UNIVERSITY of GUELPH

COLLEGE of ENGINEERING AND PHYSICAL SCIENCES

SCHOOL OF COMPUTER SCIENCE

PhD Defence

Thursday January 25, 2024 at 1PM, online via Zoom (Remote)

Amin Azmoodeh

A Framework to Enhance Security and Safety of Deep Learning Models Against Out-of-Distribution Examples

Chair: Dr. Stacey Scott
Advisor: Dr. Ali Dehghantanha
Co-Advisor: Dr. Bahram Gharabaghi [School of Engineering]
Non-Advisory: Dr. Graham Taylor [School of Engineering]
External Examiner: Dr. Benjamin Fung [McGill University]

Abstract:

In recent years, the realm of deep learning has witnessed remarkable progress, with models rooted in this paradigm surpassing traditional algorithms and, in some instances, human performance across diverse domains, encompassing autonomous navigation, image analysis, and cybersecurity. Historically, the operational premise for these deep learning algorithms was a controlled, adversary-free environment. Yet, an expanding corpus of scholarly work underscores the inherent vulnerabilities of deep learning methodologies when confronted with unpredictable and adversarial perturbations.

Much of the discourse surrounding the safety and integrity of deep learning models has been predicated on the generation of adversarial samples within a closed-world paradigm, wherein the adversarial instances adhere to a distribution analogous to the training dataset. However, it is imperative to recognize that many real-world deployments of deep learning operate in open-world settings, where models frequently encounter inputs drawn from distributions distinct from their training data, termed Out-Of-Distribution (OOD) samples.

This research endeavour seeks to rigorously explore the ramifications of adversarial scenarios, particularly those emanating from OOD sources. The proposed framework aims to: (1) discern OOD inputs and adversarial perturbations targeting deep learning architectures; (2) craft adversarial instances derived from OOD samples capable of evading deep learning defences, thereby delineating the susceptibility landscape of deep learning to OOD perturbations; and (3) introduce a refined deep learning architecture exhibiting enhanced resilience to OOD perturbations.

In order to evaluate the performance of the proposed framework, it has been assessed against a set of state-of-the-art deep learning models trained on benchmark datasets. Various evaluation metrics including Accuracy, Area-Under-Curve, Detection Rate and TPR@FPR are employed to assess the proposed framework performance. The conducted experiments demonstrated its usefulness for designed research objectives and promised the deep learning model's trustworthiness enhancement.

Our proposed framework offers the deep learning research community a systematic mechanism to ascertain the vulnerabilities of their trained architectures to both OOD and adversarial instances. Furthermore, the advanced protective model we propose augments the defence mechanisms of deep learning systems against detected OOD adversarial strategies, thereby diminishing the likelihood of successful OOD perturbations.