UNIVERSITY of GUELPH

COLLEGE of ENGINEERING AND PHYSICAL SCIENCES

SCHOOL OF COMPUTER SCIENCE

PhD Seminar 2

Monday January 9, 2023 at 2pm via Zoom

Amin Azmoodeh

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

Advisor: Dr. Ali Dehghantanha
Co-Advisor: Dr. Bahram Gharabaghi [SoE]
Advisory: Dr. Xiaodong Lin
Advisory: Kim-Kwang Raymond Choo [External]

Abstract:

During the past decade, deep learning has achieved significant advancements. Deep learning-based models have outperformed classical algorithms and human beings in various tasks ranging from autonomous vehicles to image recognition and cybersecurity. For most of its life, deep learning algorithms were assumed to work in a safe environment and the absence of any adversaries. However, a considerable body of research has been conducted to demonstrate the 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 generating or detecting malicious inputs on a closed-world system in which generated examples follow a specific distribution comparable to the training data. Notwithstanding, real-world applications of deep learning are performed in open-world environments. Many trained deep learning models receive inputs from unknown data distributions, known as Out-Of-Distribution (OOD) examples.

In this study, we investigate the effects of open-world and adversarial settings related to OOD inputs and propose a framework that consists of 1) an OOD and adversarial example detection module that recognizes malicious inputs based on deep model layers' activity pattern, 2) an OOD vulnerability identifier module to generate OOD examples that can bypass deep learning models and reveal deep learning's vulnerable areas of input space; and 3) a robustness module that leverages a new ontology-based loss function to train deep learning models and improve model's overconfident scores for OOD examples.

The proposed framework provides deep learning research and industry communities with a detection mechanism attachable to a trained deep learning model, accepts normal inputs, and rejects OOD and adversarial examples. In addition, the framework is equipped with an empirical approach to identify the vulnerability of trained models to OOD examples. Finally, the proposed protection model elevates the robustness of deep learning models through novel training/retraining techniques that cause fewer confidence scores for OOD examples.

The proposed framework is assessed using benchmark deep learning architectures and datasets. We utilize a range of evaluation metrics, including Accuracy, Detection Rate, False Positive Rate, Area Under Curve as well as statistical techniques to demonstrate the framework's efficiency.