Computer Science
Defense

Defensive Machine Learning Techniques for Countering Adversarial Attacks

Fereshteh Razmi
Emory University

Abstract: The increasing reliance on machine learning algorithms has made them a target for exploiting vulnerabilities in these systems and launching adversarial attacks. The attacker in these attacks manipulates either the training data or test data, or both, known as a poisoning attack, adversarial example, or backdoor attack, respectively. They primarily aim to disrupt the model's classification task. In cases where the model is interpretable, the attacker may target the interpretation of the model's output. ¡br¿ These attacks can have significant negative impacts; therefore, it is crucial to develop effective defense methods to protect against them. Current defense methods have limitations. Outlier detectors, used to identify and mitigate poisoning attacks, require prior knowledge of the attack and clean data to train the detector. Robust defense methods show promising results in mitigating backdoor attacks, but their effectiveness comes at the cost of decreased model utility. Furthermore, few defense methods have addressed adversarial examples that target the interpretation of the model's output. ¡br¿ To address these limitations, we propose defense methods that protect machine learning models from adversarial attacks. Our methods include an autoencoder-based detection approach to identify various untargeted poisoning attacks. We also provide a comprehensive comparative study of differential privacy approaches and suggest new approaches based on label differential privacy to defend against backdoor attacks. Lastly, we propose a novel attack and defense method to protect the interpretation of a healthcare-related machine learning model. These approaches represent significant progress in the field of machine learning security and have the potential to protect against a wide range of adversarial attacks."

Thursday, April 20, 2023, 2:00 pm
https://zoom.us/my/lxiong

Computer Science
Emory University