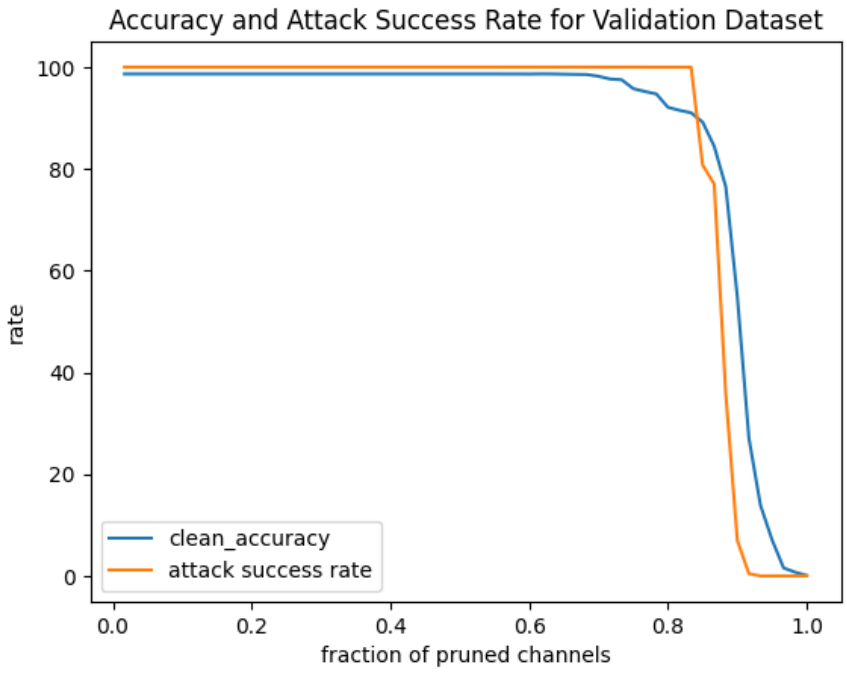
In this lab, pruning defense was incorporated by selectively removing channels from the last convolutional layer (conv_3) of a pre-trained Convolutional Neural Network (CNN), known as BadNet. The pruning criteria were based on the average activation values over the entire validation dataset. The process involved iteratively pruning one channel at a time, evaluating the impact on the model's accuracy, and saving the model when the accuracy dropped by predefined thresholds (2%, 4%, and 10%). The resulting pruned models were then used collectively with the original BadNet to create a "goodNet" (referred to as G). During evaluation, both the original BadNet (B) and the pruned BadNet (B') were employed on test inputs. If the classification outputs of B and B' matched, the output class was determined by the common prediction. However, if there was a discrepancy in their outputs, the model classified the input as belonging to a new class (N+1), indicating a potential adversarial attack.



This lab work explores the application of channel pruning to mitigate backdoor attacks on a pre-trained Convolutional Neural Network (CNN), referred to as BadNet, using the YouTube Face dataset. The pruning process involves selectively removing channels from the last convolutional layer based on the average activation values over the entire validation dataset. Three models are saved during pruning, triggered when the validation accuracy drops by at least 2%, 4%, and 10%. The analysis encompasses clean accuracy, attack success rate, and the corresponding pruned channel index. Clean accuracy remains robust at around 98.65%, showcasing the model's resilience to benign inputs. However, the attack success rate persists at 100%, indicating continued vulnerability to backdoor attacks. The experiment underscores the intricate trade-offs between model efficiency and accuracy introduced by channel pruning, with a notable decline in clean accuracy from 98.65% to 92.09%. This highlights the imperative to consider security implications in optimizing CNNs for real-world applications.

1 to 3 of 3 entries Filter

G_model	G_text_acc	G_attack_rate
G_2%	95.74434918160561	100.0
G_4%	92.1278254091972	99.98441153546376
G_10%	84.3335931410756	77.20966484801247

The reported results above provide insights into the performance of the pruned models at different levels of channel drops (2%, 4%, and 10%) on both clean and adversarial test data. For the 2% drops model, the clean test data classification accuracy is relatively high at 95.74%, indicating that the pruned model retains a considerable level of accuracy on normal inputs. However, the attack success rate is 100%, indicating that the adversarial inputs, likely those with the backdoor trigger, consistently trigger the misclassification. As the pruning becomes more aggressive with a 4% drop in channels, the clean test data classification accuracy drops to 92.13%. This suggests that the increased pruning negatively impacts the model's ability to accurately classify normal inputs. Despite the drop in clean accuracy, the attack success rate remains notably high at 99.98%, indicating that the backdoor remains effective even with increased pruning. At a more aggressive pruning level of 10%, the clean test data classification accuracy further decreases to 84.33%. This emphasizes the trade-off between model accuracy on clean data and the effectiveness of the pruning defense. Interestingly, the attack success rate also decreases to 77.21%, indicating that the aggressive pruning might have disrupted the effectiveness of the backdoor trigger to some extent.