YANIV ROMANO, Stanford University
Classification Stability for Sparse-Modeled Signals

Despite their impressive performance, deep convolutional neural networks (CNNs) have been shown to be sensitive to small adversarial perturbations. These nuisances, which one can barely notice, are powerful enough to fool sophisticated and well performing classifiers, leading to ridiculous misclassification results. We study the stability of state-of-the-art classification machines to adversarial perturbations by assuming that the signals belong to the (possibly multi-layer) sparse representation model. We start with convolutional sparsity and then proceed to its multi-layered version, which is tightly connected to CNNs. Our claims can be translated to a practical regularization term that provides a new interpretation to the robustness of Parseval Networks. Also, the proposed theory justifies the increased stability of the recently emerging layered basis pursuit architectures, when compared to the classic forward-pass.