REPORT – Lab 4

fx2078

In my methodology, I embark on a comprehensive evaluation process by first establishing the baseline accuracy and determining the success rates of attacks on the model using the clean dataset's validation split. Moving forward, a critical aspect involves identifying activations at the last pooling layer, which encompasses 60 convolution channels. To optimize model performance, I implement a pruning technique inspired by Garg et al.'s [Fine Pruning], wherein these channels are systematically pruned in ascending order based on their average activation levels over the validation dataset.

The subsequent phase involves a meticulous sequential pruning of channels, with each channel's weights and biases set to zero. This strategic approach targets the elimination of the least activated channels, a crucial step given that such channels are frequently exploited by attackers to introduce backdoors, ultimately leading to model misclassifications. Following each iteration of channel pruning, a thorough assessment of validation accuracy on the clean dataset is conducted. Models that fall below a predefined accuracy threshold are preserved as benchmarks for specific accuracy levels.

The creation of diverse models based on varying accuracy levels is integral to the methodology. This involves constructing both a Goodnet derived from the original BadNet and a BadNet_pruned. The decision-making process relies on the agreement between these models: if they align on an outcome, the prediction is accepted; conversely, a discrepancy implies that the model made a prediction based on a compromised dataset, resulting in the Goodnet having an additional dimension on the last layer.

The results obtained from this methodology unveil a consistent pattern: as more channels are pruned, there is a gradual decline in validation accuracy. Notably, beyond a certain threshold, this decline becomes pronounced. This phenomenon is attributed to the removal of inactive channels having a negligible impact, whereas the elimination of active channels significantly impairs the model's overall performance.

Regarding the accuracy of the Goodnet, it is observed that the model's accuracy mirrors the decrease seen in the accuracy of the BadNet_pruned model. Simultaneously, an increase in the attack rate is noted as the accuracy of the Goodnet diminishes. This correlation suggests a noteworthy trade-off between model accuracy and susceptibility to adversarial attacks. As the model's accuracy declines, it appears to become more vulnerable to attacks, emphasizing the delicate balance between robustness and predictive performance in the context of the Goodnet.