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

CECS SEMINAR



"Optimizing AI Performance: Bridging Algorithmic and Hardware Innovations from Inference to Training"

Professor Yu Bai

Associate Professor and Director of the Intelligent Computing Research Lab at California State University, Fullerton

Wednesday, March 12th
2:00-3:00 p.m.
Location: ISEB 1200

Abstract: Artificial intelligence (AI) continues to reshape industries and research, but its rapid growth comes with significant computational and energy costs. This talk explores the challenges and innovations in optimizing AI efficiency, bridging algorithmic advancements and hardware design to enhance performance from inference to training. The presentation covers two main avenues for improving AI efficiency: model optimization techniques and hardware specialization. Algorithmic strategies such as low-rank tensor decomposition, structured sparsity, quantization-aware training, and compression of large-scale models (e.g., LLMs and diffusion models) are discussed. These methods aim to reduce memory requirements, computational overhead, and energy consumption while maintaining model accuracy. On the hardware side, the talk highlights cutting-edge AI accelerators, dataflow optimizations, and emerging techniques like in-storage computing for AI hardware. Novel architectures such as the Tensor-Train-based Inference and energy-efficient training solutions are presented, demonstrating substantial improvements in model compression, multi-modal AI, and real-time generative AI tasks. The talk also touches on the broader implications of AI efficiency, including its impact on embedded AI platforms, cloud-based AI, and future developments in efficient AI hardware and software co-design. Through interdisciplinary collaboration and Industry and NSF-supported research, this work aims to advance the next generation of AI technology, making it both powerful and sustainable.

Biography: Dr. Yu Bai is an Associate Professor and Director of the Intelligent Computing Research Lab at California State University. His research focuses on emerging AI techniques, energy-efficient AI systems, and interdisciplinary AI applications, spanning neuromorphic computing, generative AI, deep learning for IoT devices, and high-performance computing. Dr. Bai has authored over 50 peer-reviewed publications in top AI conferences and journals, contributing significantly to efficient neural network optimization, AI hardware-software co-design, and hyperspectral imaging. His work has been recognized with numerous accolades, including the Best Researcher Award (2023) from Citation Award, the Faculty Scholarly and Creative Activity Award (2021), and inclusion in Who's Who in America. With a proven track record in securing research funding, Dr. Bai has received over $4 million in competitive grants from agencies such as the National Science Foundation (NSF), the Army Research Office, and industry-academic collaborations. His funded projects include pioneering efforts in edge AI architectures, AI applications in biological systems, and inclusive AI education strategies. A National Academies of Sciences, Engineering, and Medicine (NASEM) awardee, Dr. Bai is dedicated to advancing AI research and fostering a diverse, inclusive STEM community through his teaching, mentorship, and collaborative initiatives.

Hosted By: Prof. Fadi Kurdahi and Prof. Mohammad Al Faruque