Adversarial attacks on medical deep learning models

Lau Young Kang, Loh Brian Chung Shiong, Vong Wan Tze, Then Patrick Hang Hui

Swinburne University of Technology Sarawak Campus

ABSTRACT

Introduction: Adversarial attacks are a great threat to deep learning (DL) as they can generate imperceptible perturbations in images which severely affects model performance. More worryingly, recent works have shown that medical DL models are vulnerable to such attacks. The DL process flow is susceptible to various kinds of adversarial attacks. Specifically, causative attacks occur before a model is built, during training, and exploratory attacks occur after model training, during the inference phase. Furthermore, these attacks can be exploited to compromise overall model accuracy, or influence results on specific targeted classes. This research aims to study the impact of causative and exploratory attacks for non-targeted and targeted purposes, on medical DL models built for image classification tasks. **Methods:** Warping Based Backdoor Attack and Universal Adversarial Pattern Attack were selected due to their superior performance in generating imperceptible adversarial samples for non-targeted and targeted attacks. DL models were produced from both original and perturbed ISIC-2019 dermoscopic and COVID-NET chest X-ray image datasets. These models were subsequently evaluated on their classification performance. **Results:** Experiments on models achieving above 90.0% accuracy revealed that both causative and exploratory attacks could lower model accuracy by at least 45.0%. In the best-case adversarial attack scenario, model accuracy was reduced by up to 99.0%. **Conclusion:** These results provide a better understanding on the damaging nature of causative and exploratory adversarial attacks as well as vulnerability of medical DL models. The findings can serve as a starting point towards building effective defence approaches that are vital for medical systems utilising DL algorithms.