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, Dec 12, 2018
at 11:15

Roy Schwartz
University of Washington

Towards Interpretable Deep Learning for Natural Language Processing

Abstract:

Despite their superb empirical performance, deep learning models for natural language processing (NLP) are often considered black boxes, as relatively little is known as to what accounts for their success. This lack of understanding turns model development into a slow and expensive trial-and-error process, which limits many researchers from developing state-of-the-art models. Customers of deep learning also suffer from this lack of understanding, because they are using tools that they cannot interpret. In this talk I will show that many deep learning models are much more understandable than originally thought. I will present links between several deep learning models and classical NLP models: weighted finite-state automata. As the theory behind the latter is well studied, these findings allow for the development of more interpretable and better-performing NLP models. As a case study, I will focus on convolutional neural networks (ConvNets), one of the most widely used deep models in NLP. I will show that ConvNets are mathematically equivalent to a simple, linear chain weighted finite-state automaton. By uncovering this link, I will present an extension of ConvNets that is both more robust and more interpretable than the original model. I will then present similar observations regarding six recently introduced recurrent neural network (RNN) models, demonstrating the empirical benefits of these findings to the performance of NLP systems.

This is joint work with Hao Peng, Sam Thomson and Noah A. Smith