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

COMPUTER SCIENCE COLLOQUIUM

Towards Prevalence of On-Device AI with Full Runtime Adaptability

SPEAKER **Prof Wei Gao**

Associate Professor
Department of Electrical and Computer Engineering University of Pittsburgh Pittsburgh, Pennsylvania, USA

DATE 27 Mar, 2025 (Thu)
TIME 10:30 AM - 11:30 AM
VENUE Y5204, 5th Floor, Yellow Zone, Yeung Kin Man Academic Building, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon Tong

ABSTRACT

With the recent democratization of AI, there is a pressing need of supporting AI on mobile and embedded devices at the edge, to allow intelligent and prompt decision making autonomously on these devices. To meet the devices' constraints in computing capacity, current software solutions to on-device AI reduce the ML model's complexity, but have major weaknesses in adapting to the changes of online data patterns and environmental contexts, resulting in significant reduction of model performance in difficult learning tasks. In this talk, I will present our recent research on achieving such full runtime adaptability, as a key enabler for prevalence of on-device AI in practical systems. I will first present how we leverage explainability in AI to adaptively involve the most appropriate model structures for on-device computations, so as to support real-time inference, runtime training and LLM fine-tuning on devices with extreme resource constraints. Afterwards, I will further show how such on-device AI techniques can be applied to various application domains, including smart healthcare and embodied AI systems, to achieve high system performance with heterogeneous data characteristics and diverse environmental settings.

BIOGRAPHY

Wei Gao is currently an Associate Professor in the Department of Electrical and Computer Engineering, University of Pittsburgh. His research interests lie in the intersection between AI and computer systems, with a focus on the design and deployment of on-device AI models and algorithms on mobile, embedded and networked devices. He also has strong interests in applying the computationally efficient AI models into practical application domains to make societal impacts and benefit the human welfare. He has published more than 80 research papers at both top AI and system conference venues, including ICLR, AAAI, CVPR, ASPLOS, MobiCom, MobiSys, SenSys, etc, and received multiple best paper awards or nominations.

All are welcome!

In case of questions, please contact Prof Yuguang Michael Fang at my.fang@cityu.edu.hk, or visit the CS Departmental Seminar Web at https://www.cs.cityu.edu.hk/events/cs-seminars/recent-cs-colloquiums.