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, Jan 09, 2019
at 11:15

Jonathan Belinkov
MIT

Deep Learning Models for Language: What they learn, where they fail, and how to make them more robust

Abstract:

Deep learning has become pervasive in everyday life, powering language applications like Apple's Siri, Amazon's Alexa, and Google Translate. The inherent limitation of these deep learning systems, however, is that they often function as a "black box," preventing researchers and users from discerning the roles of different components and what they learn during the training process. In this talk, I will describe my research on interpreting deep learning models for language along three lines. First, I will present a methodological framework for investigating how these models capture various language properties. The experimental evaluation will reveal a learned hierarchy of internal representations in deep models for machine translation and speech recognition. Second, I will demonstrate that despite their success, deep models of language fail to deal even with simple kinds of noise, of the type that humans are naturally robust to. I will then propose simple methods for improving their robustness to noise. Finally, I will turn to an intriguing problem in language understanding, where dataset biases enable trivial solutions to complex language tasks. I will show how to design models that are more robust to such biases, and learn less biased latent representations.