许啸宇 2023.7.11

Blog: strint.github.io · Github: github.com/strint · 15210835395 · xiaoyulink@gmail.com

工作经历

一流科技 2020 年 7 月 – 至今

高级工程师 OneFlow 框架静态图的核心维护和开发者之一

腾讯科技 2017 年 4 月 – 2020 年 6 月

工程师 视频推荐引擎开发

教育经历

北京邮电大学，计算机技术，硕士 2014 年 – 2017 年

北京邮电大学，计算机科学与技术，学士 2009 年 – 2013 年

项目经历

深度学习框架 OneFlow 开发和优化 2021 年 4 月 – 2023 年 5 月

一流科技，高级工程师

OneFlow 的静态图执行的核心维护和开发者之一，优化对 LLM / Stable Diffusion 项目的支持：

- 大模型的执行图分离编译功能，在主节点做逻辑图编译，在每个 worker 节点分离编译执行图，大规模场景编译效率提升 10 倍以上；
- OneFlow Stable Diffusion 静态图实现的主要维护者之一，实现执行图加载以支持离线编译，实现多逻辑优化图共享以支持 Dynamic Shape；
- 在静态图上实现 1D、2D 的 ZeRO Optimizer 优化，支持 NCCL Send/Recv 算子以实现不规则通信，增加参数切分后的内存池清理；
- 实现大模型的多机多卡的分片、非对称的模型保存和加载功能，实现 tensor offload 到 CPU 的功能；
- 重新设计和实现静态图编程接口nn.Graph，支持平滑的动静转换、支持平滑的实现优化、利用 Python EvalFrame 获取 Python 代码位置支持良好的 Debug 功能；

深度学习框架 OneFlow 开发 2020 年 7 月 – 2021 年 3 月

一流科技，高级工程师

提高 OneFlow 框架的开发易用性 (C++, Python)：

- 重构 Operator 注册，实现 Op 注册 Manager 来统一管理前向、后向、Kernel 注册；采用 lazy evaluation 支持后向构图避免复杂的后向构图实现；
- OneFlow 运行时支持执行 Python kernel，支持 c++ 执行 Python （PyTorch）代码；
- 易用性性调研，认识 OneFlow/TensorFlow 这样的声明式编程在控制粒度上的局限性，推动 OneFlow 兼容 PyTorch 接口；

长视频推荐引擎优化 2018 年 7 月 – 2020 年 6 月

腾讯科技，工程师

开发和优化视频推荐系统主调度引擎（C++），支持腾讯视频 1.5 亿日活用户的长视频推荐服务：

- 独立负责 Feeds 流推荐系统、相关推荐系统、推荐重排子服务；
- 实现重排子服务支持多个推荐服务；
- 统一本地和远程索引实现召回热更新和容错；
- 索引组件无锁化改造减少约 50% 内存消耗；
- 引擎全异步化改造提高执行效率；

长视频推荐引擎开发 2017 年 4 月 – 2018 年 6 月

腾讯科技，初级工程师

开发和优化腾讯视频长视频推荐服务:

- 推动上线 LightGBM 模型，取代原有的 XGBoost 模型 (Python)，模型训练时间可以缩减 75% 因此可以训练更大模型；
- 负责腾讯视频相关性推荐引擎开发 (C++)；
- 负责实现统一的推荐协议以统一长短视频推荐服务；

深度学习框架原型系统 2016 年 10 月 – 2016 年 12 月

MSRA，*实习生* 导师：袁进辉

深度学习框架的执行图编译阶相关研发工作：

- 完善前向执行图的生成、重构了 Segment Dag；

技能

- 编程语言: C/C++, Python, CUDA
- 开发工具: Git, Linux
- 外语：英语六级