Speaker: Zara KADKHODAIE

Title: Generalization in diffusion models arises from geometry-adaptive harmonic representations

Abstract: Deep neural networks (DNNs) trained for image denoising are able to generate high-quality samples with score-based reverse diffusion algorithms. These impressive capabilities seem to imply an escape from the curse of dimensionality, but recent reports of memorization of the training set raise the question of whether these networks are learning the “true” continuous density of the data. Here, we show that two DNNs trained on non-overlapping subsets of a dataset learn nearly the same score function, and thus the same density, when the number of training images is large enough. In this regime of strong generalization, diffusion-generated images are distinct from the training set, and are of high visual quality, suggesting that the inductive biases of the DNNs are well-aligned with the data density. We analyze the learned denoising functions and show that the inductive biases give rise to a shrinkage operation in a basis adapted to the underlying image. Examination of these bases reveals oscillating harmonic structures along contours and in homogeneous regions. We demonstrate that trained denoisers are inductively biased towards these geometry-adaptive harmonic bases since they arise not only when the network is trained on photographic images, but also when it is trained on image classes supported on low-dimensional manifolds for which the harmonic basis is suboptimal. Finally, we show that when trained on regular image classes for which the optimal basis is known to be geometry-adaptive and harmonic, the denoising performance of the networks is near-optimal.