GUYUE HUANG

Email: hguyue1@outlook.com

EDUCATION

University of California, Santa Barbara, Santa Barbara, CA *09/2020 - Present*
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

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

1. Software optimizations for parameter-efficient fine-tuning (PEFT) models like LoRA and p-tuning.
2. Software optimizations for long sequence-length LLM fine-tuning.
3. Contribute to the open-source LLM framework NeMo.

NVIDIA, Santa Clara, CA. *06/2022 - 09/2022*
Deep Learning Architect Intern
Optimizing Deep Learning Recommendation Model (DLRM) training in Tensorflow.
Focus on memory footprint reduction for DLRM with SGD and Adam optimizer in XLA.

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

1. Do research about TensorCore-friendly sparse neural networks. Publish a first-author paper in DAC'22.
2. Develop and profile sparse kernel performance on GPU using NVIDIA Ampere Sparse TensorCore.

RESEARCH INTERESTS

My PhD research focuses on supporting sparsity in AI/DL on GPU. I investigate software and architecture techniques to efficiently support sparsity in AI including weight/activation sparsity, embedding operators, graphs, and Mixture-of-Experts. I also work on developing fast GPU kernels and improving DL compilers.

PUBLICATIONS

- **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.

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.

RESEARCH EXPERIENCES SELECTED

RM-STC: Row-Merge GPU Sparse Tensor Core for Energy-Efficient Sparse Acceleration

We design a novel Sparse Tensor Core architecture with two highlights: firstly, we support sparse training without the overhead of sparse transposition, and secondly, we achieve energy-efficiency and structure scalability via applying a row-wise dataflow. RM-STC achieves in average 1.93× lower energy-delay product than state-of-the-art STCs. (**Keywords: GPU Architecture, Tensor Core)**

ALCOP: Automatic Load-Compute Pipelining in Deep Learning Compiler for AI-GPUs

We develop IR transformation for asynchronous copy and shared memory swizzling on NVIDIA Ampere GPU. We improve TVM auto-tuning efficiency by leveraging analytical hardware performance model. The method achieves an average of 1.23× speedup for dense kernels against out-of-the-box TVM. (**Keywords: DL Compiler, Software Pipelining)**

Shfl-BW: Accelerating DNN Inference with Tensor-Core Aware Weight Pruning

We propose a novel parameter sparsity pattern, *Shuffled Block-wise sparsity*, to efficiently utilize TensorCores while minimizing the constraints on the weight structure. We optimize GPU kernels and achieve 4.18× speedup on NVIDIA T4 GPU for Transformer inference at 75% sparsity. (**Keywords: GPU Kernel, SpMM)**