UNIVERSITY of GUELPH

COLLEGE of ENGINEERING AND PHYSICAL SCIENCES

SCHOOL OF COMPUTER SCIENCE

PhD Seminar 1

Monday November 11, 2019 at 1PM in Reynolds, Room 3324

A Framework for Defending Deep Neural Networks Against Out-Of-Distribution Adversarial Attacks

Amin Azmoodeh

Advisor: Dr. Ali Dehghantanha

Co-Advisor: Dr. Bahram Gharabaghi [School of Engineering]

Advisory Committee: Dr. Xiaodong Lin

Advisory Committee: Dr. Raymond Choo [University of Texas San Antonio]

ABSTRACT:

Machine Learning has significant development during the past decade and machine learning-based models have outperformed classical algorithms and even human beings in variety of tasks including object recognition, malware detection, financial predictions, playing games, and medical recognition. Deep Learning is possibly the most widely adopted subset of machine learning in different tasks. Currently, deep learning plays an important role in different disciplines 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, many researchers recently showed susceptibility of deep learning algorithms to a wide range of malicious and adversarial inputs. Adversarial attacks are wisely crafted input samples that bypass a trained deep model and generate erroneous output.

A large body of research on the adversarial attack against deep learning has considered the task of generating adversarial samples on a closed-world system in which generated samples are following a specific distribution comparable to the training data. Notwithstanding, read-world applications of deep learning are performing in open-world environments and many trained deep models are receiving inputs from unknown data distributions also are known as out of distribution (OOD) samples.

In this proposal, we are investigating the effects adversarial setting related to samples generated from out-of-distribution inputs and propose a framework that 1) identifies vulnerability surface of the given deep model against OOD attacks; 2) automatically generates OOD adversarial samples that can be used to exploit detected vulnerabilities; 3) detects potential OOD attacks against the given deep model, and 4) generates an input sanitization layer that protects given deep learning models against OOD attacks. At the first stage, we proposed a model to empirically identify vulnerability surface of deep models against OOD adversarial samples. Then, in order to evaluate the usefulness of the model, we propose an OOD adversarial attack generation model to exploit detected vulnerabilities. Then, we introduce our state-full OOD adversarial attack detection model that considers the sequence of out-of-distribution samples to accurately identify attack attempts. Finally, we propose a model that generates a sanitization layer that maps the OOD adversarial inputs to a new space that significantly reduces the attack success rate of OOD samples.

The proposed framework provides deep learning research and development community with an empirical approach to identify the vulnerability of their trained model to OOD adversarial attacks. In addition, the proposed protection model elevates the security of deep learning models to identify OOD adversarial attacks. Furthermore, this study empowers deep learning models by integrating them with a prevention layer that decreases OOD attacks' probability of success.