Defense against Deep Learning Attacks

Leiqi Wang

March 2022

1 Overview

Deep Learning(DL) is widely used in Network Security because of its high accuracy. The security of DL has aroused the interest of researchers because huge losses will come if the detector is not secure. They found DL model is vulnerable to DL Attacks, like poisoning attacks and adversarial attacks, where the former occurs in the training and the latter occurs in the testing. My previous research is to detect and classify network attacks with DL models. To improve my models' safety and strengthen network security, I need to research Deep Learning Defense.

2 Proposed Methodology

Data modification, model modification, and detection using additional networks are commonly used in the prevention of DL attacks. Data modification indirectly changes the composition of training data or the feature space during the 'data collection and pre-processing' to improve the robustness. But, most of these methods require model retraining. Model modification directly modifies the target classification model. They may mask gradient information, increase the degree of model regularization, and use more sophisticated classification mechanisms to defend. However, such methods increase the complexity of the model and may decrease its accuracy. Detection using additional networks is to train additional ML models to detect adversarial samples without changing the target classification model, thereby excluding the influence of the adversarial samples on the target classifier. But this approach increases the training cost and may affect the efficiency and usefulness of the target classification model.

3 Research Plan

The methods mentioned above defend against attacks in terms of either data or models but fail to integrate. Integrated optimization of data and models will make it possible to reduce the number of retraining and achieve defense

more quickly. So I would like to propose a security method that combines preprocessing and model pruning training to improve the resistance against DL attacks. I plan to publish a paper to prove my method's effectiveness.