Yihua Liu

fwcnigo@gmail.com · (+86) 150-5345-8529 · Wendyl42

.

Education

Peking University, Beijing,China 2018.9– 2023.7

Bachelor of Science Data Science and Big Data Technology(Yuanpei college)

Experience

Wizard Quant 2023.7 – 2025.7

AI Inference Engineer

Core developer of the company internal AI Inference Framework based on C++17, supporting **High-Frequency Trading Model** running on x86-64 CPU, which provides:

- Static Computational Graphs optimized by a series of Optimization Passes
- Highly optimized CPU kernels based on SIMD instructions
- Multi-Level Benchmark Tools, Unit Testing, and Documentation

Institute of Computational Linguistics, Peking University 2021.8 – 2022.3

Research Intern

Enhanced the performance of BERT initialized by Transformer Encoder, including reproducing paper and modifying the model architecture, achieving **state-of-the-art (SOTA) performance**.

Wangxuan Institute of Computer Technology, Peking University 2022.10 – 2023.5

Research Intern

Extended the application of *MaskGIT: Masked Image Generative Transformers* from 2D to 3D Point Cloud Completion.

Project

CPU Kernel Libraries 2023.8–2024.8

- GEMV: SIMD(AVX/AMX), Mixed Precision(BF16), Kernel Selection mechanism
- Element-Wise Calculation (Activation, BatchNorm, etc.)
- general Tensor Calculation(Arithmetic, Slice, Reshape, Copy, Repeat, etc.)

Optimization Components 2024.6–2025.7

- General Optimization Passes: Constant Folding / Operator Fusion / Memory Layout Optimization
- Memory Optimization: Model weights layout / Parameter Sharing between multiple model instance / Cache Warming mechanism
- Model-Specific Optimizations: Layer Input Segmentation

CPU Model Development 2024.10-2025.7

- High-Frequency Trading Models (latency <= 100us) including GNN, RNN, etc.

Profiling Tools 2024.5-2024.7

- Measure Performance of operators based on our kernels and other libraries (OpenBLAS / MKL / AOCL)
- Compare Performance between different versions of specific model, and report profiling results

Skills

- **Programming language**: experienced in C++ / C / Python, comfortable with Rust / Golang
- **Languages**: Mandarin(Native), English(TOEFL: 101)