SUPERDEF: RESPONSE BASED KNOWLEDGE DISTILLATION ON SISR AS AN EFFICIENT DEFENSE AGAINST ADVERSARIAL ATTACK

Nazar Ahmed Amjad

A dissertation submitted in partial fulfillment 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

With the increase of computational power and large gathering of data, Deep Learning has become an important part of today's business and industrial tasks to the extent of outperforming and replacing humans with machines. With that said, the question to be asked is can so much trust be put on Deep Learning? For example, consider autonomously driving cars, if the Deep Learning algorithm equipped fails identify obstacles it would end up in a disaster. Considering these risks, the research community has focused on investigating the threat of associated to Deep Learning.

In the recent past, it was identified that deep learning-based models can be fooled by simply adding a noise to the image. Following that research investigated on approaches for crafting this type of noise. Hence, raised the domain of Adversarial Attacks. Various researchers put forwards ways of crafting these attacks and some of they were able to nullify the accuracy of Deep Learning models. Further into research domain it was shown that this risk also exists in the Deep Learning models which was used in the real-world application.

Following the growth of the domain Adversarial Attacks the research community investigated on the countermeasures and was successful to increase the robustness of Deep Learning against the adversarial attacks. But most of the highly robust model consumes high computational resources and some of they are practically not deployable. In the other hand, various research has shown that model compression can reduce the complexity by maintaining the accuracy and performance of the model. Linking these dots, the author hypothesis that effective using model compression will reduces the resource consumption without affecting the robustness that much.

In this research the author was able to successfully implement SuperDef by apply Knowledge Distillation on RCAN model for single image super resolution task and use that as defense against adversarial attacks. The benchmark results of SuperDef outperforms the previous implementation of defense which uses Image Super Resolution in terms of performance and robustness. The protype of the SuperDef includes web application to show the results of the defense against various adversarial attacks.

Keywords

Adversarial Defense, Knowledge Distillation, Single Image Super Resolution