

Peiming Yang

University of Toronto

✉ peiming.yang@mail.utoronto.ca

🌐 github.com/ypm1999

☎ +1 2498728733

RESEARCH INTERESTS

○ Machine Learning System

○ Computer System

○ Computer Architecture

EDUCATION

ACM Honor Class, Shanghai Jiao Tong University

Bachelor of Computer Science at Zhiyuan College, GPA 87.4 overall, 92.5 for last year

Member of ACM Honor Class, an elite CS program for top 5% talented students.

Shanghai

Sep 2017 — Jun 2021

University of Toronto

Master of Applied Science in Electrical & Computer Engineering, GPA 4.0/4.0

Toronto

Sep 2021 — Aug 2023

University of Toronto

PhD of Computer Science, with Professor Gennady Pekhimenko

Toronto

Sep 2023 — Present

PUBLICATIONS

Horizontally Fused Training Array: An Effective Hardware Utilization Squeezer of Training Novel Deep Learning Models

Shang Wang, **Peiming Yang**, Wentao Wang, Yan Hong, Liqing Zhang

4th Conference on Machine Learning and Systems

Image Editing via Segmentation Guided Self-Attention Network

Jianfu Zhang, **Peiming Yang**, Wentao Wang, Yan Hong, Liqing Zhang

IEEE Signal Processing Letters (Volume: 27)

AWARDS AND COMPETITIONS

Zhiyuan Honorary Scholarship, top 5% of 17,000 students in SJTU

2017 - 2021

China Collegiate Programming Contest, Gold Award, top 4% of 250+ teams

Qinhuangdao, 2017

International Collegiate Programming Contest(ACM-ICPC), Sliver Award

Maynila, 2017

The 33nd China National Olympiad in Informatics(NOI), Sliver Award

Mianyang, 2016

EXPERIENCE

EcoSystem Lab, University of Toronto

2021 - Present

- Design a new cache layer with data compression for the filesystem to keep all metadata staying in the memory. After Compression, the slow disk operation is removed from the critical path and applications like ML and big data are allowed to work with huge datasets. It is transparent for users and doesn't require any data movement.
- Making real machine leaning jobs running on REAL Processing-In-Memory(PIM) architecture more efficient and convonient, developing PIM-related tools for graph neural network(GNN) and large language mode(LLM)

Research Internship - EcoSystem Lab, University of Toronto

2020

- Propose Horizontally Fused Training Array, which optimize speed and memory usage for hyper-parameter search by fusing operators in neural network. It gains $1.33\times$ - $4.88\times$ speedup and 33%-86% memory saving. This paper has been accepted by MLSys 2021.

Research Internship - Brain-like Computing and Machine Intelligence Lab, SJTU

2019 - 2020

- Use segmentation guided self-attention neural network to simplify image editing. Our model combines segmentation information and hand-drawn sketch for erased parts to generate a new picture.

Teaching Assistant - Data Structures

2019

Teaching Assistant - Introduction to Computer Science

2018

SKILLS

- **Proficient in Python and C++** Familiar with Python and C++ programing. Being good at basic data structures like list, dict and set in python and STL library in C++.
- **Data structures and Algorithm** Being good at basic data structures and algorithms. Familiar with heap, stack, link list and binary search tree. Have solved more than 1000 problems from leetcode, codeforces and other online judge websites.
- **Digital circuit design** Familiar with FPGA and Verilog. And I have designed a RISCv CPU in the course project.
- **Knowledge of processor architecture** Familiar with multi-stage CPU design. And have finished a RISCv CPU with verilog.
- **Fundamentals of Image processing** Knows basic image processing methods including traditional image filters and advanced Machine Learning methods.

PROJECTS

HFTA [\[code\]](#)

[Website](#)

- By horizontally fuse the same operator in parallel running models, it gives up to 4.88x speedup and 86% memory saving on model training .

MX-Compiler [\[code\]](#)

(Score 99/100) Course project

- A simple complier of C-like language with NSAM assembly output. I implemented several optimization so that it is **almost 2 times faster than 'gcc -O1'**.

RISCv CPU [\[code\]](#)

Course project

- A cpu implemented by **Verilog**, with **RISCv-32I** instruction set. The cpu uses five-stage pipeline architecture, with 1KB instruction cache, running on Basys3 FPGA board.