מכון ויצמן למדע

THE WEIZMANN INSTITUTE OF SCIENCE

FACULTY OF MATHEMATICS AND COMPUTER SCIENCE

Machine Learning and Statistics Seminar

Room 1 ,Ziskind Building
on Wednesday, Apr 19, 2017
at 11:15

Naftali Tishby
The Hebrew University of Jerusalem

A Deeper Understanding of Deep Learning

Abstract:

By analytical and numerical studies of Deep Neural Networks (using standard TensorFlow) in the "Information Plane" - the Mutual Information the network layers preserve on the input and the output variables - we obtain the following new insights.

1. The training epochs, for each layer, are divided into two phases: (1) fitting the training data - increasing the mutual information on the labels; (2) compressing the representation - reducing the mutual information on the inputs. The layers are learnt hierarchically, from the bottom to the top layer, with some overlaps.
2. Most (~80%) of the training time - optimization with SGD - is spent on compressing the representation (the second phase) - NOT on fitting the training data labels, even when the training has no regularization or terms that directly aim at such compression.
3. The convergence point, FOR EVERY HIDDEN LAYER, lies on or very close to the Information Bottleneck IB) theoretical bound. Thus, the mappings from the input to the hidden layer and from the hidden layer to the output obey the IB self-consistent equations for some value of the compression-prediction tradeoff.
4. The main benefit of adding more hidden layers is in the optimization/training time, as the compression phase for each layer amounts to relaxation to a Maximum conditional Entropy state, subject to the proper constraints on the error/information on the labels. As such relaxation takes super-linear time in the compressed entropy, adding more hidden layers dramatically reduces the training time. There is also benefit in sample complexity to adding hidden layers, but this is a smaller effect.

I will explain these new observations and the benefits of exploring Deep Learning in the "Information Plane", and discuss some of the exciting theoretical and practical consequences of our analysis.

Joint work with Ravid Ziv and Noga Zaslavsky.