Frank Young

2679frank@gmail.com | github.com/frank2679.io | Linkedin/FrankYoung

Personal Summary

- Over six years of experience in heterogeneous platform (CPU/GPU/DSP/NPU) operator acceleration library, including DNN, BLAS, FFT, RAND. Deeply involved in 0-1 chip software and hardware development.
- Two years of team leadership experience, leading a team of 5 people.
- Strong learning ability, pursuit of excellence, and a passion for triathlons.

Education

National Taiwan University |Telecom Institute| *Master's Degree* 2014.09—2017.01

[1] **Yang, Hang**, D.-J. Deng, and K.-C. Chen, "On energy saving in ieee 802.11 ax," *IEEE Access*, vol. 6, pp. 47 546–47 556, 2018.

[2] **Yang, Hang**, D.-J. Deng, and K.-C. Chen, "Performance analysis of ieee 802.11 ax ul ofdma-based random access mechanism," in *GLOBECOM 2017-2017 IEEE Global Communications Conference*, IEEE, 2017, pp. 1–6.

Chongqing University |Communication Engineering| *Bachelor's Degree* 2010.09—2014.06

Top 5% of the program, outstanding graduate, multiple scholarship recipient.

Technical Skills

- **Technology Stack**: Familiar with operator optimization, deep learning model deployment optimization, including CNN, large language models.
- **Tools**: Proficient in C++, Python, CUDA, CMake, Verdi, Vim, Git, JIRA, and working in an English environment.

Work Experience/6 Years

Leading GPU Manufacturer in China/3 Years |*Software Engineer/Team Leader* 2021.1—Present

- Developed AI/HPC acceleration libraries from scratch based on GPGPU platform.
- Led and managed a team of around 10 people, providing guidance and direction for project development.

Leading company of video surveillance products/3.5 Years |*Software Engineer* 2017.7—2020.12

- Developed high-performance convolutional neural network (CNN) library for heterogeneous platforms, with a focus on evaluating chip performance.
- Held end-to-end responsibility for algorithm-side projects, deeply analyzing business requirements, and building efficient application solutions to accelerate intelligent algorithm implementation.

Project Experience

BLAS Library in C++ |*Development and Maintenance from Scratch* 2021.11—Present

- Led the project, responsible for design, construction, development, testing, CICD, documentation, and management.
- Optimized the performance of various large language model GEMM scenarios, benchmarked against A100.
- Designed and implemented kernel selection algorithms.
- Implemented operator registration framework, logging system, and core operator development.

Ultimate Optimization of GEMM |*based on verdi/zebu* 2022.9—2022.11

- Handwritten assembly to achieve ultimate performance GEMM on Vcore, achieved full utilization for specific shapes by observing waveforms through Verdi.
- Implemented simulated FP32 GEMM algorithm based on Tcore, achieved full utilization for specific shapes, and outperformed A100 in performance.

DNN Framework Development |*Self-defined yaml to serialize operator and graph* 2022.3—Present

- Developed a deep learning framework based on Caffe.
- Self-defined operator/graph expression based on YAML referencing ONNX.
- Compared precision with mainstream industry solutions and designed custom precision comparison solutions.
- Deployed precision verification, and performance dashboard.