HUNG-YANG CHANG

hychangee@gmail.com, james.chang@cerence.com (1) 438-509-9099 LinkedIn
Github Google scholar Montreal/Brossard, QC, Canada

PROFESSIONAL EXPERIENCE

Cloud Platform Innovations team, Cerence (Hybrid) Montreal, Canada
Senior AI software developer, full-time permanent *Aug. 2023 - now*

- Developed and integrated LLM function call algorithms (Back-end) with automotive console UX/UI (Front-end) to enhance in-car user experience
- Designed and implemented a robust testing framework for LLM function calls, including unit tests, contract tests, and user acceptance tests to ensure system reliability and performance

Bittensor (Open-source ML), Opentensor (Remote) Ontario, Canada
Machine Learning Engineer, full-time contract with Mining team *Mar. 2023 - Jul. 2023*

- Fine-tuned multiple language models and built an ensemble model of them on Bittensor project , outperforming GPT-4 and other models on text prompting based on Bittensor reward mechanism
- Built multi-modality of the Bittensor network, which includes text-to-image, text-to-video, text-to-music, and more subnets

McGill Edge Intelligence Lab, McGill Quebec, Canada
Graduate Research Assistant, advised by Professor Warren Gross, full-time *Sep. 2020 - Feb. 2023*

- Proposed a pipeline framework to utilize the heterogeneous resources available in edge device, achieving an average 49% of higher throughput and 61% of lower energy-delay product in edge BERT inference than the best homogeneous configuration JSPS'22
- Integrated Neural Architecture Search and pipeline on BERT model, achieving 9x higher inference throughput with only a 1.3% decrease in accuracy in edge BERT inference than the best homogeneous configuration EIW'22 GLSVLSI'23

Neuromorphic Devices and Architectures Research Group, IBM San Jose, CA, USA
Research Intern, mentored by Dr. Geoffrey W. Burr, full-time *Oct. 2018 - Apr. 2019*

- Analyzed power behavior of the circuit and modified the power-hungry structure to achieve up to 12 to 14 TOPs/s/W energy efficiency for training. IBM Journal of R&D'19

DSML Group, National Chiao Tung University Hsinchu, Taiwan
Research Assistant, advised by Chair Professor Steve S. Chung *Oct. 2017 - Oct. 2018*

- Built ideal-linearity neuromorphic synapses on FinFET Platform with a wide tuning-window (20×) of weight-tuning capability Symposia on VLSI'19

Signal Sensing and Application Lab, NTHU Hsinchu, Taiwan
Undergraduate Researcher, advised by Professor Chih-Cheng Hsieh *Feb. 2017 - Feb. 2018*

- Taped-out chip of CMOS image sensor readout circuit Chip Report
- Cooperated with Industrial Technology Research Institute (ITRI) with USD 30,000 project funding

EDUCATION

McGill University, Canada *Sep. 2020 - Feb. 2023*
MSc (thesis) in Electrical & Computer Engineering Overall GPA: 4/4

National Tsing Hua University, Hsinchu (NTHU), Taiwan *Sep. 2014 - Jun. 2018*
B.S. in Electrical Engineering Overall GPA: 3.99/4.3 (3.87/4)

SELECTED PUBLICATION

PipeBERT: High-throughput BERT Inference for ARM Big.LITTLE Multi-core Processors, Journal of Signal Processing Systems, IEEE SiPS 2022

- Hung-Yang Chang, Seyyed Mozafari, Cheng Chen, James Clark, Brett Meyer, and Warren Gross

AI hardware acceleration with analog memory: micro-architectures for low energy at high speed, IBM Journal of Research and Development

- Hung-Yang Chang and Geoffrey W. Burr, Pritish Narayanan, Stefano Ambrogio, et al.

High-Throughput Edge Inference for BERT Models via Neural Architecture Search and Pipeline, GLSVLSI 2023 (Poster Presentation)

- Hung-Yang Chang, Seyyed Mozafari, James Clark, Brett Meyer, and Warren Gross

A Novel Architecture to Build Ideal-linearity Neuromorphic Synapses on a Pure Logic FinFET Platform Featuring 2.5ns PGM-time and 10^{12} Endurance, 2019 Symposium on VLSI Technology (Oral Presentation)

- E.R Hsieh, H. Y. Chang, Steve S. Chung, S. Simon Wong et al.

SELECTED PROJECTS

Hardware Aware Efficient Training Competition *Feb. 2021 - Mar. 2021*
ICLR 2021 Workshop *ECSE, McGill*

- Built a modified ResNet with mixed-precision training to achieve 95%, 77% validation accuracy for CIFAR-10 and CIFAR-100, respectively in 5 minutes with single V100 GPU

Exploring Super-Converge in Analog NNs with IBM tool *Feb. 2021 - Apr. 2021*
Deep Learning (ECSE 552) *ECSE, McGill*

- Explored super-convergence phenomena in IBMs Analog Hardware Acceleration Kit for in-memory training of DNNs
- Applied cyclic learning rates to VGG8, ResNet18, and LeNet architectures on MNIST and CIFAR10

System C implementation: Design of MPSoC *Oct. 2020 - Dec. 2020*
Design of Multiprocessor System-on-chip (ECSE 541) *ECSE, McGill*

- Utilized IBMs Analog Hardware Acceleration Kit and PyTorch to simulate in-memory FCNN & CNN computations for MNIST Dataset

AWARDS & HONORS

Graduate Excellence Fellowship ECSE, McGill, 2022

Awarded with 4600 CAD for 10 selected graduated students

Outstanding Project Award in Contest of Implementation EECS, NTHU, 2018

- Top 10 of Research project competition with more than 250 student competitors

International Volunteer Certification Ministry of Education Taiwan, 2015

- Awarded with $1000 USD funding to host 100 people classes, and school anniversary fair in Malaysia

RELATED SKILLS

Programming Language: Python, Pytorch, Pytest, Tensorflow, TVM, Matlab, SystemC, LaTeX
Engineering Tools: DevOps, gPRC, Iceberg Catalog, Flask, Docker, Kubernetes, Jira,