COLLEGE of ENGINEERING AND PHYSICAL SCIENCES

SCHOOL OF COMPUTER SCIENCE

PhD Qualifying Examination

Amin Azmoodeh

Monday July 19, 2021 at 2:30PM via Zoom

A Framework to Enhance the Robustness of Deep Learning Models Against Out-of-Distribution Samples

Chair: Dr. Joe Sawada
Advisor: Dr. Ali Dehghantanha
Advisory: Dr. Xiaodong Lin
Non-Advisory: Dr. Graham Taylor [Engineering]
Non-Advisory: Dr. Hassan Khan

Abstract:

During the past decade, deep learning has achieved significant advancements and deep learning-based models have outperformed classical algorithms and even human beings in variety of tasks ranging from autonomous vehicles to image recognition and even cybersecurity. For most of its life, deep learning algorithms were assumed working in a safe environment and in the absence of any adversaries. However, a considerable body of researches has been conducted to demonstrate susceptibility of deep learning algorithms to a wide range of uncontrollable and adversarial inputs.

A large body of research on the safety and security of deep learning has considered the task of generating malicious inputs on a closed-world system in which generated examples are following a specific distribution comparable to the training data. Notwithstanding, real-world applications of deep learning are performing in open-world environments and many trained deep learning models are receiving inputs from unknown data distributions also are known as Out-Of-Distribution (OOD) examples.

In this proposal, we are investigating the effects of open-world and adversarial setting related to examples generated from OOD inputs and propose a framework that 1) detects OOD inputs and adversarial attacks against deep learning models; 2) generates adversarial examples from OOD examples that can bypass deep learning models and identifies deep learning's vulnerability surface for OOD inputs; 3) presents an alternative deep learning model that is more robust against OOD inputs.

The proposed framework provides deep learning research and development community with an empirical approach to identify the vulnerability of their trained models to OOD and adversarial examples. In addition, the proposed protection model elevates the robustness of DL models to identified OOD adversarial attacks and decreases OOD inputs' probability of success.