Protection of Deep Neural Networks against Adversarial Attacks with Application to Facial Recognition

Guo, Alice (School: Morgantown High School)

The creation of deep neural networks (DNNs) in the field of artificial intelligence, containing tens to hundreds of layers to learn any complex function, give computers the ability to outperform humans on a wide spectrum of things, face recognition as one. However, it has been proven very recently that these networks are especially vulnerable to attacks, i.e., the so-called adversarial examples, which are imperceptible to humans but can fool DNNs easily. Thus, it is vital to develop defense technologies to make AI systems robust against various attacks. In this project, I propose a novel approach to defend against said adversaries in response to various attacks that generate adversarial examples. To contrast, the often used defense method based on adversarial learning requires days to train a model, whereas the newly proposed idea takes at most a few hours, utilizing the attacks themselves as a defense mechanism to mimic attacked examples (rather than using originals) for feature extraction. The task of face recognition is adopted to validate the novel idea and approaches experimentally. Based on a large-scale dataset with 6,000 pairs of face images, this new defense handles the adversarial attacks efficiently, improving facial recognition accuracies from about 0% under attack to above 80% after defense, leading to a very promising future.

Awards Won:
National Security Agency Research Directorate : Second Place Award "Science Security" of $1,000