Artificial Intelligence Initiative (Aii)

https://ai.ucf.edu/



Invited Speaker

Mr. Yuzhang Shang, Illinois Institute of Technology

"Towards AI Democratization: From Efficient Large AI Models to Network Compression Theory"

Monday, March 10, 2025 • 11:00AM • HEC 101A • Zoom

Abstract:
Artificial Intelligence (AI) models are growing increasingly complex and resource-intensive, posing significant challenges for widespread deployment. This research focuses on efficient AI as a crucial pathway towards democratizing large AI models, making their benefits accessible to a global audience. This presentation will explore three key areas of efficient AI: (i) Accelerating Generative Models: I will discuss novel training-free methods to enhance the efficiency of state-of-the-art generative models, specifically diffusion models. (ii) Compressing Large Language Models: My research addresses the acceleration of large language models, improving their performance and accessibility by implementing low-rank decomposition on LLMs. (iii) Neural Network Compression Theory: my research goals are not only to propose effective approaches for deep neural network compression, but also to explore the possibility towards well-defined and explainable compression methods. By improving the efficiency of these critical AI domains, we aim to lower the barriers to entry for AI adoption and foster innovation across diverse fields and industries.

Bio:
Yuzhang Shang is a final-year Ph.D. candidate in Computer Science at the Illinois Institute of Technology (IIT) under the supervision of Prof. Yan Yan. His research focuses on Efficient/Scalable AI. During his Ph.D., he spent a year as a visiting student at the University of Wisconsin-Madison hosted by Prof. Yong Jae Lee. He was a Research Scientist Intern at Google DeepMind, working on training inference-efficient large language models. He also interned as a Research Scientist at Cisco Research, where he developed efficient computer vision models and created the model slimming toolbox for Cisco.