HKU-TCL Joint Research Centre for AI Workshop

Speaker:
Dr. Lingpeng KONG
Assistant Professor
Department of Computer Science, Faculty of Engineering, HKU

Title:
Pretraining Algorithms for Natural Language Processing

Abstract:

In the past two years, we've seen a sudden and significant leap in the performance of natural language processing (NLP) systems, with state-of-the-art models such as BERT, RoBERTa, GPT-3, among many others. These advances are enabled by the same technology called pretraining. The core idea of pretraining is to leverage a large amount of raw text to build a general model of language understanding. This model can then be easily adapted to specific NLP tasks, such as sentiment analysis and question answering.

However, relatively little research has shed light on the algorithmic side of pretraining, whilst recent advances are mostly seen as a triumph of large scale computation with a surge of the number of parameters and the size of the training data. The sequence-level language modeling (sLM) objective, despite its massive success, has apparent drawbacks and should not be simply taken for granted. Language is compositional and hierarchical in nature. The meaning of a sentence is constructed from smaller parts (e.g., words in a sentence) in a structured way. Language is also generated in a way that constantly reuses the knowledge acquired long ago---information beyond the sequence level. However, the sLM objective often fails to capture any of these, leading to naive mistakes (e.g., misreading the negation in a sentence).

In this talk, we will focus on the algorithmic side of pretraining by investigating two potential solutions to the drawbacks of the sLM objective. In the first half of the talk, we will discuss our newly proposed understanding the objective function of pretraining from the mutual information maximization perspective. Our new framework allows us to construct pretraining tasks around more complex views of the sentence (e.g., phrases or semantic parses) beyond solely the word-level. In the second half of the talk, we will focus on how we incorporate syntactic bias from a generative model (RNNG) into the pretraining stage by constructing a baseline employing distillation approach.

Bio of Speaker:

Dr. Lingpeng Kong is an assistant professor in the Department of Computer Science at the University of Hong Kong (HKU).

His research tackles the core problems in natural language processing (NLP) by designing representation learning algorithms that exploit linguistic structures. His work lies at the intersection of deep learning and structured prediction, with an application focus on syntactic parsing, speech recognition, social media analysis and machine translation. Before joining HKU, he was a senior research scientist at Google DeepMind. Dr. Kong obtained his Ph.D. from Carnegie Mellon University in 2017, co-advised by Noah Smith and Chris Dyer.