NJIT | INSTITUTE FOR DATA SCIENCE

Department of Data Science

Data Science Seminar Series

Towards Trustworthy Knowledge Sharing of Language Models

Chenguang Wang, Ph.D.
Postdoctoral Researcher
UC Berkeley

Date: Thursday, March 3rd, 2022
Time: 11:00 AM – 12:00 PM EST
Location: Zoom Virtual Room
Web Link: Zoom Meeting Room Link

Pretrained language models such as BERT and GPT-3 have revolutionized 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 trustworthy for the following reasons. First, this mechanism shares latent 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 trustworthy knowledge sharing of pretrained language models that solve pressing problems. My talk will start by presenting my work in interpreting latent 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 trustworthy text understanding techniques to real-world scenarios. Finally, I will conclude with a vision of future directions for trustworthy knowledge sharing.

Chenguang Wang is a postdoc in Computer Science at UC Berkeley. Before that, he was a Research Scientist at Amazon AI and a Research Staff Member at IBM Research-Almaden. He received his Ph.D. degree from Peking University in 2016. He was also a visiting Ph.D. student at UIUC. His research interests span the areas of data science, natural language processing, security, systems, and machine learning. His recent work is focused on trustworthy text understanding. He has created several impactful open-source systems, including GluonNLP (one of the most popular deep learning for NLP systems with 220,000 downloads) and AutoGluon (4,136 GitHub stars). He is the recipient of several academic awards such as ACM China Doctoral Dissertation Award Honorable Mention (one of the two national winners). His research has resulted in real-world impact and has been used by Amazon, Microsoft, UPMC Hillman Cancer Center.