Evaluating the point-wise significance of sparse adversarial attacks

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. Relevant papers: "Sparse and imperceivable adversarial attack".
Project Description: In this project, we discuss the point selections produced by sparse adversarial attacks and aim to evaluate their point-wise significance in the corresponding input sample.
Prerequisites: Deep learning course
Supervisor: Yaniv Nemcovsky (yanemcovsky@gmail.com)