

宋宝玉

联系方式: 15215365890 邮箱: songbaoyu1996@foxmail.com 期望岗位: 模型相关岗位 (C++)

教育经历:

聊城大学(通信工程, 学士)

2015-2019

北京工业大学(信息与通信工程, 硕士)

2019-2022

工作经历:

浪潮软件集团有限公司

2022.7.12-2026.7.25

项目经历:

某市政府视频监控项目:

具体责任:模型部署, 视频文件硬解码。

负责昇腾 AI 服务器 (Atlas 800I A2) 基础环境的搭建, 包括驱动与固件安装。

参与语音大模型在昇腾 AI 推理卡 (310P3 96G*2) 上进行分布式部署, 辅助进行违规判定。模型使用 FP16 精度, 开启多个服务端口, 请求通过 Nginx 均衡负载, 实现约 32~35 倍音频分析速率。

负责昇腾 310P3 上的硬解码模块, 基于昇腾 DVPP+VPC SDK 库, 使用 C/C++ 开发, 实现对 H264/H265 视频文件的多线程快速解码并后置图片的相关前处理操作, 上线解码模块可达单路 300+FPS 并稳定运行。

负责昇腾 om 模型的转换。通过使用 AIPP 将模型输入中的 resize/padding/归一化等操作合并入模型, 编写 Python 脚本将 NMS 操作合并入模型, 达到减少 CPU 利用率, 提高模型推理速度的目的。

多异构硬件 AI 推理服务平台: 国内线缆行业缺陷检测领域, 基于视觉。

参与推理服务 C 端与 S 端开发。C 端基于 JSON 配置驱动的 Pipeline 引擎, 支持多种模型组合, 只负责模型前后处理; S 端通过条件编译封装多种 AI 芯片 SDK, 同时使用工厂模式结合 libevent 框架实现多任务注册与分发, 只负责模型推理。

具体责任: 推理引擎开发, 模型部署, 边缘设备。

1.项目构建: 项目使用 CMake 跨平台构建, 支持条件编译选择硬件后端, 通过 Pybind11 封装 Python 接口交由 Python 中间件调用。项目使用 CPack 生成 deb 安装包, 配合 shell 脚本实现基于 Docker 的一键部署。

2.性能优化: 使用线程池复用线程资源, 使用对象池代替频繁 new/delete 减少开销, 使用异步拷贝与流配合页锁定内存减少推理耗时, 使用 CUDA 编写 NMS 等后处理减少 CPU 利用率, 使用短自旋休眠替代原生阻塞, 使用 C++11/14/17 新特性。

3.多硬件适配: 包括寒武纪 MLU270/MLU370/MLU365、昇腾 310P3/910B4、算能推理卡、NVIDIA 30/40/50 系显卡等。使用 CNRT/CANN/BMRuntime/TensorRT/等相关 SDK 编写推理模块并集成到推理服务 S 端中, 并设计板卡调度算法进行板卡调度, 减少硬件负载倾斜。

4.模型部署与算法集成: 主要负责传统视觉模型 (YOLO/STDC/ResNet/OCR/等) 以及视觉大模型 (DINOv3) 的转换、量化、加密与前后处理编写。模型转换主要包括 ONNX 模型转化为各板卡适配的格式 (magicmind/om/bmodel/engine/rknn 等), 模型皆量化为 FP16 格式, 兼顾推理速度与精度, 加密采用国密 SM2, 模型前后处理使用 OpenCV 编写 (尺度变换、色域转换、归一化、通道置换等)。

5.边缘设备适配(ARM): 包括 Jsteson/算能 SE5/SE7/瑞芯微 3576。 负责边缘设备的推理服务开发, 采用极简架构, 构建多线程 Pipeline, 利用设备特点实现数据零拷贝, 同时使用 SDK 进行前处理操作以减少 CPU 利用率, 提升推理速度, 最后通过 CMake 与 Pybind11 封装为动态库, 解决开发板性能较低的问题。

掌握技能:

熟悉大模型推理优化与部署:

1. 熟悉大模型量化、剪枝、蒸馏、微调等常见推理优化手段,熟悉模型分布式部署。
2. 熟悉主流大模型架构, 了解 VLLM 原理, 使用过 TensorRT-LLM。
3. 熟悉大模型训练与推理流程。

具备使用 C++语言在 Linux 系统进行项目开发的能力:

1. 熟练使用 C++11 新特性, 能够使用 CMake 构建较复杂项目, 使用 Git 协同开发。
2. 熟练使用常用数据结构,熟悉常用设计模式。
3. 熟练使用 Linux 系统, 熟练 Docker 容器技术, 熟悉 Linux 命令。
4. 熟悉开发辅助工具, 能够快速定位项目问题 (GDB/ Valgrind/Vim/Shell 等)。

具备视觉模型应用的能力:

1. 熟悉 OpenCV, 能够使用 OpenCV 进行图像前后处理操作。
2. 熟悉 CUDA, 熟悉 GPU 架构, 能进行算子编写 (GEMM, 卷积, 规约等), 并对 CUDA 性能优化了解, (共享内存、线程束分化、Bank Conflict、异步执行与流) 了解高性能库 (CUTLASS), 使用过 Nsight 工具。
3. 熟悉相关 AI 推理卡 SDK 接口 (CNStream、TensorRT、CANN、RKNN、SOPHON 等)。