NJIT | INSTITUTE FOR DATA SCIENCE

Data Science Seminar Series

In Collaboration with
The Department of Data Science

Towards Ultimate Efficiency in Ubiquitous ML Powered Intelligence and Green AI

Hosted by Shuai Zhang

Peiyan Dong

Northeastern University

Date:	**Wednesday, February 21, 2024**
Time:	**2:30 PM - 3:30 PM (Coffee served at 2:15 PM)**
Location:	**GITC Building Room 4402 (4th floor Seminar Room)**
Web Link:	**Zoom Meeting Link**

As AI techniques continue to advance, the efficient deployment of deep neural networks on resource-constrained devices becomes increasingly appealing yet challenging. Simultaneously, the proliferation of powerful AI technologies has raised significant concerns about sustainability and fairness, demanding increased attention from the community. This talk presents two novel software-hardware co-designs for improving the efficiency and sustainability of deep learning models. The first part introduces a hardware-efficient adaptive token pruning framework for Vision Transformers (ViTs) on embedded FPGA, HeatViT, which achieves significant speedup under similar model accuracy compared to the state-of-the-art. HeatViT is the first end-to-end accelerator for ViT on embedded FPGA and also achieve practical speedup by data-level compression for the first time. The second presents PackQViT and Agile-Quant, a paradigm of the efficient implementation for transformer-based models by sub-8-bit packed quantization and SIMD-based optimization for computing kernels. Our framework can achieve better task performance than state-of-the-art ViTs and LLMs with significant acceleration on edge processors, such as mobile CPU, Raspberry Pi and RISC-V. This work not only marks the first successful implementation of the LLM on the edge but also addresses the previous limitation where edge processors struggled to efficiently handle sub-8-bit computations. At the conclusion of the presentation, the speaker will discuss today's challenges related to AI sustainability and fairness and outline her research plans aimed at addressing these issues.

Peiyan (Peggie) Dong is a final-year Ph.D. Student at Northeastern University, Boston, advised by Prof. Yanzhi Wang. Her research area is the intersection of Software-Hardware Co-design, Hardware Architecture, and Efficient Emerging Devices, such as superconducting devices and quantum circuits. Her work has been published broadly in top conference and journal venues (e.g., DAC, ICCAD, MICRO, HPCA, ICS, ISSCC, AAAI, ICML, NeurIPS, CVPR, IJCAI, ECCV, RTAS, TCAS-I, TCAD, etc.) She has received Rising Star Award in EECS 2023, three Oral Paper Awards, one Spotlight Paper Award, and also the inventor of one U.S. patent.