Ke Tang

+8618911648482 | 509634578tk@gmail.com
https://www.linkedin.com/in/ke-tang-582656286/

PROFESSIONAL EXPERIENCE

Beijing Megvii Technology Co., Ltd Apr 2022 - Jun 2024
High Performance Computing Engineer MegEngine Beijing
Project A- LLM Inference Framework(open source) (2024.1-2024.4)

- Used I8mm to calculate the ARM operator for int8 multiplied by int4, utilizing NEON intrinsics to optimize the matrix multiplication kernel. For problem size (M = 20, N = 11008, K = 4096), achieved a performance increase of approximately 38% (21540ms vs. 34999ms, -O3 build, mi12 pro, Armv8a), compared to dotprod.
  - Status: Merged into internal GitLab, not yet synchronized to GitHub.
  - Responsibilities: Personally implemented the foundational computation code and optimized it through performance testing iterations. Focused on data rearrangement, reducing the vld instructions required for vmmlaq calculations. Used zip to increase data width and reduced the vst instructions during write-back, resulting in significant performance improvements.

Project B- MPP Inference Framework(DeepLearning inference) (2022.12-2024.6)

- Implemented a correctness verification module in C++ to verify the accuracy of various platform operators (e.g., computer-vision preprocessing, model inference, video image codec), based on gtest. Simplified the code required for testing newly added platforms, reducing redundancy.
- Integrated convert_color, crop_resize_pad, and warp_affine operators in x86 OpenCV, with support for both RGB and YUV formats (NV12, YU12).
- Integrated ffmpeg for H.264 video decoding and wrote a basic implementation for an H.264 stream parser, removing anti-competition bytes and SODB.
- Integrated Atlas-related operator compilation and added operator support.

Project C- MegEngine(open source) (2022.4-2023.5)

- Added the Norm operator, implementing CUDA/x86 forward calculations with support for FP16/FP32, including inf/-inf/0/1/p norms. Combined with the existing reduce operator implementation, achieving performance on par with PyTorch.
  - Commit: https://github.com/MegEngine/MegEngine/commit/b55942a94df02c59b1bdb7c6ec8c09baa98c9c1a
- Collaborated on the implementation of the Region Restricted Convolution operator using CUDA C, and later independently migrated it to CUTLASS (v2.8).
  - Used CUDA C to implement Fprop & Dgrad kernels, and encapsulated the calling interface from C++ to Python in MegEngine. Wgrad is completed by colleagues in CUTLASS(v2.8)
  - Updated the Fprop & Dgrad implementation to CUTLASS, utilizing SIMT instead of MMA or WMMA PTX.
  - Status: Merged into internal GitLab, not yet synchronized to GitHub.
- Added additive noise types: Gaussian, Laplace, and Poisson. Implemented a Python-side operator interface.
  - Commit: https://github.com/MegEngine/MegEngine/commit/ba9f67eb49e46d2a81b353b71ca3270e7fbdabc8

Beijing Megvii Technology Co., Ltd Jul 2021 - Apr 2022
DeepLearning Engineer Person Re-identification Team Beijing

- Trained and deployed human body quality models using the End-to-End Multi-Task Learning with Attention network to handle low-quality filtering tasks. Deployed to the business line.

Aerospace Long March Rocket Technology Co., Ltd Jul 2014 - May 2018
Technology management Confidential Technology Division Beijing

- Reviewed and managed projects involving confidential technologies, collaborating with development providers.

SKILLS LIST

A. Familiar with the ARM neon intrinsic and assembly code
B. Proficient in RV32I architecture, QEMU OS
C. Proficient in C/C++(~14)/CUDA/Python
D. Other

- [https://github.com/ijpq/MyCS61C]
  - Use logisim to implement RV32I two-stage pipeline cpu. Use RISCV asm to implement relu,argmax,dotprod,matmul, and so on.
- [https://github.com/ijpq/MIT_ OS]
  - All QEMU OS labs except FileSys
- [https://github.com/ijpq]
  - github home page

EDUCATION

Beijing Jiaotong University 211 Double 1st-Class Sep 2018 - Jun 2021
Computer Science Master Beijing
GPA: 3.2/4
second-class scholarship (2018)
third-class scholarship (2019)

Beijing Institute of Electronic Science and Technology Sep 2010 - Jun 2014
Computer Science Bachelor Beijing
GPA 3.0/4
school-level second-class scholarship (2011,2012)

MISCELLANEOUS

- **Languages:** English(IELTS-G, 6.5)