A Biologically Inspired Game Theoretic Adversarial Training Method

Lee, Simon (School: Whittle School And Studios)

Recent research has shown that neural networks are vulnerable to errors caused by adversarial attacks; when carefully crafted, imperceptible noise is added to inputs, a neural network classifier’s accuracy is damaged significantly. This property is concerning when deep learning models are being applied in increasingly high-risk and security-critical environments with potential for human harm, such as autonomous vehicles. The state-of-the-art adversarial defense is adversarial training, but it is computationally inefficient considering its inability to confer robustness against unseen, diverse types of attacks. In this work, I improve on vanilla adversarial training by developing a 5-player minimax game in which a dual generator generative adversarial network (GAN) adversarially attacks an image classifier, and the classifier trains on generated adversarial examples to robustify in response. I use a custom loss function combining Gabor filter bank, orthogonality, and decision boundary guidance regularizers, which helps the classifier to learn robust representations that are similar to primate primary visual cortex behavior whilst balancing standard accuracy tradeoff. Experimental results on MNIST and CIFAR10 show that GAN-based adversarial training regularized in this manner successfully robustifies deep classifiers against a wide-ranging benchmark consisting of state-of-the-art white-box attacks and common image corruptions.