Zhicheng Fang

+86 19967447386 | fangzhicheng1115@gmail.com | cyberkillor.github.io

Education

Shanghai Jiao Tong University 2018/09 - 2022/06

- Bachelor of engineering in Computer Science and Technology
- skills：C++, Python, CUDA, TensorFLow, PyTorch, Machine Learning, HPC

Work Experience

TEMU Shanghai, China

Software Engineer, fulltime, searchrec 2022/11 - Present

- Transfer from Pinduoduo domestic rec group, focus on online rank service and offline model training
- **Rank Service** Support rank framework using OnnxRuntime as inference backend to serve a series of PyTorch models online. Optimize TensorFlow compute graph, split user and item features on LHUC structure, CVR model infer latency reduced 25%. Support MatMul + BiasAdd + BatchNorm op fusion pattern. Based on FasterTransformer, develop Llama TensorRT plugin, PyTorch op and dynamic batching, support Llama-7B generating embeddings online. Develop flame graph visualization based on brpc builtin service and perf.
- **Query Process Service** Develop C++ ver. tokenizer based on HuggingFace transformers open source repository, including Marian / BPE / Clip / Llama Tokenizer. Develop and optimize Beam Search sampling on GPU, translation model latency reduced 45%.
- **Sparse Training Framework** Investgate the principles of open source training framework such as HugeCTR SOK and Deeprec. As one of the core developers, develop and optimize the sparse training framework from scratch. Support single-GPU and multiple-GPU on single worker training, with performance exceeding HugeCTR SOK. Responsible for migrating existing CTR model training tasks from PS-Worker CPU architecture to gpu framework based on model and data parallelism. Develop feature filter and fp16 embedding.

Pinduoduo Shanghai, China

Software Engineer, fulltime, xrec 2022/07 - Present

- Focus on online serving, including Feature Generation and TensorFlow model inference.
- **Feature Engineering** Support int32 type feature computation. Optimize string type feature hashing, migrate to fore stage, increase overall qps by 4%. Develop FG framework export TF format graph, visualized bt Netron. Support using brpc builtin service to download TensorRT sub-graph dumped online, make it convenient for debugging.
- **TensorFlow Inference** Based TensorRT Plugin and cuBLAS BatchedGemm, optimize by MLP operation fusion, increase qps by 15%, save 20% of GPU usage on a single machine. Support transformer model on GPU, fix TF grappler shape inference fail problem.

Software Engineer, intern, xrec 2021/12 - 2022/02

- **Feature Engineering** Optimize FG compute graph by fusing Seek op for position map lookup with different features.

Alibaba Cloud Hangzhou, China

Data Warehouse R&D Intern, Hologres 2021/06 - 2021/09

- Develop Python multithread scripts, implement failover tests for distributed nodes. Design a simple mock filesystem for error injection.

Research

John Hopcroft Center for Computer Science 2020 - 2022

Improving the robustness of analog deep neural networks through a Bayes-optimized noise injection approach

- N Ye, L Cao, L Yang, Z Zhang, **Z Fang**, Q Gu, GZ Yang | Communications Engineering

BayesFT: Bayesian Optimization for Fault Tolerant Neural Network Architecture

- N Ye, J Mei, **Z Fang**, Y Zhang, Z Zhang, H Wu, X Liang | 2021 58th ACM/IEEE Design Automation Conference