COMPUTER SCIENCE & ENGINEERING
UNIVERSITY OF MICHIGAN

DISSERTATION DEFENSE

Haizhong Zheng

Bridging Data and Hardware Gap for Efficient Machine Model Scaling

Wednesday, October 16, 2024
3:30pm – 5:30pm
3725 Beyster
Hybrid – Zoom

ABSTRACT: The rise of large language models marks a watershed moment in artificial intelligence, led by the exponential growth of model sizes and data volumes. Despite our hope for further performance gains through scaling up model and dataset sizes, today's large models are reaching scaling limits in two aspects: First, large model training is data-hungry. The cost of creating large volumes of high-quality human feedback data is prohibitively expensive, creating bottlenecks in scaling up large model training. Second, naive reliance on more powerful hardware is inadequate, as hardware improves at a much slower rate than what the growth in model size demands.

In this talk, I will discuss my dissertation research in bridging the gap between the rapid scaling of models and the slower scaling in high-quality data and hardware. I will begin by introducing coreset algorithms we developed for data-efficient deep learning, highlighting the importance of data coverage in deep learning coreset selection. Our novel coverage-centric selection algorithm delivers up to a 5.2% accuracy improvement on coreset selection for ImageNet. Following this, I will introduce my recent research on label-free coreset selection, which aims to save human annotation effort in deep learning pipelines. Next, I will present our work on building contextual sparse models for inference efficiency. This involves designing training algorithms to create hardware-friendly and efficiency-aware contextual sparse models, aiming for fast model inference by treating inference efficiency as an optimization goal. Our algorithm achieves up to 50% structured contextual sparsity with only marginal performance drops. Finally, I will discuss the challenges of building more efficient and accessible AI in the future. I will conclude this talk with my vision and plan to address those challenges from model design, system optimization, and data collection perspectives.

CHAIR: Prof. Atul Prakash