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 10, 2018
at 11:15

Yonatan Belinkov
MIT

Understanding Internal Representations in Deep Learning Models for Language and Speech Processing

Abstract:

Language technology has become pervasive in everyday life, powering applications like Apple's Siri or Google's Assistant. Neural networks are a key component in these systems thanks to their ability to model large amounts of data. Contrary to traditional systems, models based on deep neural networks (a.k.a. deep learning) can be trained in an end-to-end fashion on input-output pairs, such as a sentence in one language and its translation in another language, or a speech utterance and its transcription. The end-to-end training paradigm simplifies the engineering process while giving the model flexibility to optimize for the desired task. This, however, often comes at the expense of model interpretability: understanding the role of different parts of the deep neural network is difficult, and such models are often perceived as "black-box". In this work, we study deep learning models for two core language technology tasks: machine translation and speech recognition. We advocate an approach that attempts to decode the information encoded in such models while they are being trained. We perform a range of experiments comparing different modules, layers, and representations in the end-to-end models. Our analyses illuminate the inner workings of end-to-end machine translation and speech recognition systems, explain how they capture different language properties, and suggest potential directions for improving them. The methodology is also applicable to other tasks in the language domain and beyond.