方 志 成

github.com/cyberkillor | linkedin | scholar | fangzhicheng1115@gmail.com | 19967447386

教育经历

上海交通大学 2018/09 - 2022/06

- 本科，计算机科学与技术
- 参与高性能计算中心和深度学习科研（导师：叶南阳）的工作，热衷于GPU、MLSys以及深度学习。

工作经历

TEMU **上海**

服务端研发工程师，全职，搜索推荐 searchrec 2022/11 - Present

- 从拼多多国内业务转岗，负责搜推模型的在线serving和离线training。
- **RTP**服务（**Rank Service**）支持现有框架以OnnxRuntime为推理引擎，在线使用GPU/CPU推理算法的一系列PyTorch模型。在线TensorFlow计算图优化，支持LHUC模型结构上的结构化特征，相关模型推理平响降低**25%**。支持在线MatMul + BiasAdd + BatchNorm的算子融合。基于FasterTransformer封装和开发，实现Llama TensorRT插件、PyTorch OP以及动态Batching，支持在线使用Llama-7B模型生成embedding。
- **QP**服务（**Query Process, NLP**）基于HuggingFace transformers开源库从零实现Tokenizer的C++版本，包含：Marian / BPE / Clip / Llama Tokenizer等，并通过brpc、TF封装进QP服务。实现Beam Search的GPU版本，使得翻译服务平响降低**45%**。
- 稀疏训练框架 调研HugeCTR SOK、Deeprec等开源稀疏训练框架原理。作为主要开发之一，从零开发训练框架，目前支持单机单卡，单机多卡的训练，且性能超过SOK。负责迁移现有CTR模型任务 (内部定制的TF1.12，PS架构) 至新训练框架 (稀疏 + TF2.12，DP + MP) 进行测试。

拼多多 **上海**

服务端研发工程师，全职，推荐 xrec 2022/07 - 2022/11

- 参与推荐模型的在线serving工作，包含Feature Generation和TensorFlow模型推理。
- **FG框架** 支持int32类型特征计算。将string类型特征的id哈希，从TF迁移到FG中，使整体**qps**提升**4%**。支持FG框架导出TF格式的计算图，借助Netron可视化。支持通过brpc builtin service，下载在线机器保存的TensorRT子图，方便debug。
- **TensorFlow推理** 在CVR模型上，使用TensorRT Plugin和cuBLAS BatchedGemm，实现模型多头的mlp算子融合，**qps**提升**10%**，单机节省GPU 20%使用率。支持Attention模型上GPU，修复TF grappler无法推理Tensor形状的问题。

服务端研发工程师，实习，推荐 xrec 2021/12 - 2022/02

- **Feature Engineering** 抽取seek操作并融合，减少查表操作。参与FG到TF serving的开发。

阿里云 **杭州**

基础平台研发工程师，实习，数据仓库 Hologres 2021/06 - 2021/09

- 编写Python多线程脚本，实现对分布式节点的failover测试。设计一个简单的mock filesystem实现错误注入。

科研经历

John Hopcroft Center for Computer Science 2020 - 2022

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

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

BayesFT: Bayesian Optimization for Fault Tolerant Neural Network Architecture Link

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