北京大学前沿计算研究中心
Center on Frontiers of Computing Studies, Peking University

静园5号院
青 年 讲 座

Towards Responsible Knowledge Sharing of Language Models

Dr. Chenguang Wang

Assistant Professor
Washington University in St. Louis

Host：王鹤 助理教授、张铭 教授
2023年4月24日 星期一 19:30–21:00
静园五院204室

Abstract

Pretrained large language models such as ChatGPT have significantly advanced text understanding over the last few years. These models share knowledge present in the training data via parameter sharing with downstream tasks (e.g., question answering and code completion). However, it is difficult to deploy these models in real-world applications, since the current knowledge sharing mechanism is not responsible for the following reasons. First, this mechanism shares hidden model knowledge that is not explainable. Second, it is not robust to distribution shifts arising in real-world tasks. Third, this also raises concerns for broader societal impacts, such as bias. In this talk, I will describe my research in responsible knowledge sharing of pretrained language models that solve the pressing problems. My talk will start by presenting my work in interpreting hidden knowledge in pretrained language models as human-readable knowledge. I will then introduce benchmarks and algorithms that enhance the robustness of knowledge sharing from those models. The talk will also discuss the applications of our responsible text understanding techniques to real-world scenarios. Finally, I will conclude with a vision of future directions for responsible knowledge sharing.

Biography

Chenguang Wang is an Assistant Professor in the Department of Computer Science and Engineering at Washington University in St. Louis, and the lead of the Natural Language Processing Group at Washington University in St. Louis. Previously, he was a postdoc in Computer Science at UC Berkeley. He received his Ph.D. degree from Peking University. His research interests span the areas of natural language processing, data science, security, and machine learning. His recent work is focused on responsible text understanding. He has created several impactful open-source systems, including GluonNLP and AutoGluon. He is the recipient of several academic awards such as ACM China Doctoral Dissertation Award Honorable Mention. His research has influenced real-world settings ranging from science to industry.

http: //cfcs.pku.edu.cn/