Yibai Meng

Mountain View, CA ✉ yibai@meng.engineer in linkedin.com/in/yibai-meng ☎ +1 6692009099

Skills

Programming Languages: Python, C/C++, Go

Machine Learning: PyTorch, Tensorflow, Jax, XLA, MLIR, Transformer/LLM, quantization

GPU: CUDA, OpenAI Triton, CUTLASS, Nsight Compute

Industry Experiences

Waymo Jan 2023 - Present
Software Engineer *Mountain View, California*

- Working on ML model optimization and inference infrastructure.
- **GPU kernel development**: Implemented high-performance transformer kernels, including flash attention and quantized feed-forward operations, in OpenAI Triton. Profiled and fine-tuned these kernels using Nsight Compute, identifying subtle code generation issues within our underlying stack. Conducted an in-depth analysis of the generated PTX code and achieved bit-for-bit accuracy with the JAX XLA reference implementation. This optimization led to a 3x end-to-end speedup compared to XLA, enabling the deployment of a critical foundation model for on-car usage. Additionally, designed and developed continuous integration testing and benchmarking infrastructure.
- **LLM quantization**: Implemented quantization for LLM transformer kernels, including 8-bit weight-and-activation, 4-bit weight-only, and 4-bit weight-and-activation quantization. For 4-bit weight-only quantization, utilized bitwise operations and inline assembly to address the lack of native Triton support. For 4-bit weight-and-activation quantization, employed CUTLASS to develop a high-performance fused quantized projection kernel.
- **Model optimization**: Improved the latency and stability of on-vehicle machine learning models using techniques such as quantization and operation fusion. One such operation fusion led to a 30% latency reduction compared to XLA. Modernized graph manipulation workflows with technologies like MLIR.
- **Foundation model support**: Unblocked foundation model adoption by overcoming longstanding limitations due to Tensorflow implementation details. Made numerous contributions to Tensorflow codebase.
- **Developer tooling**: Created tool for inspecting models and providing optimization suggestions. Made tool to verify numerics after graph manipulation, leveraging existing integration testing infrastructure.

TikTok May 2022 - Aug 2022
Software Engineer Intern *Mountain View, California*

- Worked on software defined network, implemented a novel data plane network verification algorithm in C++ from scratch, based on a recent academic paper. Drastically increased the performance of network topology modelling and invariant checks (loop) by 500+ times

Education

University of California, Berkeley Aug 2021 - Dec 2022
Master of Engineering in Electrical Engineering and Computer Science *Berkeley, California*

Peking University Sep 2016 - May 2020
Bachelor of Science in Electronics and Information Science and Technology *Beijing, China*

Academic Experiences

Center for Energy-Efficient Computing and Applications, Peking University July 2020 -- June 2021
Research Assistant *Beijing, China*

- GPU acceleration of elfPlace using CUDA, an algorithm for the placement phase of FPGA physical synthesis. Shortened runtime by 7 times on average.
- Ported the algorithm to an PyTorch based framework, framing this nonlinear nonconvex optimization problem as training a neural network. Used C++ & CUDA extension on Python to speed up critical segments while maintaining low code complexity. Refactored the individual functionalities into “operators”, following the paradigm of high-cohesion low-coupling.
- Resulted in two academic publications in top journals.

Publications

- **elfPlace: Electrostatics-based Placement for Large-Scale Heterogeneous FPGAs**: Yibai Meng, Wuxi Li, Yibo Lin and David Z. Pan. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, 2021
- **Multi-Electrostatic FPGA Placement Considering SLICEL-SLICEM Heterogeneity and Clock Feasibility**: Jing Mai, Yibai Meng, Zhixiong Di and Yibo Lin. *Design Automation Conference (DAC)*, 2022

Updated 2024-09-24, see https://www.mengyibai.com/cv.pdf for current version.