INFORMATICS INSTITUTE OF TECHNOLOGY

In Collaboration with

UNIVERSITY OF WESTMINSTER

ATERT : A General Defense Framework for Defending against Adversarial Attacks and Physical World Adversaries on Autonomous Driving

A dissertation by

Mr. K.T.Yasas Mahima

W1742097 - 2018362

Supervised by

Mr. Guhanathan Poravi

Submitted in partial fulfilment of the requirements for the BEng (Hons) Software Engineering degree at the University of Westminster.

May 2022

ABSTRACT

Though a wide range of domains has been influenced by the rise of deep learning and machine learning technologies, recent research works have identified these intelligent models are vulnerable to intentionally synthesized adversarial perturbations by attackers that are reliable enough to alter the prediction output without appealing a noticeable change in the input image to the human eye. With the advent of autonomous vehicles, this has earned higher attention and while moving deeper into the research domain, it can identify that, apart from adversarial attacks, the physical world itself acts as a performance degradation producer by constructing different adversarial constraints such as illumination changes, noises .etc.

This research aims to design, develop and evaluate a general model robustness approach for both man-made and physical world adversaries without changing the given model architecture or no usage of auxiliary tools in the inference primarily on the autonomous vehicle domain. As a result, the models that are robustified by the suggested approach are capable of easily integrating into any application without hesitating about the improvements in computational resource consumption. Grounded on the literature review, the author has proposed a combined two-step training approach (ATERT) of Projected Gradient Descent l_∞ based adversarial training and an improved version of the mix-up image transformation method named ERT. The experiment results demonstrate that the ATERT is capable of improving the resilience against both adversarial types without affecting the standard models' performance. In particular, ATERT improves the robustness for both digital and physical world adversaries up to 5-30% and 5-25% respectively on the evaluated models. Besides, a separate study conducted using Explainable AI further confirms that the ATERT improves the network's ability to capture pixel feature attributes under adverse conditions.

Key Words – Security of Intelligent Systems, Robustness of Machine Learning, Adversarial Machine Learning

ACM Subject Descriptions

1. Computing methodologies >> Artificial intelligence >> Computer vision >> Computer vision problems >> Object recognition
2. Computing methodologies >> Machine learning >> Machine learning approaches >> Neural networks
3. Security and privacy >> Software and application security