Frontier of Artificial Network

Department of Data Science
香港城市大學
City University of Hong Kong

Extreme Discretization: Towards Efficient Intelligence and Systems in the Scaling Era

Invited Speaker

Haotong Qin

ETH Zurich

Date: December 17, 2025 (Wednesday)

Time: 16:00-17:30 (Hong Kong Time)

Zoom Meeting: 756 768 6492

Biography

Haotong Qin is a postdoctoral researcher at ETH Zurich, Switzerland. His research focuses on efficient AI/ML systems, including neural network compression, efficient large-scale models, low-power sensors, and embodied computing. He has published over 60 papers in top-tier conferences and journals such as IEEE TPAMI, IJCV, ICML, NeurIPS, ICLR, and CVPR, with more than 3,500 citations on Google Scholar. He has received numerous awards, including Electronics Best PhD Thesis, MLCommons/Meta ML and System Rising Star, IJCAI-GLOW Workshop Best Paper Award, Baidu PhD Fellowship, ByteDance PhD Fellowship, and the Yunfan Award at the World Artificial Intelligence Conference (WAIC). He also serves as Area Chair for major conferences such as CVPR, NeurIPS, and ACM MM, and as Guest Editor for journals including Neural Networks.

Abstract

Artificial intelligence (AI) is entering a scaling era, in which advanced models such as GPT now contain tens of trillions of parameters. While the continuously growing scale has endowed AI with extraordinary capabilities, it has also introduced severe efficiency challenges in terms of latency, memory, and energy consumption. Discretization is a key insight for breaking the efficiency bottleneck in AI. Techniques such as quantization, tokenization, and chain-of-thought represent forms of discretization that improve AI efficiency by decomposing and compressing computation, data, and reasoning processes. This talk will introduce our work on achieving efficient intelligence and systems through extreme discretization. These efforts aim to propose and refine theoretical paradigms, model architectures, and data characteristics related to discretization in AI. Beyond improving efficiency, these works require overcoming limitations in expressiveness and optimization of extreme discretization within finite, non-continuous representation spaces. The goal is to enable discrete AI systems that operate efficiently, sustainably, and accurately. Looking ahead, we aspire to realize the next generation of discrete AI in real-world settings, leveraging discrete reasoning, computation, and theory to advance AI efficiency and sustainability further.