AdvMed: Detecting Adversarial Attacks in Medical Deep Learning Systems

Mehrotra, Ayushi (School: Troy High School)

Deep neural networks are used in the medical industry as tools to diagnose skin cancer from photographic images and detect the severity of diabetic retinopathy. Alongside these steps in medical deep learning, adversarial attacks emerge as a threat, characterized by images minutely altered to produce misclassification while the perturbations are imperceptible to the human eye. Medical images have distinct characteristics, making adversarial examples more effective with less alteration. We propose to solely use the gradient of the medical image and the output of the deep learning model ResNet50 to detect adversarial examples. We develop four novel gradient-based functions, along with their proofs, as our detection methods. We test our detection methods against three different attacks on datasets of skin lesions and diabetic retinopathy on Amazon Sagemaker. Moreover, we attack our detection methods using the state-of-the-art attack called □□□2, which tries to mimic the gradient of a benign image while producing misclassification. We show through experiments that our defense is robust against this attack. Finally, we compare our collection of detection methods against Feature Squeeze, the currently accepted detection method, and show that our defenses outperform the state-of-the-art by over 300%.