Reevaluating adversarial defences

Background: Adversarial attacks were first discovered in the context of deep neural networks (DNNs), where the networks' gradients were used to produce small bounded-norm perturbations of the input that significantly altered their output. Such attacks target the increase of the model's loss or the decrease of its accuracy and were shown to undermine the impressive performance of DNNs in multiple fields. The usually considered accessibility setting for adversarial attacks is ``white-box" attacks, in which the attacks can access the weights and gradients of the model. However, attacks have also been shown to exist in a ``black-box" setting, in which they can only access the input and output of the model. Relevant papers: “Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples”, “Explaining and harnessing adversarial examples.”, “A data augmentation-based defense method against adversarial attacks in neural networks”.
Project Description: Adversarial defenses are methods aiming to mitigate the effect of such attacks and usually utilize either adversarial training or data augmentation-based approaches. Multiple models report relatively high robustness to adversarial attacks, however, some of the defense effects may be due to gradient obfuscation which gives a false sense of security. In this project, we aim to reevaluate the robustness of models by utilizing black-box adversarial attacks.
Prerequisites: Deep learning course
Supervisor: Yaniv Nemcovsky (yanemcovsky@gmail.com)