GUYUE HUANG

Email: hguyue1@gmail.com • Phone: (+1) 571-523-6855 • Linkedin • Google Scholar

EDUCATION

University of California, Santa Barbara, Santa Barbara, CA 09/2020 - 04/2024
Ph.D. Degree, Department of Electrical and Computer Engineering.

Tsinghua University, Beijing, China 09/2016 - 07/2020
Bachelor Degree, Department of Electronic Engineering.
Graduate with honor (Top 10%).

EXPERIENCES

Senior Deep Learning Architect, NVIDIA, Santa Clara, CA. 05/2024 - Present

- Develop software systems to improve Large Language Model (LLM) training performance on GPU clusters.
- Perform benchmarking of Deep Learning training workloads to help future GPU architecture design.

Machine Learning Engineer Intern, Meshy LLC, Santa Clara. 01/2024 - 03/2024

- Improve the performance of text-to-3D and image-to-3D diffusion model inference and training.

Deep Learning Architect Intern, NVIDIA, Santa Clara, CA. 06/2023 - 09/2023

- Work on a software feature that uses PyTorch 2.0 compiler to accelerate huggingface Parameter-Efficient Fine-Tuning (PEFT) models. The key method is to debug the torch dynamo tracing process to reduce graph breaks.
- Work on a software feature that accelerates the LLM long context fine-tuning workloads.
- Participate in the development and maintenance for the PEFT feature in the NVIDIA NeMo software package. Work on bug fixing related to the distributed training package (Megatron-LM).

Deep Learning Architect Intern, NVIDIA, Santa Clara, CA. 06/2022 - 09/2022

- Work on a software feature that reduces the memory usage of the Deep Learning Recommendation Model in Tensorflow.

Research Intern, Alibaba DAMO Academy, Shanghai China 08/2020 - 09/2021

- Work on a research project that develops an algorithm-software co-optimization framework for accelerating sparse DNN inference on GPUs. Publish a first-author paper in DAC'22.
- Participate in the designing and modeling of the sparse acceleration hardware feature on internal DNN accelerator.

PUBLICATIONS

- Xu J., Huang S., Li J., **Huang G.**, Xie Y., Wang Y., Dai G. Enabling Efficient Sparse Multiplications on GPUs with Heuristic Adaptability. In IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD), 2024.
- Zhao T., Ning X., Fang T., Liu E., **Huang G.**, Lin Z., Yan S., Dai G., Wang Y. MixDQ: Memory-Efficient Few-Step Text-to-Image Diffusion Models with Metric-Decoupled Mixed Precision Quantization. In Proceedings of European Conference on Computer Vision (ECCV), 2024.
- Wang Z., Wang Y., Feng B., **Huang G.**, Mudigere D., Muthiah B., Li A., Ding Y. OPER: Optimality-Guided Embedding Table Parallelization for Large-scale Recommendation Model. In Proceedings of USENIX Annual Technical Conference (ATC), 2024.
- **Huang G.**, Wang Z., Tsai P., Zhang C., Ding Y., Xie Y. RM-STC: Row-Merge Dataflow Inspired GPU Sparse Tensor Core for Energy-Efficient Sparse Acceleration. In Proceedings of 56th IEEE/ACM International Symposium on Microarchitecture (MICRO), 2023.
- Wang Y., Feng B., Wang Z., **Huang G.**, Ding Y. TC-GNN: Bridging Sparse GNN Computation and Dense Tensor Cores on GPUs. *To appear* in the USENIX Annual Technical Conference (ATC), 2023.

- **Huang G.**, Bai Y., Liu L., Wang Y., Yu B., Ding Y., Xie Y. ALCOP: Automatic Load-Compute Pipelining in Deep Learning Compiler for AI-GPUs. In Proceedings of the Sixth Conference on Machine Learning and Systems (MLSys), 2023.
- Wang X., Wei Y., Xiong Y., **Huang G.**, Qian X., Ding Y., Wang M., Li L. LightSeq2: Accelerated Training for Transformer-based Models on GPUs. In Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC), 2022.
- **Huang G.**, Li H., Sun F., Qin M., Ding Y. and Xie Y. Shfl-BW: Accelerating Deep Neural Network Inference with Tensor-Core Aware Weight Pruning. In Proceedings of the 59th ACM/IEEE Design Automation Conference (DAC), 2022.
- Dai G., **Huang G.**, Yang S., Yu Z., Zhang H., Ding Y., Xie Y., Yang H. and Wang Y. Heuristic Adaptability to Input Dynamics for SpMM on GPUs. Design Automation Conference (DAC), 2022. (**Best Paper Nominee**).
- Zhang H., Yu Z., Dai G., **Huang G.**, Ding Y., Xie Y. and Wang Y. Understanding GNN Computational Graph: A Coordinated Computation, IO, and Memory Perspective. In Proceedings of Fifth Conference on Machine Learning and Systems (MLSys), 2022.
- **Huang G.**, Dai G., Wang Y., Ding Y. and Xie Y. Efficient Sparse Matrix Kernels based on Adaptive Workload-Balancing and Parallel-Reduction. Poster, ACM Student Research Competition (SRC), 2021.
- Yu Z., Dai G., **Huang G.**, Wang Y. and Yang H. Exploiting Online Locality and Reduction Parallelism for Sampled Dense Matrix Multiplication on GPUs. In Proceedings of the IEEE International Conference on Computer Design (ICCD), 2021.
- **Huang G.**, Hu, J., He, Y., Liu, J., Ma, M., Shen, Z., Wu, J., Xu, Y., Zhang, H., Zhong, K., & others. Machine learning for electronic design automation: A survey. ACM Transactions on Design Automation of Electronic Systems (TODAES), 26(5), 1-46. 2021.
- **Huang, G.**, Dai, G., Yu, W., Huazhong, Y. GE-SpMM: General-purpose Sparse Matrix-Matrix Multiplication on GPUs for Graph Neural Networks. In Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC), 2020.
- Li Q., Zhu H., **Huang G.**, Yu Z., Qiao F., Wei Q., Liu X., Yang H., Low-power in-pixel buffer circuit for smart image sensor. Sensor Review, 2020.

PATENTS AND PATENT APPLICATIONS

- Sun F., Qin M., Li H., Zhu G., Gao Y., **Huang G.**, Zhang Y. Systems and methods for neural network training with weight sparsity. US Patent Application US20230306257A1.
- Gao Y., Sun F., Li H., **Huang G.**, Zhang C., Zhong R. Warp execution method and associated gpu. US Patent Application US20230394617A1.

AWARDS

- Best paper candidate, DAC 2022.
- ACM Student Research Competition (SRC), Graduate Student Third Place, 2021.
- MICRO'20 Student Research Competition (SRC), Undergraduate Student First Place.
- Travel grant of MLSys'23, MLSys'22.