Patch adversarial attacks with optimized location

Background: Adversarial perturbations 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). Sparse adversarial perturbations constitute a setting in which the perturbations are limited to affect a relatively small number of points in the input. Patch adversarial attacks are then sparse attacks in which the perturbed points are additionally limited to a given structure and location. Relevant papers: “Sparse and imperceivable adversarial attack”, “Physical passive patch adversarial attacks on visual odometry systems”.
Project Description: In this project, we aim to implement a patch adversarial attack in which we optimize both the location of the patch and the corresponding perturbation.
Prerequisites: Deep learning course
Supervisor: Yaniv Nemcovsky (yanemcovsky@gmail.com)