

Wentao Hou

PHD STUDENT · UNIVERSITY OF WISCONSIN-MADISON

1210 W. Dayton Street, Madison, WI 53706

✉ taoh@cs.wisc.edu | 🏠 thoment.github.io | 📄 github.com/Thoment | 🔗 www.linkedin.com/in/wentao-hou-55825a220

Education

University of Wisconsin-Madison

PHD IN COMPUTER SCIENCES

- Advisor: Ming Liu

Madison, United States

Sep 2022 - present

Tsinghua University

BE IN ELECTRONIC ENGINEERING

Beijing, China

Sep 2018 - Jun 2022

Publications

Wentao Hou*, Kai Zhong*, Shulin Zeng, Guohao Dai, Huazhong Yang, Yu Wang, "NTGAT: A Graph Attention Network Accelerator with Runtime Node Tailoring", 28th Asia and South Pacific Design Automation Conference (ASP-DAC 2023) (Accepted)

Awards and Honors

2022 **Scholarship for excellence in the academy (top 30%)**, Tsinghua University

Dec. 2018 **35th National College Physics Competition (group of non-physics major), First prize**,
Beijing Institute of Physics

Research Experience

Tsinghua University - Dept of Electronic Engineering

ADVISOR: PROF. YU WANG

Beijing, China

Oct 2021 - Jul 2022

- Proposed a runtime node tailoring algorithm based on attention coefficients sorting to accelerate graph attention network.
- Proposed a hardware-efficient pipeline insertion sorting scheme for fast node tailoring.
- Designed an accelerator architecture and dedicated processing units for Graph Attention Convolution.
-

University of Virginia - Department of Computer Science

ADVISOR: PROF. SAMIRA KHAN

Remote

Jun 2021 - Oct 2021

- Conducted a breakdown performance evaluation of pre-processing in DNN with an image dataset.
- Measured the overheads of different pre-processing steps and looked for bottlenecks in certain scenarios.
- Profiled overheads of differential privacy in machine learning.

Tsinghua University - Dept of Electronic Engineering

ADVISOR: PROF. YU WANG

Beijing, China

Dec 2020 - Oct 2021

- Worked on a software and hardware co-designed GNN accelerator which optimizes loading scattered features.
- Wrote several modules of the GNN and implemented a data path between host, device and on-chip memory via openCL and AXI channel in Xilinx SDx environment. Measured the bandwidth and delay of reading features over different granularities.
- Found the bandwidth saturates at 1KB per addressing in continuous memory access.

Tsinghua University - Dept of Electronic Engineering

Beijing, China

ADVISOR: PROF. YONGPAN LIU

Dec 2020 - Oct 2021

- Simulated a CNN accelerator with gem5. Simulated sparse acceleration by rounding small weights and skipping all-zero weight groups.
- Simulated a binary neural network on RRAM array by modifying with gem5. Simulated the effects of random noise in RRAM, and measured the relation between noise amplitude and accuracy. Designed an algorithm to reuse weights on different layers when on-chip memory is enough.