Unadversarial attacks as adversarial defence

Background: Adversarial attacks are small bounded-norm perturbations of a network's input that aim to alter the network's output and are known to mislead and undermine the performance of deep neural networks (DNNs). Adversarial defenses then aim to mitigate the effect of such attacks, and unadverserial are the self-application of attacks for improved performance. Relevant papers: “Unadversarial examples: Designing objects for robust vision”.
Project Description: In this project, we aim to use unadverserial attacks as an adversarial defense, targeting the improving robustness of models to adversarial attacks.
Prerequisites: Deep learning course
Supervisor: Yaniv Nemcovsky (yanemcovsky@gmail.com)