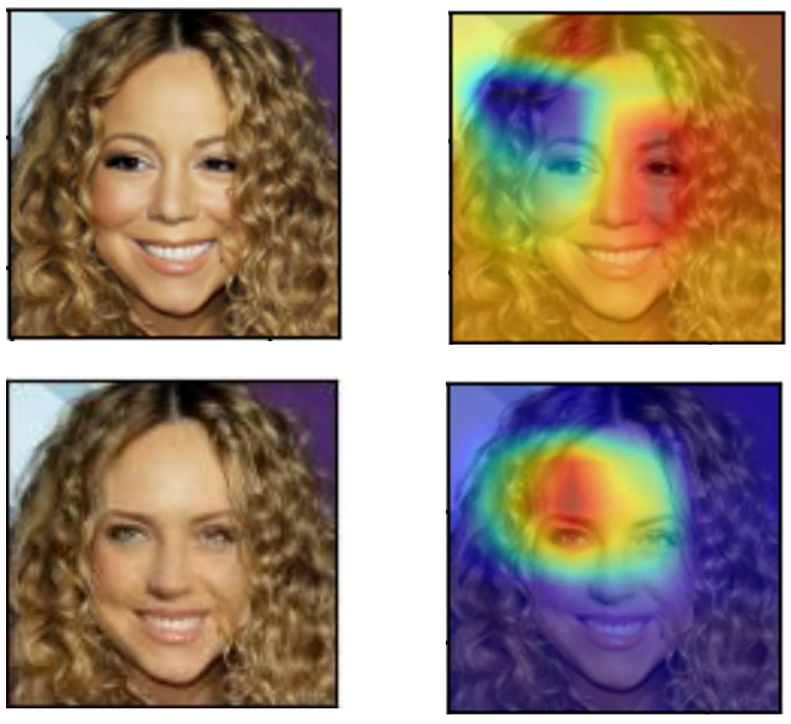
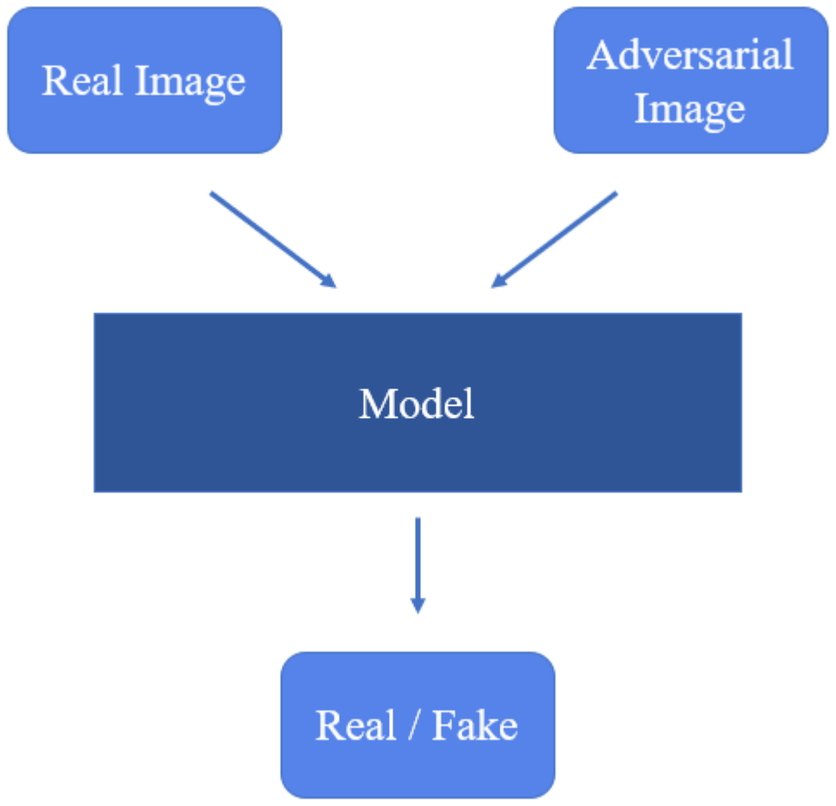
NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

School of Computer Science and Engineering
College of Engineering

Defence on unrestricted adversarial examples

Student: Chan Yan Cheng Jarod Supervisor: Ast/Prof Zhao Jun





Project Objectives:

Neural networks used for image classification can be fooled by unrestricted adversarial examples, which are crafted through methods such as noise corruption, generative models and more. The aim of this project is to investigate if unrestricted adversarial examples made from generative models can be easily distinguished from real images using CNNs. Identification of unrestricted adversarial examples can be essential in defending against adversarial attacks through filtering out blatant attacks against neural networks, preventing adversarial attacks from reaching the neural network to be classified.

Methodology:

This model can distinguish between adversarial examples and real images and will be trained through transfer learning from a Generative Adversarial Network (GAN) that creates unrestricted adversarial examples. During training of the GAN, its discriminator will learn to determine if the image produced by its generator is realistic, thus learning the useful features to classify adversarial examples. From here, we will fine tune the model to further train it in classifying adversarial examples.