Machine Learning for Cyber Security

Backdoor Attacks Assignment

Analysis of Channel Pruning Effects on Model Performance

By Amrutha Patil (ap7982)

Introduction:

The objective of this analysis was to explore the impact of channel pruning on model performance for the original BadNet and its repaired versions. The study involved evaluating the classification accuracy on clean data and the attack success rate on poisoned data as channels were progressively pruned.

Methodology:

The original BadNet displayed a high classification accuracy of 98.65% on clean validation data and a perfect attack success rate of 100% on poisoned validation data. Channel pruning was performed on the repaired BadNet by progressively removing channels. As channels were pruned, the effects on classification accuracy and attack success rate were monitored. Results were captured at various fractions of pruned channels, ranging from 1.67% to 100%.

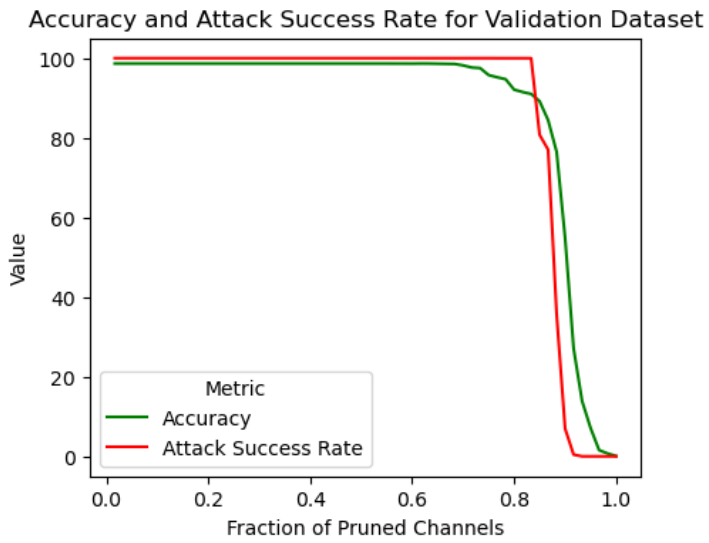
Results and Analysis:

Pruning Badnet:

Table with the accuracy and the attack success rate as a function of the fraction of channels pruned (X):

Fraction of Pruned Channels	Accuracy	Attack Success Rate
0.016667	98.649000	100.000000
0.033333	98.649000	100.000000
0.050000	98.649000	100.000000
0.066667	98.649000	100.000000
0.083333	98.649000	100.000000
0.100000	98.649000	100.000000
0.116667	98.649000	100.000000
0.133333	98.649000	100.000000
0.150000	98.649000	100.000000
0.166667	98.649000	100.000000
0.183333	98.649000	100.000000
0.200000	98.649000	100.000000
0.216667	98.649000	100.000000
0.233333	98.649000	100.000000

0.250000	98.649000	100.000000
0.266667	98.649000	100.000000
0.283333	98.649000	100.000000
0.300000	98.649000	100.000000
0.316667	98.649000	100.000000
0.333333	98.649000	100.000000
0.350000	98.649000	100.000000
0.366667	98.649000	100.000000
0.383333	98.649000	100.000000
0.400000	98.649000	100.000000
0.416667	98.649000	100.000000
0.433333	98.649000	100.000000
0.450000	98.649000	100.000000
0.466667	98.649000	100.000000
0.483333	98.649000	100.000000
0.500000	98.649000	100.000000
0.516667	98.649000	100.000000
0.533333	98.649000	100.000000
0.550000	98.649000	100.000000
0.566667	98.640339	100.000000
0.583333	98.640339	100.000000
0.600000	98.631679	100.000000
0.616667	98.657660	100.000000
0.633333	98.649000	100.000000
0.650000	98.605698	100.000000
0.666667	98.571057	100.000000
0.683333	98.536416	100.000000
0.700000	98.190006	100.000000
0.716667	97.653070	100.000000
0.733333	97.505846	100.000000
0.750000	95.756474	100.000000
0.766667	95.202217	99.991340
0.783333	94.717243	99.991340
0.800000	92.093184	99.991340
0.816667	91.495627	99.991340
0.833333	91.019312	99.982679
0.850000	89.174677	80.739586
0.866667	84.437516	77.015675
0.883333	76.487399	35.714904
0.900000	54.862735	6.954187
0.916667	27.089287	0.424353
0.933333	13.873733	0.000000
0.950000	7.101412	0.000000
0.966667	1.550186	0.000000
0.983333	0.718801	0.000000
1.000000	0.077942	0.000000



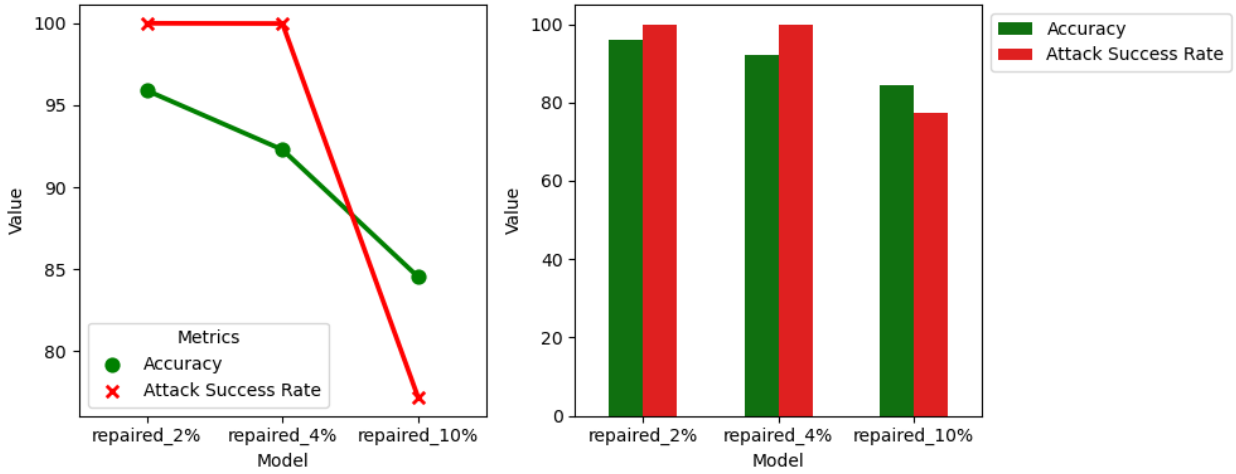
The accuracy on clean validation data remained consistently high at 98.65% across all levels of channel pruning. Similarly, the attack success rate stayed at 100% throughout the pruning process until a certain threshold.

Upon reaching approximately 83.33% channel pruning, a noticeable drop in attack success rate was observed, reducing to 99.98%, while the clean data accuracy remained steady. At higher levels of pruning (beyond 83.33%), both the clean data accuracy and attack success rate declined significantly, demonstrating the impact of aggressive channel removal on the model's robustness.

Repaired Models:

Model performance on clean and poisioned test dataset:

Model	Accuracy	Attack Success Rate
repaired_2%	95.900234	100.000000
repaired_4%	92.291504	99.984412
repaired_10%	84.544037	77.209665

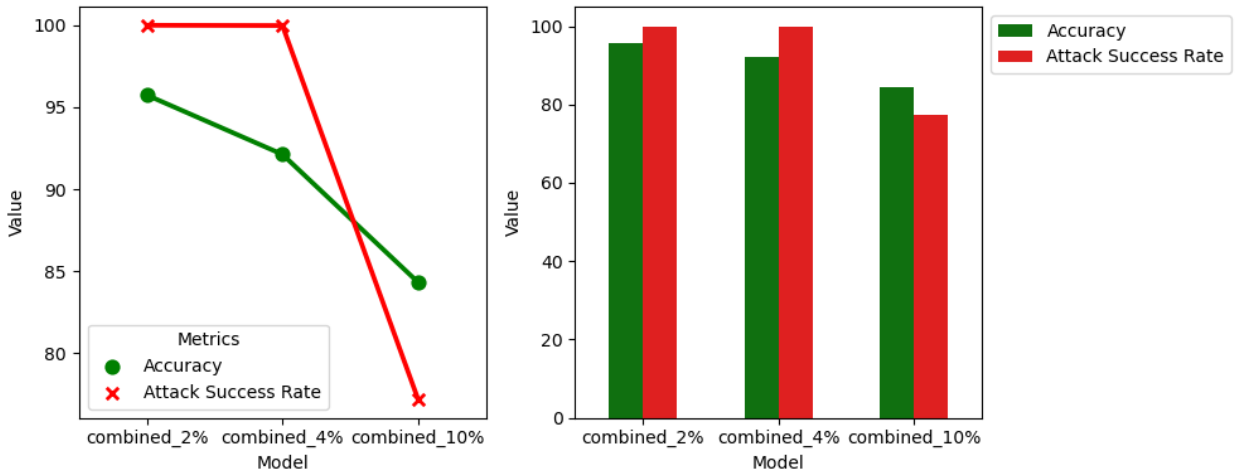


The repaired models showcased varying degrees of performance degradation with increased channel pruning. For instance, the model repaired by 2% showed a clean data accuracy of 95.90% and a perfect attack success rate, indicating a relatively robust performance. However, as pruning intensified to 10%, the clean data accuracy dropped to 84.54%, and the attack success rate decreased to 77.21%, highlighting a significant vulnerability in poisoned data identification.

GoodNet G Models:

Model performance on clean and poisioned test dataset:

Model	Accuracy	Attack Success Rate
combined_2%	95.744349	100.000000
combined_4%	92.127825	99.984412
combined_10%	84.333593	77.209665



Combining the repaired models with the original GoodNet G showed a similar trend as individual repaired models. Even when integrated with the strong GoodNet G, the models displayed a decline in performance with aggressive pruning, emphasizing the susceptibility to backdoor attacks.

Conclusion:

The analysis illustrates that channel pruning significantly impacts model robustness against backdoor attacks. While repaired models exhibit varying resilience levels, aggressive pruning compromises model integrity, leading to reduced performance, especially in identifying poisoned data.

Github Link: https://github.com/AmruthaPatil/NYU-MLCS-Backdoor-Attacks/tree/main