CECS
CENTER FOR EMBEDDED & CYBER-PHYSICAL SYSTEMS
UNIVERSITY OF CALIFORNIA · IRVINE

CECS Seminar

"Sparse Computing and Large Language Model: from AI 1.0 to AI 2.0"

Guohao Dai

Associate Professor at
Shanghai Jiao Tong University

Friday, November 3rd
10:30-11:30 p.m. PST
Location: ISEB 4020

Abstract: General Artificial Intelligence is going through a new development stage from Computation AI, Perception AI, to Cognition AI, with the new trends of big data, multi-modality, and multi-task. However, due to the slowdown of Moore's Law, there are bottlenecks in hardware computility, leaving a huge gap between the supply of computility and the demand for the computility in intelligent computing. Because the AI function has the learnable characteristics, the amount of calculation can be reduced through the sparsification method, thereby breaking through the hardware computility bottleneck of intelligent computing while ensuring task accuracy. This report will focus on the software and hardware co-design method of sparse computing. This report will introduce the definition and process of sparse computing, as well as the challenges faced in intelligent computing from algorithm, compilation, and architecture perspectives. This report will elaborate on the multi-level design method for sparse computing in AI and introduce three typical models in the AI 1.0 era (Graph Neural Networks, Point Cloud Networks, and Conventional Convolution Neural Networks). Finally, this report takes the Large Language Models (LLMs) algorithm in the AI 2.0 era as an example to illustrate the application of the above design method in large language model inference optimization. Compared with the existing model, the calculation speed on dense hardware architecture is increased by 2 times.

Biography: Guohao Dai is a tenure-track Associate Professor at Shanghai Jiao Tong University, he received his B.S. and Ph.D. (with honor) degrees from Tsinghua University, Beijing, in 2014 and 2019. Prof. Guohao Dai has published more than 50 high-level international journals and conference papers in the fields of Electronic Design Automation (EDA), heterogeneous computing, and system/architecture design. He has received more than 1000 citations in Google Scholar. The published paper won the ASP-DAC 2019 Best Paper Award, DATE 2023/DAC 2022/DATE 2018 Best Paper Nomination, and WAIC 2022 Outstanding Youth Paper Award. He has personally won honors such as the WAIC 2022 Yunfan Award, the global champion of the NeurIPS21 BIGANN competition, Beijing's outstanding doctoral graduates, Tsinghua University's outstanding doctoral graduates, and Tsinghua University's outstanding doctoral thesis. Professor Dai has participated in guiding students to rank third in the world in ACM 2021 SRC and first in the world in MICRO 2020 SRC. Currently, he serves as PI/Co-PI for several projects with a personal share of over RMB 10 million.

Hosted By: Sitao Huang