Invited Talk

Accelerating Artificial Intelligence for Data-Driven Discovery

Speaker: Professor Shih-Chieh Hsu

Professor of Physics and Adjunct Professor of
Electrical and Computer Engineering,
University of Washington

Date : December 25th (Wednesday), 2024

Time : 10:00AM – 11:00AM

Place: Room EE-106, Department of Electrical Engineering, NCHU

Abstract:
As scientific data sets become progressively larger algorithms to process this data quickly become more complex. In response Artificial Intelligence (AI) has emerged as a solution to efficiently analyze these massive data sets. Emerging processor technologies such as graphics processing units (GPUs) and field-programmable gate arrays (FPGAs) allow AI algorithms to be greatly accelerated. The Accelerated AI Algorithms for Data-Driven Discovery (A3D3) Institute sponsored by the National Science Foundation under the Harnessing the Data Revolution program is established to enable real-time AI at scale for broad applications. In this talk I will give an overview about the challenges of high energy physics, multi-messenger astrophysics and neuroscience regarding AI across latency and throughput regimes. I will introduce various techniques for model compression using state-of-the-art techniques such as pruning and quantization for edge computing. I will demonstrate that that acceleration of AI inference as a web service represents a heterogeneous computing solution. Finally I'll discuss how A3D3 can bring together disparate communities that are threaded by common data-intensive grand challenges to accelerate discovery in Science and Engineering.

Bio:
Shih-Chieh Hsu (徐士傑) is a Professor of Physics and Adjunct Professor of Electrical and Computer Engineering at the University of Washington. He directs the NSF HDR Institute for Accelerated AI Algorithms for Data-Driven Discovery (A3D3 - https://a3d3.ai). With degrees from National Taiwan University and UC San Diego, Dr. Hsu specializes in experimental particle physics focusing on dark matter searches, neutrino measurements and AI applications in data-intensive research. His work utilizes the Large Hadron Collider and incorporates real-time AI for rapid data analysis and decision-making across multiple scientific discipline. He has received recognition for his innovative research, mentorship, and contributions to real-time AI applications in scientific discovery.