Cunyang Wei

Email: weicunyang@gmail.com | Mobile: (+86) 17719250299
Address: No.6 Kexueyuan South Road Zhongguancun, Haidian District, Beijing 100190, China

EDUCATION BACKGROUND

The Institute of Computing Technology (ICT) of the Chinese Academy of Sciences **09/2020 – 06/2023**
M. S. in Computer Technology GPA: 3.50/4.0

Zhengzhou University **09/2016 – 06/2020**
B. S. in Mathematics and Applied Mathematics GPA: 3.18/4.0

RESEARCH INTERESTS

- High-performance computing

RESEARCH EXPERIENCE

Optimization of LLM Inference Framework on Mobile GPU **07/2023 – Present**
Research Assistant, PerfXlab, Supervisors: Professor Haipeng Jia, Dr. Xianyi Zhang

- Optimized the inference part of Llama 7B on Qualcomm Snapdragon 8gen2's Adreno 740 GPU, with a processing time of less than 2 seconds for the Step 0 phase with 32 token inputs and 8 tokens/s for the Step N phase.
- Enhanced the performance of tall-and-skinny matrix multiplication, a computational hotspot in the step0 phase. It provides up to 4.0× speedups compared to CLBlast, by implementing a carefully planned tiling strategy for more efficient computation and on-chip memory optimization.
- Improved the efficiency of GEMV, a key computational process in the stepN phase, reaching over 90% of peak bandwidth capability.

IrGEMM: An Input-Aware Tuning Framework for Irregular GEMM on ARM and X86 CPUs **10/2022 – 04/2023**
Research Assistant, ICT, Supervisor: Professor Haipeng Jia

- Generated hundreds of highly optimized assembly kernels for diverse irregular GEMM types based on computing templates, the instruction mapping rules between templates and assembly codes, and pipeline optimization strategies.
- Abstracted tiling problems of GEMM into bin packing problems that utilizes a dynamic programming approach to minimum memory access of Irregular GEMM and maximum computational memory access ratio.
- Built a load-balanced multithreaded scheduling framework for processing batch matrix multiplication to achieve the ultimate multi-threaded speedup.
- Implemented a high-performance irregular matrix multiplication library for ARMv8 and Intel cascade Lake architectures.
- Increased the speed-up ratio of irregular DGEMM in a single-threaded environment to 2.3x, 2.7x, and 2.5x in comparison to Intel MKL, ARMPL, LIBXSMM, and BLIS; increased the speed-up ratio of irregular DGEMM in a multi-threaded environment to 3.4x, 14.6x, and 14.3x in comparison to Intel MKL, ARMPL, LIBXSMM, and BLIS.

IATF: An Input-Aware Tuning Framework for Compact BLAS Based on ARMv8 CPUs **10/2021 – 04/2022**
Research Assistant, ICT, Supervisor: Professor Haipeng Jia

- Proposed computing kernel templates for GEMM and TRSM based on the SIMD-friendly data layout and analyzed the compute-to-memory-access ratio to find the optimal kernel size; and optimized instruction selection.
- Carefully designed the data packing kernel so that the memory accesses of the computing kernel are contiguous.
- Proposed an adaptive tuning framework to chooses an appropriate number of matrices for batch operation each time according to L1 cache size and matrix size, and chooses the optimal data packing kernel and computing kernel according to the input matrix properties.
- Increased the speed-up ratio of GEMM and TRSM to 4x and 5x in comparison to ARMPL under double-precision floating-point operation.

LBBGEMM: A Load-Balanced Batch GEMM Framework on ARM CPUs **05/2022 – 10/2022**
Research Assistant, ICT, Supervisor: Professor Haipeng Jia

- Designed high-performance small GEMM kernels without data packaging to greatly reduce the memory accessing overhead.
- Presented a load-balanced multi-thread task scheduling strategy for batch GEMM to improve multi-core performance dramatically.
- Increased the speed-up ratio of DGEMM_Batch to 2.3x for a single thread and 4.2x for 48 threads in comparison to ARMPL.

High-performance Image Processing Algorithms Optimization Based on ARMv8 CPUs **10/2020 – 10/2021**

Research Assistant, ICT, Supervisor: Professor Haipeng Jia

- Sorted image processing algorithms into three types (data irrelevant algorithm, data sharing algorithm and irregular memory access algorithm).
- Built a high-performance image processing algorithms library by writing the underlying code with Arm Neon Intrinsic and optimizing multi-threaded performance with OpenMP.
- Presented optimized image processing algorithm library based on ARMv8 architecture and substantially improved the image processing performance by optimizing the algorithms, memory access, SIMD, and assembly instruction.
- Increased the speed-up ratio of cvtColor, Resize and Filter modules to 1.2x, 2x, and 2x in comparison to the OpenCV algorithms library.

PUBLICATIONS

- **Cunyang Wei**, Haipeng Jia, Yunquan Zhang, Jianyu Yao, Chendi Li, Wenxuan Cao. 2023. *IrGEMM: An Input-Aware Tuning Framework for Irregular GEMM on ARM and X86 CPUs*. IEEE Transactions on Parallel and Distributed Systems (**TPDS**). (Under review)
- **Cunyang Wei**, Haipeng Jia, Yunquan Zhang, Liusha Xu, and Ji Qi. 2022. *IATF: An Input-Aware Tuning Framework for Compact BLAS Based on ARMv8 CPUs*. In 51st International Conference on Parallel Processing (**ICPP**), 2022.
- **Cunyang Wei**, Haipeng Jia, Yunquan Zhang, Kun Li, Luhan Wang. 2022. *LBBGEMM: A Load-Balanced Batch GEMM Framework on ARM CPUs*. The 24th IEEE International Conference on High Performance Computing & Communications (**HPCC**), 2022. (**Acceptance rate 17.6%**)
- Luhan Wang, Haipeng Jia, Yunquan Zhang, Kun Li, **Cunyang Wei**. 2022. *EgpuIP: An Embedded GPU Accelerated Library for Image Processing*. The 24th IEEE International Conference on High Performance Computing & Communications (**HPCC**), 2022.

HONORS AND AWARDS

Outstanding Graduate of Beijing, Beijing Municipal Education Commission	**2023**
Outstanding Graduate, University of Chinese Academy of Sciences (Top 3%)	**2023**
National Scholarship (top scholarship in China), Ministry of Education of the People's Republic of China	**2022**
First Prize Scholarship, University of Chinese Academy of Sciences	**2022**
Merit Student, University of Chinese Academy of Sciences	**2022**
First Prize Scholarship, Zhengzhou University	**2017**
Merit Student, Zhengzhou University	**2017**

EXTRACURRICULAR ACTIVITIES

Session Chair for IEEE HPCC'22 **12/2022**

Academic Conference Host & Coordinator for Conference of China Computer Federation Technical Committee on High Performance Computing (CCF TCHPC) 2021 **10/2021**

PROFESSIONAL SKILLS

- Mastered ARM assembly, X86 assembly, and programming with C/C++.
- Proficient in OpenMP, OpenCL, Arm Neon, Intel AVX512 and etc.
- Solid knowledge in Linux commands, data structure and computer architecture.