מכון ויצמן למדע

THE WEIZMANN INSTITUTE OF SCIENCE

FACULTY OF MATHEMATICS AND COMPUTER SCIENCE

Machine Learning and Statistics Seminar

Room 1 ,Ziskind Building
on Tuesday, Aug 13, 2019
at 14:15

NOTE THE UNUSUAL TIME

Stefano Soatto
University of California Los Angeles

The Information in the Weights of a Deep Network, and its consequences for transfer learning and critical learning periods

Abstract:

The information in the weights of deep networks plays a key role in understanding their behavior. When used as a regularizer during training (either explicitly or implicitly) it induces generalization through the PAC-Bayes bound. Rather than being computed and minimized explicitly, it can be directly controlled by injecting noise in the weights, a process known as Information Dropout. It can also be shown that stochastic gradient descent (SGD), when taken to the continuous limit and interpreted in the Wasserstein metric, adds the information in the weights as inductive bias, even if not explicitly present in the loss function. The Emergence Bound shows that, provided that a trained network has sufficient capacity, minimizing the information in the weights, which is a function of the training set consisting of data seen in the past, guarantees minimality, invariance, and independence of the components of the representation of the test (future) data. The trace of the information in the weights during training shows that, rather than increasingly monotonically through the learning process, as one would expect, first increases rapidly in the first few epochs, and then decreases to an asymptote that is a fraction of its peak. This unusual behavior qualitatively follows the sensitivity to critical learning periods observed in biological systems, from cats to humans, as well as recently in deep artificial networks. Together with the Emergence Bound and the PAC-Bayes bound, this shows that forgetting is a necessary part of the learning process, and happens while precision increases monotonically. The information in the weights can also be used to defined a topology in the space of tasks, and an asymmetric distance that can be used to predict the cost and performance of fine-tuning a network trained for a different task, without actually performing the experiment. These phenomena collectively point to the importance of the dynamics of learning, and suggests that studying the transient behavior can yield insight beyond those that can be gleaned from the asymptotics. Depending on the context, we use Shannon, Fisher, or Kolmogorov information to prove the results described.