Hanxian Huang

hanxianhuang@meta.com | (+1) 858 214 0677 | https://hanxian97.github.io

EDUCATION

University of California San Diego, Department of Computer Science and Engineering — San Diego, CA, USA
Ph.D. student in Computer Science, *GPA: 4.00/4.00* — Dec. 2022 - Dec. 2024

University of California San Diego, Department of Computer Science and Engineering — San Diego, CA, USA
M.S. in Computer Science, *GPA: 3.96/4.00* — Sep. 2019 - Dec. 2022

Peking University, School of Electronics Engineering and Computer Science — Beijing, China
B.S in Computer Science and Technology, *Major GPA: 3.75/4.00* — Sep. 2015 - Jul. 2019

RESEARCH INTERESTS

Efficient ML; Model sparsity and quantization; ML system and compiler; ML-system co-design and co-optimization

AWARDS

ACM SIGSOFT Distinguished Paper Award	2024
Machine Learning and Systems Rising Star	2024
Powell Fellowship	2019
Schlumberger Scholarship (top 3%)	2018
"Merit student" at Peking University (top 3%)	2018
Second Prize in the 6th PKU Young Scientists Symposium on Informatics (top 5%)	2018

WORK EXPERIENCE

CoreAI team, Reality Lab, Meta — Burlingame, CA,USA
AI Research Scientist, report to Raghuraman Krishnamoorthi & Igor Fedorov — Jan. 2025 – current

- Leading LLM model optimization efforts, delivering SOTA on-device models for both open-source releases (MobileLLM Pro, MobileMoE, ParetoQ) and production.
- Leading LLM architecture search tailored for mobile and edge devices, achieving optimal trade-offs between model quality and efficiency (MobileLLM-Flash).

PyTorch Core Performance team, Meta — Menlo Park, CA,USA
Research Intern, Supervised by Dr. Andrew Or and Supriya Rao — Jun. 2024 – Sep. 2024

- Bayesian Optimization for Mixed-Precision Quantization
  - Developed a Bayesian Optimization tool to automatically search mixed-precision weight-only quantization configuration, by leveraging the layer-wise sensitivity prior knowledge. Delivered the tool under TorchAO.
  - Reduced 20.1% model size with 2.8% perplexity reduction, and improves 15.1% inference throughput with 3.2% perplexity reduction on the Llama3-8B model compared to int8 uniform quantization.

Gray System Lab, Microsoft Research — (remote) Redmond, WA,USA
Research Intern, Supervised by Dr. Yuanyuan Tian Jun. 2022 – Sep 2022 (full-time), Oct. 2022 - May 2023 (part-time)

- SIBYL: Forecasting Time-Evolving Query Workloads (SIGMOD'24)
  - Designed SIBYL, an end-to-end learning-based framework accurately forecasts sequences of future queries, with the entire query statements, in various prediction windows.
  - Addressed the challenge of large prediction windows (up to 10k), demonstrating high scalability over large workloads with highly varying query arrival rates.
  - SIBYL achieves high forecasting accuracy with an 87.3% median F1 score and results in 1.7× and 1.3× performance improvement when applied to materialized view selection and index selection applications.

Y-tech Lab, Kwai Inc. — (remote) Palo Alto, CA, USA
Machine Learning Intern, Supervised by Dr. Xin Chen — Jun. 2021 – Sep. 2021

- Fasor: A Fast Tensor Program Optimization Framework (ICS '24)
  - Identified the bottleneck in efficient DNN compilation is the on-device measurement for cost model training, which can take $\sim$ 80% of the optimization time.
  - Developed Fasor, a fast tensor program optimization framework, reducing optimization time with a transferable cost model, a search space shrinking module, and a deep reinforcement learning-based schedule search engine.
  - Improved compilation efficiency of DNNs on the Intel CPU and NVIDIA GPU by up to 10.24× and 8.17×, respectively, compared to TVM, with better or equal output code latency performance.

Research Intern, Supervised by Dr. Xin Chen — Jun. 2020 – Sep. 2020

- ADMM-based Model Pruning

– Proposed a new 4-step model pruning pipeline consisting of pre-train, ADMM-softcut, prune, and fine-tune, outperforming traditional 3-step pruning in terms of both accuracy and efficiency (model compression ratio).
– Designed a novel softcut method based on the alternating direction method of multipliers (ADMM) to redistribute knowledge from the pruned subnet to the target subnet while preserving target subnet accuracy.
– Achieved 70% pruning ratio on several model architectures with no accuracy loss on the ImageNet dataset.

Heterogeneous and Extreme Computing (HEX) group, Microsoft Research Asia Beijing, China
Research Intern, Supervised by Dr. Chen Zhang Oct. 2018 – May 2019

- Sparsity/Quantization for Larger-scale Neural Network
  – Studied and evaluated the model and activation sensitivity to quantization and sparsity and explored the relationship among quantization, sparsity, and model architecture.
  – Designed an end-to-end algorithm to automatically search the compact objective model architecture according to the layer's sensitivity during model training while maintaining high accuracy comparable to full-size models.

RESEARCH EXPERIENCE

STABLE Lab, University of California San Diego San Diego, CA, USA
Graduate Student Researcher, Supervised by Prof. Jishen Zhao Sep. 2019 - Present

- Multi-Agent LLMs for Verilog RTL Code Generation (Hot Chips 2024 Tutorial, DAC 2025) May 2024
  – Designed a multi-agent framework with a specialized RTL code-generation agent, a test/verification agent, and a debugging/code refining agent. Enabling a feedback loop to achieve self-verification and self-correction, to improve the pass rate of the RTL generated code, by up to 10.4% on the pass@5 score.
  – By mimicking human designers' behavior, LLMs generate test benches with test cases, walk through the generated code with test cases to reason the code behavior considering timing, and refine the generated code.
- WASMREV: Multi-modal Learning for WebAssembly Reverse Engineering (ISSTA '24) Dec. 2023
  – Developed WASMREV, a multi-modal learning model capturing the relationship among source code, code documentation, and WebAssembly binaries to enhance WebAssembly reverse engineering.
  – Deployed WASMREV for various reverse engineering tasks, e.g., type recovery (74.9% accuracy), function identification (93.1% F1 score), and binary code documentation (87.3% BERTScore-F1).
- Triple: Efficient Vision Transformer (ViT) Training and Scaling (ICCV'23) Dec. 2022
  – Explored scalable ViT training and identified one key to reuse pre-trained models is preserving optimizer states.
  – Collaborated on building Triple, an efficient ViT training and scaling framework through pre-trained model reuse, progressive learning, and knowledge distillation.
  – Saved up to 80% of training time compared to from-scratch training when scaling ViTs by 8×.
- Q-gym: A Equality Saturation-based DNN Inference Framework (PACT'22) May 2022
  – Collaborated on building Q-gym, a DNN inference framework leveraging equality saturation and exploiting weight repetition to generate efficient expressions for convolutional layers.
  – Achieved a significant reduction (70%) in the number of operations and achieved 2.56×/1.78× inference speedup on CPU / GPU compared to OneDNN and PyTorch.
- GateNet (ICLR Workshop '21) and GateFlow: Accelerating BNN evaluation under FHE Dec. 2020
  – Developed GateNet, a new Fully Homomorphic Encryption (FHE)-friendly BNN model with group convolution to reduce the cost of 'popcount' and advanced non-linear functions to preserve model accuracy.
  – Developed GateComp, a new FHE compiler that leverages weight repetition feature of BNNs to reuse computations during FHE-based BNN evaluation, bridging the gap between BNN evaluation and FHE.
  – Achieved an average speedup of 13.41× over the SOTA BNNs under FHE evaluation.
- Learn-to-Share: Efficient Multi-NLP Tasks Training (ICML'21) Jan. 2021
  – Collaborated on designing Learn-to-Share, a framework that leverages both parameter and computation sharing across multiple tasks for NLP by a novel neural architecture search.
  – Designed a novel delta-pruning in the early stage of model fine-tuning based on a salient criterion based on connection sensitivity, allowing highly parameter sharing and adding only 1.4% of extra parameters per task.
  – Reduced the computation by 49.5% on GLUE benchmarks compared to full fine-tuning on each task.
- Ayudante: Assisting Persistent Memory Programming (USENIX ATC'21) Dec. 2020
  – Developed Ayudante, a Reinforcement Learning-based assistant to select APIs based on volatile C/C++/Java code, generating persistent memory-aware code.
  – Developed a code refining pipeline consisting of advanced persistency checkers to parse the generated code and provide users with a report for further program testing and performance optimization.
  – Achieved a high persistency checker pass rate (78.7% $\sim$ 100%) and comparable performance to expert code.

Microprocessor Architecture Researchers LAB, University of California Los Angeles Los Angeles, CA, USA
Research Intern, Supervised by Prof. Glenn Reinman Jul. 2018 – Sep. 2018

- Efficient Face Recognition Application on Computing Hierarchy

– Developed a video analysis pipeline by leveraging the computation and memory characteristics of applications to better fit into the Computing Hierarchy, achieving high throughput on Alpha Data FPGA Board.
– Mapped video decoder and face detection stage to Near-Storage accelerator, face recognition stage to the Near-Memory accelerator, and face verification stage to the Near-Cache accelerator.

Center for Energy-efficient Computing and Applications, Peking University Beijing, China
Research Assistant, Supervised by Prof. Guojie Luo Sep. 2017 – Mar. 2018

- Adaptive-Precision Framework for SGD Using Deep Q-Learning (ICCAD'18) Mar. 2018
  – Proposed a framework for adaptive-precision adaptation using Q-learning, automatically trading off precision for throughput and accelerating SGD.
  – Employed re-configurable devices (FPGAs) to support adaptive precision representations generated by Q-learning, increasing throughput by up to 4.3× compared to 32-bit floating point setting.
- FPGA-based Real-Time Super-Resolution System (FCCM'18) Jan. 2018
  – Designed an algorithm to automatically decide the usage of accurate but complex CNNs or fast but naive interpolation for real-time super-resolution of ultra-high-definition videos.
  – Implemented an FPGA-based accelerator, balancing the resource utilization, the attainable frame rate, and the resolution quality to achieve efficient real-time (30 fps) super-resolution for Ultra High-Definition (4k) videos.

PUBLICATIONS

22. Y. Chen, **H. Huang**, E. Chang, J. Szwejbka, D. Desai, Z. Liu, V. Chandra, R. Krishnamoorthi "MobileMoE: Scaling On-Device Mixture of Experts". Arxiv.

21. **H. Huang**, I. Fedorov, A. Gromov, B. Beckerman, N. Suda, D. Eriksson, M. Balandat, R. Conway, P. Huber, C. Sankar, A. Dalmia, Z. Liu, L. Wu, T. Elgamal, A. Sagar, V. Chandra, R. Krishnamoorthi "MobileLLM-Flash: Latency-Guided On-Device LLM Design for Industry Scale Deployment". ACL 2026 Industry Track.

20. P. Huber, E. Chang, W. Wen, I. Fedorov, T. Elgamal, **H. Huang**, N. Suda, C. Sankar, V. Vogeti, Y. Wang, A. Gladkov, K. S. Tai, A. Elogeel, T. Hefny, V. Chandra, A. Aly, A. Kumar, R. Krishnamoorthi, A. Sagar. "MobileLLM-Pro Technical Report". Arxiv.

19. Z. Liu, C. Zhao, **H. Huang**, S. Chen, J. Zhang, J. Zhao, S. Roy, L. Jin, Y. Xiong, Y. Shi, L. Xiao, Y. Tian, B. Soran, R. Krishnamoorthi, T. Blankevoort, V. Chandra "ParetoQ: Scaling Laws in Extremely Low-Bit LLM Quantization". *The Thirty-Ninth Annual Conference on Neural Information Processing Systems (NeurIPS 2025).*

18. Z. Yu, H. Zhang, Y. Zhao, **H. Huang**, M. Yao, K. Ding, J. Zhao "OrcaLoca: An LLM Agent Framework for Software Issue Localization". *International Conference on Machine Learning (ICML), 2025.*

17. Y. Zhao, H. Zhang, **H. Huang**, Z. Yu, J. Zhao "MAGE: A Multi-Agent Engine for Automated RTL Code Generation". *ACM/IEEE Design Automation Conference (DAC), 2025.*

16. Y. J. Soh, **H. Huang**, Y. Tian, J. Zhao "You Only Use Reactive Attention Slice for Long Context Retrieval". *EMNLP Findings, 2025.*

15. X. Chen, **H. Huang**, Y. Gao, Y. Wang, J. Zhao, K. Ding "Learning to Maximize Mutual Information for Chain-of-Thought Distillation". *ACL Findings, 2024.*

14. **H. Huang**, X. Chen, J. Zhao "Fasor: A Fast Tensor Program Optimization Framework for Efficient DNN Deployment". *International Conference on Supercomputing (ICS), 2024.*

13. **H. Huang**, J. Zhao "Multi-Representation Learning for WebAssembly Reverse Engineering". *ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA), 2024.* **ACM SIGSOFT Distinguished Paper Award**.

12. **H. Huang**, T. Siddiqui, R. Alotaibi, C. Curino, J. Leeka, A. Jindal, J. Zhao, J. Camacho-Rodríguez, Y. Tian "SIBYL: Forecasting Time-Evolving Query Workloads". *ACM SIGMOD International Conference on Management of Data, 2024.*

11. C. Fu, **H. Huang**, Z. Jiang, Y. Ni, L. Nai, G. Wu, L. Cheng, Y. Zhou, S. Li, A. Li, J. Zhao "TripLe: Revisiting Pretrained Model Reuse and Progressive Learning for Efficient Vision Transformer Scaling and Searching". *International Conference on Computer Vision (ICCV), 2023.*

10. C. Fu, **H. Huang**, B. Wasti, C. Cummins, R. Baghdadi, K. Hazelwood, Y. Tian, J. Zhao, H. Leather "Q-Gym: An Equality Saturation Framework for DNN Inference Exploiting Weight Repetition". *International Conference on Parallel Architectures and Compilation Techniques (PACT), 2022.*

9. **H. Huang**, Z. Wang, J. Kim, S. Swanson, J. Zhao "Ayudante: A Deep Reinforcement Learning Approach to Assist Persistent Memory Programming". *USENIX Annual Technical Conference (ATC), 2021.*

8. C. Fu, **H. Huang**, X. Chen, Y. Tian, J. Zhao "Learn-to-Share: A Hardware-Friendly Transfer Learning Framework Exploiting Computation and Parameter Sharing". *International Conference on Machine Learning (ICML, Long Presentation), 2021.*

7. C. Fu, **H. Huang**, X. Chen, J. Zhao "GateNet: Bridging the Gap Between Binarized Neural Network and FHE Evaluation". *ICLR Workshop on Security and Safety in Machine Learning Systems, 2021.*

6. W. Zhang*, **H. Huang***, J. Zhang, M. Jiang, G. Luo "Adaptive-Precision Framework for SGD Using Deep Q-Learning". *International Conference on Computer-Aided Design (ICCAD), 2018.*

5. Z. He*, **H. Huang***, M. Jiang, Y. Bai, G. Luo "FPGA-Based Real-Time Super-Resolution System for Ultra High-Definition Videos". *IEEE International Symposium on Field-Programmable Custom Computing Machines (FCCM), 2018.*

4. G. Luo, Z. He, **H. Huang**, Y. Bai, H. Jia, M. Jiang. "FPGA-Based Real-Time Super-Resolution Method and System". *Patent CN108765282B, 2020.10.09.*

3. Z. Peng, Y. Liu, **H. Huang**, Y. Ren, J. Yang, L. Liu, X. Chen "Multi-Level Intermediate Representation Decoder for Heterogeneous Platforms". *Patent US11928446B2, 2024.03.12.*

2. Y. Tian, **H. Huang**, R. Alotaibi, T. Siddiqui, J. Leeka, J. Camacho-Rodríguez, C. Curino "Characterizing and Forecasting Evolving Query Workloads". *Patent US12197405, 2024.11.07.*

1. **H. Huang**, Z. Lin, Z. Wang, X. Chen, K. Ding, J. Zhao "Towards LLM-Powered Verilog RTL Assistant: Self-Correction and Self-Verification". *Hot Chips 2024 Tutorial on AI for Chip Design, 2024.*

*Contributed equally