DEFENSE MECHANISM AGAINST ADVERSARIAL ATTACKS FOR OPTICAL CHARACTER RECOGNITION SYSTEMS

Fathima Shiffna Merza

A dissertation submitted in partial fulfilment of the requirement for Bachelor of Engineering (Honours) degree in Software Engineering

Department of Computing
Informatics Institute of Technology, Sri Lanka
in collaboration with
University of Westminster, UK

2021

Abstract

Deep neural networks are widely being employed for Machine learning related tasks like Optical Character Recognition. Modern OCR is a computer vision task which adopts DNN and are found to be vulnerable against adversarial samples. Adversarial text images can successfully mislead the model to produce erroneous outputs. These perturbations are crafted in a way which is benign to the human eye. Number of defenses have been proposed in the literature for image classification models. However, these approaches are not directly applicable to OCR. This research attempts to employ an image compression and transformation defense approach against the CRNN model to overcome this issue in a considerable way. Image transformation techniques are used to transform the images by compression before it is fed into the CRNN network. This eliminates the perturbations from the input level itself. This research project facilitates varying levels of compression. The author conducted experiments and results showcases that the defense was able to eliminate most of the perturbations for attacks like FGSM and recognize the misclassified text accurately. A much faster defense which can be seamlessly integrated with most of the models compared to the existing defenses in literature.

Keywords: Adversarial machine learning, Optical character recognition software, Data compression, Deep learning