The Information Theory of Deep Learning - Rethinking large scale learning theory

Naftali (Tali) Tishby, The Hebrew University

Abstract

Deep Neural Networks pose new challenges to Computational learning theory. The much larger scale of the problem, both in terms of input size and parameter space, is forcing us to abandon the "old" worse case generalization bounds and look again at problem dependent but algorithm independent typical behavior theory. This type of analysis requires the use statistical physics or information theoretic tools which differ from uniform convergence tools. The information bottleneck theory of deep learning is an example of this approach. It gives new understanding and different type of computational bounds on sample and time complexity. Interestingly, the successes of Deep Learning stems mainly from the dynamic properties of stochastic gradient descent (SGD) in high dimension. It can also explain in a new way the importance of many hidden layers, which turns out to be more Computational (time complexity) than the expersivity of the Networks.