Hanxian Huang

hah008@ucsd.edu | (+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 (expected)

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

Intersection of machine learning with computer systems, compilers, and programming languages; ML4Sys and Sys4ML

AWARDS

Machine Learning and Systems Rising Star 2024 — 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
Student travel grant: HPCA 2020, PACT 2022, ICCV 2023, SIGMOD 2024, ASPLOS 2024

WORK EXPERIENCE

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

- GeoT: Tensor Centric Library for Graph Neural Networks In Progress
  - Collaborate on designing a customized tiling algorithm for segment reduction in GNNs and developing advanced thread workload mapping and GPU-specific optimizations.
  - Deploy a data-driven, ultra-low overhead decision tree for configuration rule selection, which offers heuristic adaptability to dynamic input scenarios.
  - Achieving an average operator speedup of 1.80× and an end-to-end speedup of 1.68× compared to PyTorch.
  - GeoT is being integrated into PyTorch.
- Exploiting CS Education Theory for LLMs Prompting in Programming Tasks In Progress
  - Investigate and categorize programming failures, quantify programming struggles, evaluate multi-turn programming, and measure repeated error density of existing Large Language Model (LLM)-based code generation tools.
  - Exploit CS education methodologies, e.g., collaborative programming, incremental learning, analogical learning, and reverse learning to enhance LLM prompting for programming tasks.
- LLMs for Verilog RTL Code Generation In Progress
  - Design an auto-prompt system to improve the accuracy of the RTL generated code by LLM through self-correction and self-verification.
  - Explicitly prompt LLM to generate test benches with test cases, walk through the generated code with test cases to symbolically 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 exploring the relationship among source code, comments, 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).
- WASMBert: First Transferable WebAssembly Language Model Aug. 2023
  - Developed WASMBert, a novel WebAssembly language model for generalized WebAssembly analysis tasks.
  - Extracted comprehensive syntactic and semantic features from binary code sequences and code property graphs and tailored pre-training tasks, enhancing WASMBert's ability to comprehend intricate WebAssembly structures.
  - Achieved up to 55% and 16% recall improvement for WebAssembly similarity and vulnerability detection tasks compared to conventional analysis tools, delivering user-friendly tools with superior code inspection suggestions.
- 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 GPU.
- 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.
  - Achieved an average speedup of 13.41× over the SOTA BNNs under FHE evaluation, bridging the gap between BNN evaluation and FHE.
- 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

1. X. Chen, **H. Huang**, Y. Gao, Y. Wang, J. Zhao, K. Ding "Learning to Maximize Mutual Information for Chain-of-Thought Distillation", *Annual Meeting of the Association for Computational Linguistics (ACL) Findings, 2024*
2. **H. Huang**, X. Chen, J. Zhao "Fasor: A Fast Tensor Program Optimization Framework for Efficient DNN Deployment", *International Conference on Supercomputing (ICS), 2024*
3. **H. Huang**, J. Zhao "Muti-representation Learning for WebAssembly Reverse Engineering", *ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA), 2024*
4. **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*
5. 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*
6. C. Fu, **H. Huang**, B. Wasti, C. Cummins, R. Baghdadi, K. Hazelwood, Y. Tian, J. Zhao, and H. Leather "Q-gym: An Equality Saturation Framework for DNN Inference Exploiting Weight Repetition", *In the Proceedings of the International Conference on Parallel Architectures and Compilation Techniques (PACT), 2022*
7. **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", *In the Proceedings of International Conference on Machine Learning (ICML, Long Presentation), 2021*
9. 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*
10. W. Zhang*, **H. Huang***, J. Zhang, M. Jiang, G. Luo "Adaptive-Precision Framework for SGD using Deep Q-Learning", *In the Proceedings of International Conference on Computer Aided Design (ICCAD), 2018*
11. 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*
12. 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*
13. 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*
14. **H. Huang**, J. Zhao "Neural WebAssembly Comprehension: A Transferable WebAssembly Learning for Generalized Analysis Tasks", *Under Review*
15. C. Fu*, **H. Huang***, H. Chen, Y. Li, J. Zhao "GateFlow: Bridging The Gap Between Binarized Neural Network And Fully Homomorphic Encryption", *Under Review*
16. Z. Yu, G. Zhang, **H. Huang**, K. Ding, J. Zhao, "GeoS: Building Efficient Tensor Centric Library for Graph Neural Network via Segment Reduction", *Under Review*
17. **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", *US Patent Application Filed*

*Contribute equally