5349f67a86444afb1Xd52ty9EVBTwYS_Vv6aWOGrlvbTNhFm2A~~

邹加贝

男 | 年龄：26岁 | 13141055288 | 937128703@qq.com

4年工作经验 | 求职意向：cuda算子开发 | 期望薪资：25-30K | 期望城市：北京

个人优势

1.具备 cuda 编程经验，熟悉常见 CUDA 优化手段，包括 memory coalescing、bank conflict等。

2.具备vllm大模型部署经验，实现tp、pp等部署以及量化等优化、并进行性能分析。

3.具备了解k8s使用经验，可编写yaml文件实现k8s部署。

4.具备c++、Python 编程经验，熟悉 Linux 操作。

工作经历

北京航天拓普科技有限公司 脚本开发 2021.07-2024.06

在工作期间主要负责仿真想定的脚本编写、项目开发及文档编写工作

1.在项目组完成仿真开发工作，使用脚本语言编写仿真想定代码；

2.工作期间自学的 CUDA，编写了 GPU 并行计算大规模矩阵相乘的优化项目；

浙江省公众信息产业有限公司 AI工程师 2024.06-2025.06

在工作期间主要vllm模型部署优化及模型推理性能分析

1. 部署vllm大模型推理并优化，使用nsys工具分析优化效果。

2. 构建包括算子异常等推理异常case、使用nsys工具分析异常情况。

3. 部署实现ring、Ulysses并行、dp并行，使用nsys通信算子数据。

项目经历

调用GPU并行计算大规模矩阵相乘 开发 2024.01-2024.02

使用 NVIDIA CUDA 调用 GPU 并行计算大规模矩阵相乘的项目，利用 GPU 的强大并行计算能力，显著提升了计算性能。 通过使用共享和代码，成功实现了在 RTX2080显卡上特定 case 下的带宽利用率提升。 在 RTX2080显卡上，特定 case 下带宽R/O达到了383GB/330GB，L2带宽达到了1700GB 。 成功实现了95%的硬件峰值性能，超越了 cublas 库的表现，为项目带来了显著的性能提升。

vllm调优 开发 2024.07-2025.06

1.成功完成 vllm 模型的张量并行（TP）、流水线并行（PP）数据并行（DP）、pd分离等多模式部署，并通过模型量化技术实现大模型的推理优化。

2. 设计并构建了多项测试案例，包括算子异常、CPU与GPU资源故障及通信异常等，以全面评估模型的稳定性和性能。

3. 部署实现ring、Ulysses并行。

3. 利用Nsys及perfetto工具对以上case进行性能数据采集与分析，识别TP、PP、量化、算子异常等case对推理性能的影响。

教育经历

北京信息科技大学 本科 计算机科学技术 2017-2021