Huang Hantao

Phone: +86 13162605119
Email: hhuang013@e.ntu.edu.sg, Huang Hantao - Google Scholar

Personal Statements

I am the co-founder of a startup company, Metachip, which is currently valued at 10 million RMB. Prior to this, I worked as a research associate professor at the Southern University of Science and Technology (SUSTech). Before my work at SUSTech, I held positions as an algorithm department manager at Canaan in Shanghai and MediaTek in Singapore. My expertise lies in neural network hardware-software co-design, quantization algorithms, sparse algorithms, and FPGA implementations (from algorithms to hardware architecture). I have authored a book titled "Compact and Fast Machine Learning Accelerator for IoT Devices", as well as over 40 papers published in top-tier IEEE/ACM journals and conferences. Additionally, I filed four US patents.

Education Background

Ph.D. from Nanyang Technological University, Singapore 2014.01 – 2018.05

- Thesis：Compact and Fast Machine Learning Accelerator on IoT Devices.
- Advisor：Prof. Yu Hao, Prof. Goh Wang Ling

First-class bachelor's degree from Nanyang Technological University, Singapore 2009.08 – 2013.07

- PRC-Singapore Ministry of Education Scholarship (SM3)

Visting Scholar from Georgia Technology University, USA. 2017.06 – 2017.12

- Topic: 3D-IC AI Accelerator for Binary Neural Networks
- Advisor: Prof. Sung-Kyu Lim

Working Background

Start-up company Meta chip (Research Associate Professor in Southern University of Science and Technology, Shenzhen China) **Shenzhen**

Fundraising Startup seed round: 2023.07 – Now

- Raised 5 million RMB (Shenzhen Government Subsidiary), 2 million Shenzhen Leaguer (力合科创) and 4 million RMB signed from other investment fund.

Large Language Model Optimization and customization 2023.07 – Now

- Adopt Retrieval Augmented Generation (RAG) for LLMs to support economic teaching course. Using an external knowledge base (course PPT/PDF) to ground ChatGLM3-6B on the most accurate, up-to-date information and to give users the answers to course-related questions.
- Design post-training quantization (APTQ) and pruning algorithms (Sparse GPT) for large language models, advanced from GPTQ and achieve mixed bit quantization (4bit/2bit) on ChatGLM-6B and Llama-7B models.DAC'24 paper is submitted.
-

Multi-precision AI Chip Design 2023.07 – Now

- Support AI algorithm and architecture design and achieve a 29.12 TOPS/W and 1.13 TOPS/mm2 NAS-Optimized Mixed-Precision DNN Accelerator with Vector Split-and-Combination Systolic in 28nm CMOS and published in CICC'24.
- Architecture design for large language models and adopts 4bit weights and 16-bit activation to achieve high DRAM bandwidth utilization on FPGA Xilinx VCU-128.

Canaan, AI Algorithm, Department Manager **Shanghai**

Large Language Models 2022.12 – 2023.07

- Design customized large language model with consideration of commercial license and performance, develop Bloomz-3B/7B fine-tuned large language models. Our Bloomz-7B model achieves the same performance as ChatGLM-7B. Open sourced at kendryte/Toucan-LLM: Self-trained Large Language Models based on Meta LLaMa (github.com)

Quantization and Pruning 2022.9 – 2023.07

- Develop a smooth quant based transformer/large language model quantization strategy and achieve 4 times model compression with nearly 0 loss of accuracy.
- Develop post training quantization and sparse pruning tool and achieve 4/8/16 bit mixed quantization, 4:2 semi-structured pruning and channel pruning.

Smart Dictionary Pen (Translation) 2021.9 – 2023.07

- Develop neural machine translation algorithm for Chinese-to-English and English-to-Chinese and achieves comparable to Baidu commercial solutions.
- Develop text to speech algorithm based on Fastspeech-v2 and Hifi-Gan and achieves MOS score 4.4 of 5.

MediaTek, System Platform AI Algorithm Manager **Singapore**

Low bit and Mixed Bit Quantization tool 2017.12 – 2022.09

- Develop quantization tool in NeuroPilot SDK to support 4/8/16 mixed bit optimization and support Tensorflow V1, V2 and PyTorch.
- Support application deployment using 4/8 mixed bit quantization. For face detection, AI accelerator achieves 15% speed-up. For image classification, 38% speed-up and 13 power-saving are achieved on the MediaTek AI accelerator.

Voice and Language Model Optimization and Development 2017.12 – 2022.09

- Benchmark Optimization (ETHZ-V5, MLperf): for ETHZv5 LSTM model, 3 times speed-up is achieved by analyzing DRAM bandwidth (operation sequence). For MLperf, Mobile BERT achieves nearly 3 times speed-up from 8.1ms to 2.6ms.
- Customer NPU Deployment and Optimization
  1) Automatic speech recognition landing: co-develop with Vivo on project Vivo X90 and Dimensity 9200 achieve the fist landing of speech recognition on NPU, with 36% power saving and 76% speed-up.
  2) Text to speech landing: by replacing un-supported OP and fine-tuning of text-to-speech models (Hifi-gan), 31% power saving and 48% speed-up is achieved.

Speech Denoise and speech recognition

- Speech denoise project: by analyzing the echo, noise generation process, design a data flow to collect data and synthesize data; design a LSTM model to achieve SOTA denoise algorithm POLQA 3.3.
- Speech recognition project: utilize the opensource code Facebook wave2letter C++ framework and design a voice based remote control for TV (14MB ASR model) and 6.5% word error rate.

Research Projects

Nanyang Technological University, Ph.D. program

Tensor-train based matrix decomposition and accelerator design 2016 – 2018

- Proposed tensor-train based decomposition algorithm and achieve the state-of-the-art compression performance. Based on this algorithm, an RRAM in memory computing algorithm is proposed.
- Algorithm paper published in IEEE TNNLS and AI hardware accelerator published in IEEE TNANO。

Patent

1. Hardware-Aware Mixed-Precision Quantization, US Patent App. 17/852,484, 2024
2. AI-Assisted Power Amplifier Optimization, US Patent US-20230006611-A1, 2023
3. Heterogeneous Computing for Hybrid Acoustic Echo Cancelation, US Patent App. 17/688,600, 2023
4. Calibration of Analog Circuits For Neural Network Computing, US Patent US-20220230064-A1, 2022

Selected Publications

Books:

1. **Hantao Huang**, Hao Yu," Compact and Fast Machine Learning Accelerator for IoT devices", Springer 2019 (Order link: https://www.springer.com/us/book/9789811333224)

Algorithms and Applications:

1. Ziyi Guan*, **Hantao Huang***, Yupeng Su, Hong Huang, Ngai Wong and Hao Yu. "APTQ: Attention-aware Post-Training Mixed-Precision Quantization for Large Language Models", IEEE/ACM Design and Automation Conference (**DAC**), 2024 (submitted, * equal contribution)

2. Ziyi Guan, Boyu Li, Yuan Ren, Muqun Niu, **Hantao Huang**, Graziano Chesi, Hao Yu, and Ngai Wong, “An Isotropic Shift-Pointwise Network for Crossbar-Efficient Neural Network Design”, ACM/IEEE Design Automation and Test Conference in Europe (**DATE**), March 2024.
3. Zikun Wei, Tingting Wang, Chenchen Ding, Bohan Wang, Ziyi Guan, **Hantao Huang***, and Hao Yu , “FMTT : Fused Multi-head Transformer with Tensor-compression for 3D Point Clouds Detection on Edge Devices”, ACM/IEEE Design Automation and Test Conference in Europe (**DATE**), March 2024. (Corresponding authors)
4. Han, Wei**, Hantao Huang** and Xiaoxi Yu. “TAPL: Dynamic Part-based Visual Tracking via Attention-guided Part Localization.” British Machine Vision Conference (**BMVC**), 2021.
5. Wei Han*, **Hantao Huang*** and Tao Han, “Finding the Evidence: Localization-aware Answer Prediction for Text Visual Question Answering”, The 28th International Conference on Computation Linguistics (**COLING**), 2020 (Oral presentation accepted, * equal contribution)
6. **Hantao Huang**, Tao Han, Wei Han, Deep Yap and Chiangmeng Chiang, “Answer-checking in Context: A Multi-modal Fully Attention Network for Visual Question Answering”, The 25th International Conference on Pattern Recognition (**ICPR**), 2020
7. **Hantao Huang** and Hao Yu, “Layer-Wise Training of Tensorized Neural Network”, IEEE Transactions on Neural Networks and Learning Systems (**TNNLS**), 2018.
8. **Hantao Huang**, Hang Xu and Hao Yu, “Hantao Huang*, Yuehua Cai, Hang Xu and Hao Yu, “Distributed Machine Learning on Smart-gateway Network towards Real-time Smart-grid Energy Management with Behavior Cognition”, Design Automation of Electronic Systems (**TODAES**), 2018.
9. **Hantao Huang**, Yuehua Cai, Hang Xu and Hao Yu, “A Multi-agent Minority-game based Demand-response Management of Smart Buildings towards Peak Load Reduction”, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (**TCAD**), 2016. (doi: 10.1109/TCAD.2016.2571847)
10. **Hantao Huang**, Yuehua Cai, and Hao Yu, “Distributed-neuron-network based Machine Learning on Smart-gateway Network towards Real-time Indoor Data Analytics”, ACM/IEEE Design Automation and Test Conference in Europe (**DATE**), March 2016.

AI Accelerator and Architecture:

1. Kai Li, **Hantao Huang**, Mingqiang Huang, Chenchen Ding, Longyang Lin, Liebing Ni, Hao Yu, “A 29.12 TOPS/W and 1.13 TOPS/mm2 NAS-Optimized Mixed-Precision DNN Accelerator with Vector Split-and-Combination Systolic in 28nm CMOS”, IEEE Custom Integrated Circuits Conference (**CICC**), 2024
2. Sai Manoj Dinakarrao, **Hantao Huang**. and Hao Yu, "Energy-Efficient and Error-Resilient Cognitive I/O for 3-D-Integrated Manycore Microprocessors," in IEEE Design & Test (**D&T**), vol. 38, no. 6, pp. 88-95, Dec. 2021
3. **Hantao Huang**, Leibin Ni and Hao Yu, “A Highly-parallel and Energy-efficient 3D Multi-layer CMOS-RRAM Accelerator for Tensorized Neural Network”, IEEE Transactions on Nanotechnology (**TNANO**), 2017.
4. **Hantao Huang**, Rai Sulman Khalid and Hao Yu, “IoT Network Intrusion Detection by Online Sequential Machine Learning Accelerator”, International Conference on Hardware/Software Codesign and System Synthesis (**CODES+ISSS**), 2017.
5. **Hantao Huang**, Leibin Ni, and Hao Yu, “TNN: An Energy-efficient Machine Learning Accelerator on 3D CMOS-RRAM for Tensorized Neural Network”, IEEE International System-On-Chip Conference (**ISOCC**), 2017
6. **Hantao Huang**, Leibin Ni and Hao Yu, “A 3D Multi-layer CMOS-RRAM Accelerator for Multi-layer Machine Learning”, IEEE International Conference on Solid-State and Integrated Circuit Technology (**ICSICT**) (Invited), October 2016.
7. Leibin Ni, **Hantao Huang**, Zichuan Liu, Rajiv V. Joshi and Hao Yu, “Distributed In-Memory Computing on Binary RRAM Crossbar”, ACM Journal on Emerging Technologies in Computing System (**JETC**), 2016. (doi: 10.1145/2996192)
8. Wei Pang, **Hantao Huang**, Fengwei An, and Hao Yu, "Low-power and Real-time Computer Vision On-chip", IEEE System-on-Chip Conference (**ISOCC**) (Invited), Korea, October 2016
9. Yuhao Wang, **Hantao Huang**, Leibin Ni, Hao Yu, Mei Yan, Chuliang Weng, Wei Yang and Junfeng Zhao, “An Energy-efficient Non-volatile In-Memory Accelerator for Sparse-represented Face Recognition”, ACM/IEEE Design Automation and Test Conference in Europe (**DATE**), March 2015.
10. **Hantao Huang**, Sai Manoj P.D., Dongjun Xu, Hao Yu, and Zhigang Hao, "Reinforcement Learning based Self-adaptive Voltage-swing Adjustment of Through-silicon Interposer I/Os for Many-core Microprocessor and Memory Communication," IEEE/ACM International Conference of Computer-Aided-Design (**ICCAD**), November 2014.

Invited Talks：

1. Waimon Wong and **Hantao Huang**, “Enabling Edge AI”, NTU-MediaTek IC Design Workshop, Singapore 2019
2. **Hantao Huang**, “Make a Power-efficient Voice UI on Edge Devices”, Speech Signal Processing Workshop, Taiwan, China 2020
3. **Hantao Huang,** “面向大模型的高能效并行存算大算力芯片”, 全球 AI 芯片峰会 (GACS), Shenzhen China, 2023.