For favour of Posting

THE HONG KONG UNIVERSITY OF SCIENCE & TECHNOLOGY

Department of Mathematics

SEMINAR ON DATA SCIENCE

AI Safety by the People, for the People

By

Prof. Hongyang ZHANG
University of Waterloo

Abstract

The rapid advancement in large language models has heightened the importance of AI safety, a concept that encompasses two critical dimensions: ensuring safety 'by the people' and 'for the people'. 'By the people' focuses on the responsibility of developers to use legitimate training data. 'For the people' emphasizes the alignment of AI models with core human values — helpfulness, harmlessness, and honesty

In the first part of tha talk, we will delve into strategies for addressing concerns regarding the legitimate use of data in AI training. A case study from March 2023 highlights OpenAI's challenge: demonstrating the legitimacy of their training data and logic while maintaining the confidentiality of ChatGPT's weights and data. We introduce zkDL (Zero-Knowledge Deep Learning), an innovative solution offering efficient zero-knowledge proofs for deep learning. This technology allows for the creation of proofs in a second for neural networks with 20M parameters, achieving a 2000× speedup on an NVIDIA's A100 GPU.

The second part of the talk explores how integrating self-evaluation and rewind mechanisms in unaligned large language models (LLMs) can produce outputs that resonate with human preferences through self-boosting. We present the Rewindable Auto-regressive INference (RAIN) framework, enabling pre-trained LLMs to assess their own outputs and utilize these evaluations to inform and refine their response generation. This innovative approach is notable for its ability to enhance AI safety without the need for additional alignment data, training, gradient computations, or parameter updates.

Biography

Hongyang Zhang is a tenure-track assistant professor at University of Waterloo and Vector Institute for AI. He received his PhD in 2019 from Machine Learning Department at Carnegie Mellon University and did a Postdoc at Toyota Technological Institute at Chicago. He is the winner of NeurIPS 2018 Adversarial Vision Challenge, CVPR 2021 Security AI Challenger, Amazon Research Award, WAIC Yunfan Award, etc. His paper on adversarial security ranks 13/6729 (top 0.19%) for its citations among all accepted papers in ICML 2018-2023. He also served as an area chair for NeurIPS, ICLR, AISTATS, AAAI, ACM CCS, and an action editor for DMLR.

Date : 30 November 2023 (Thursday)
Time : 11:00am
Venue : Room 1103 (near Lift 19)

All are Welcome!