The Information Bottleneck theory of Deep Learning and the computational benefit of the hidden layers

Naftali Tishby, Hebrew University of Jerusalem

A proposed theory of large scale learning with Deep Neural Networks, based on the correspondence between Deep Learning and the Information Bottleneck framework, has several intriguing predictions. (1) Large scale learning requires rethinking Learning theory. In particular, we need to move from worse case distribution independent generalization bounds to distribution dependent but algorithm independent bounds. (2) For large scale Deep Neural Networks the mutual information of the encoder and optimal decoder of each hidden characterize the irrelevant information filtered out and the specific features encoded by each layer.
(3) Stochastic Gradient Descent, as used in Deep Learning, can push the layers to the Information Bottleneck optimal bound, through the weights diffusion in the second phase of training. (4) A direct prediction of this insight is that the convergence time (number of weight updates) to good generalization scales like a NEGATIVE power of the number of effective layers - more layers converge faster.
In this talk I will focus on the last point, which provides a new computational understanding of the effectiveness of Deep Neural Networks.
Based partly on works with Ravid Shwartz-Ziv, Amichai Painsky, and Noga Zaslavsky.