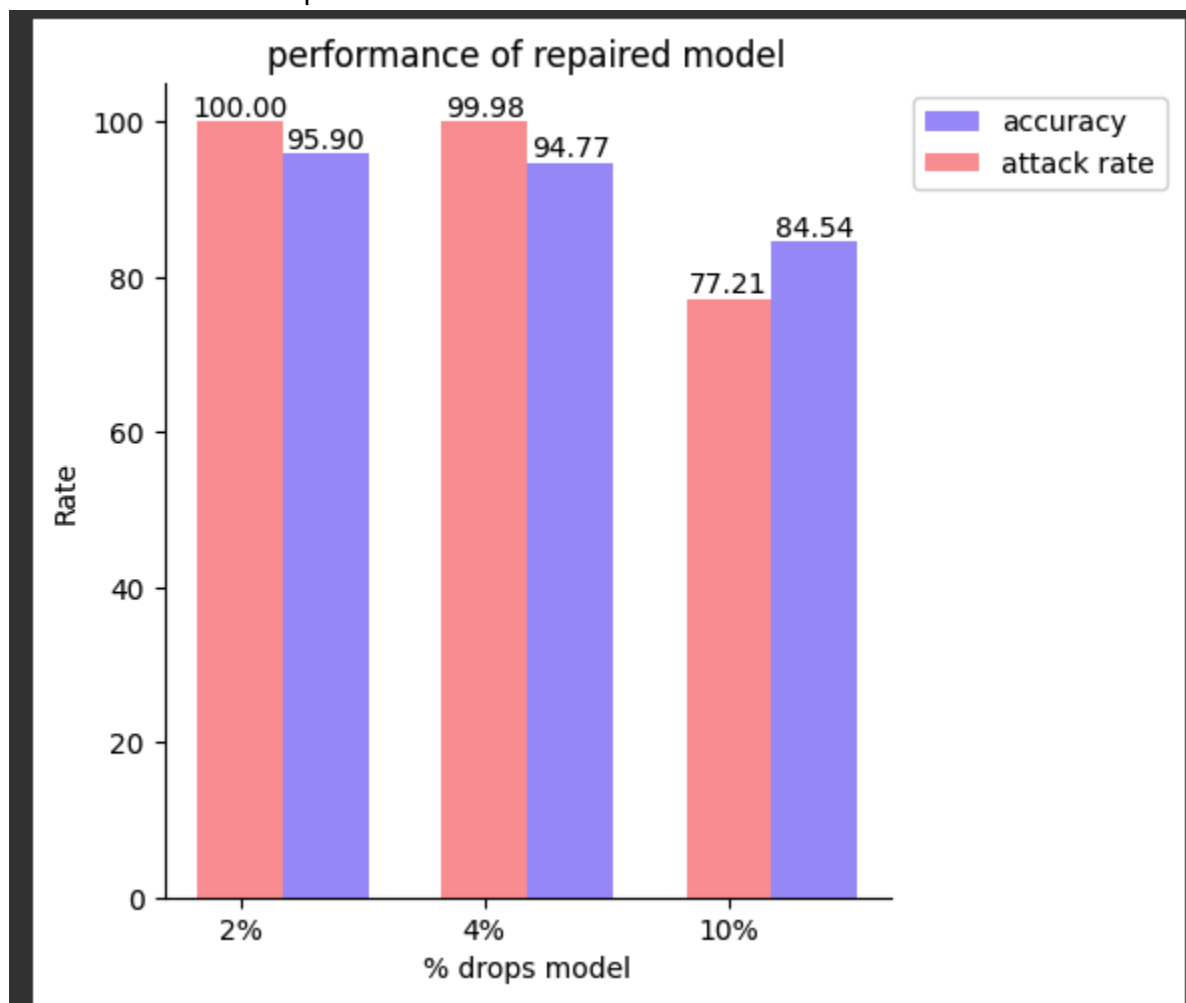


The Attack Success rate when the accuracy drops at least 30% is 6.954.

For the repaired net, which is by pruning out the bad net model by 2%, 4% and 10%, we received test accuracies as well as the attack rate for the following 3 repaired net models we have trained:

	test_accuracy	attack_rate
model		
repaired_2%	95.900234	100.000000
repaired_4%	94.770070	99.984412
repaired_10%	84.544037	77.209665

Performance of the Repaired Net:



For combined Good net models:

Combined 2% drops model, the clean test data Classification accuracy: 95.90023382696803

Combined 2% drops model, Attack Success Rate: 100.0

Combined 4% drops model, the clean test data Classification accuracy: 94.77007014809041
Combined 4% drops model, Attack Success Rate: 99.98441153546376
Combined 10% drops model, the clean test data Classification accuracy: 84.54403741231489
Combined 10% drops model, Attack Success Rate: 77.20966484801247

Test Accuracies and attack rates for models combined:

Goodnet_model	Good Net_test_acc	Goodnet_attack_rate
G_2%	95.900234	100.000000
G_4%	94.770070	99.984412
G_10%	84.544037	77.209665

Performance of the GoodNet Model:

