Yannan (Nellie) Wu

606 Antioch Terrace, CA, 94085
Personal website: *https://nellie-wu.github.io*
Email : nellieywu@gmail.com
Mobile : +1-607-379-2186

Education

Masssachusetts Institute of Technology — Cambridge, MA
Ph.D. *in Computer Science (GPA: 5.0/5.0)* — ***June 2023***
M.S. *in Computer Science (GPA: 5.0/5.0)* — ***Feb. 2020***
Advisors: Prof. Joel Emer & Vivienne Sze

Cornell University — Ithaca, NY
B.S. *in Electrical & Computer Engineering (GPA: 4.02/4.3; 4.0=A)* — ***May 2017***

Summary and Objectives

I am a machine learning accelerator modeling engineer at Google. Before joining Google, I obtained my Ph.D. from MIT in computer architecture and systems. I have extensive research experience in co-designing energy-efficient hardware accelerators for deep neural networks, in both academic and industrial settings. My works have led to significant contributions to open-source industrial code bases, publications/tutorials at top-tier conferences/journals.

Selected Work Experience

- **Google TPU Modeling Engineer** — June, 2023 -
  - Developed analytical modeling infrastructure for various components in TPU chips.
  - Developed large language model training benchmarks in Tensorflow to aid TPU performance modeling.
  - Performed co-design studies to improve TPU compute performance for large language models.
- **NVIDIA Computer Architecture Research Intern** — May 2021 - Aug. 2021; May 2020 - Aug. 2020
  - Developed a statistical approach to analytically model various sparse deep learning workloads' energy consumption.
  - Enhanced Ampere GPU's tensor core support to accelerate more structured sparsity patterns and degrees.
  - Developed pruning and fine-tuning procedures using PyTorch to realize various sparsity structures.

Skills

- C++, Python, Open-source Management(with Git), Docker, MATLAB, C, Verilog, Linux, LaTex, Markdown, HTML

Selected Patents and Publications

- **HighLight: Efficient and Flexible DNN Acceleration with Hierarchical Structured Sparsity**
  Yannan Nellie Wu, Po-An Tsai, Saurav Muralidharan, Angshuman Parashar, Vivienne Sze, Joel S. Emer
  IEEE/ACM International Symposium on Microarchitecture (MICRO), Oct. 2023
- **Sparseloop: An Analytical Approach to Sparse Tensor Accelerator Modeling**
  Yannan Nellie Wu, Po-An Tsai, Angshuman Parashar, Vivienne Sze, Joel S. Emer
  IEEE/ACM International Symposium on Microarchitecture (MICRO), October 2022 ***(Distinguished Artifact Award)***
- **An Architecture-Level Energy and Area Estimator for Processing-In-Memory Accelerator Designs**
  Yannan Nellie Wu, Vivienne Sze, Joel S. Emer
  IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS), April 2020
- **Accelergy: An Architecture-Level Energy Estimation Methodology for Accelerator Designs**
  Yannan Nellie Wu, Joel S. Emer, Vivienne Sze
  IEEE/ACM International Conference on Computer-Aided Design (ICCAD), Nov. 2019
- **Pruning and Accelerating Neural Networks with A Novel Sparsity Structure**
  Yannan Wu, Po-An Tsai, Saurav Muralidharan, Joel S. Emer
  US Patent Application Number: 63/236,629

Conference Tutorials

- **ISCA21 Tutorial: Sparse Tensor Accelerators: Abstraction and Modeling**
  Yannan Nellie Wu with Joel S. Emer, Vivienne Sze, Po-An Tsai, and Angshuman Parashar
- **ISCA20, ISPASS20, MICRO19 Tutorial: Tools for Evaluating DNN Accelerator Designs**
  Yannan Nellie Wu with Joel S. Emer, Vivienne Sze, Angshuman Parashar, and Po-An Tsai

Selected Awards

- MICRO22 Distinguished Artifact Award — Oct. 2022
- MIT Jacob's Presidential Fellowship — Sept. 2017 - May. 2018
- Cornell ECE Early Career Scholarship — June. 2014 - Aug. 2014