Yoav Levine

Theoretical and practical principles for designing, training, and deploying huge language models

Abstract:

The field of natural language processing (NLP) has been advancing in giant strides over the past several years. The main drivers of this success are: (1) scaling the Transformer deep network architecture to unprecedented sizes and (2) "pretraining" the Transformer over massive amounts of unlabeled text. In this talk, I will describe efforts to provide principled guidance for the above main components and further thrusts in contemporary NLP, aimed to serve as timely constructive feedback for the strong empirical pull in this field. I will begin by describing our theoretical framework for analyzing Transformers, and present results on the depth to width tradeoffs in Transformers, identified bottlenecks within internal Transformer dimensions, and identified biases introduced during the Transformer self-supervised pretraining phase. This framework has guided the design and scale of several of the largest existing language models, including Chinchilla by Deepmind (70 billion learned parameters), Bloom by BigScience (176 billion learned parameters), and Jurassic-1 by AI21 (178 billion learned parameters). Then, I will describe our works on leveraging linguistic biases such as word senses or frequent n-grams in order to increase efficiency of the self-supervised pretraining phase. Subsequently, I will describe novel principles for addressing a present-day problem stemming from the above success of scaling, namely, how to deploy a huge language model such that it specializes in many different use cases simultaneously. Finally, I will comment on future challenges in this field, and will relatedly present a recent theoretical result on the importance of intermediate supervision when solving composite NLP tasks. This talk is based on works published in NeurIPS 2020, ACL 2020, ICLR 2021 (spotlight), ICML 2021, ICLR 2022 (spotlight), ICML 2022 (workshop), as well as on several recent preprints.