34b0088b7b76a9721Xd72Ni5FVNQwIS5Wf-WWOGrlvbTNhFm2A~~

曹振杰

男 | 年龄：28岁 | 18753912032 | 741266385@qq.com | 籍贯：临沂

2年工作经验

个人优势

精通：Bert、Transformer、GPT、等NLP基础框架算法；

熟练：llama，deepseek,qwen，百川等开源大模型；

熟练：ollama，vllm，mindie等推理引擎部署大模型；

熟练：Pytorch 等常见的开发框架；

熟练：使用 AI 智能体与应用开发平台（Coze 和 Dify），创作有用、好用的智能体和应用

熟悉：Python 、linux 编程命令；

熟悉：国产芯片昇腾相关组件；

熟悉：使用 Docker 和自定义 Dockerfile 对大模型进行容器化封装，实现可复用、可扩展的部署环境；

熟悉：大模型的预训练、微调和大模型的部署，通过轻量化微调、模型量化剪枝等技术，辅助优化模型工程化能力

工作经历

华为技术有限公司 大模型部署工程师 2023.11-至今

1. 负责把客户的AI算法模型迁移到昇腾技术体系（昇腾Ascend，华为自研人工智能处理器）

2.基于华为推理框架/vllm框架/ollama框架部署llama,deepseek,qwen等大模型等客户业务场景；

3.基于客户业务场景，使用modelmate/dify/maxkb等软件搭配本地服务化的openai接口，完成RAG的开发与使用；

4.基于客户数据集，使用mindspeed框架完成llama,deepseek，qwen等大模型的预训练、全参微调/Lora微调/QLora微调；

5.基于客户现场资源与需求，规划训练与推理所需要的卡数资源，对模型进行量化，且跟踪维护客户在预训练、微调、推理过程中遇到的问题；

6.对昇腾技术体系在不同场景下的应用效果进行跟踪评估，撰写专业的评估报告，为后续项目的开展提供数据支持与经验借鉴；

7.针对迁移过程中出现的技术难题，开展专项技术攻关，与研发团队紧密协作，确保问题及时有效解决。

齐鲁理工学院 教师 2023.06-2023.09

1. 担任大学生教授计算机相关课程,带领本科生参加计算机相关的比赛,如互联网+,创新创业大赛 ；

2. 为本科生计算机相关的考试进行培训.如计算机等级考试,计算机软考。

深圳光沦科技有限公司 算法工程师 2021.12-2022.07

1. 负责计算机视觉、机器学习(包括深度学习)算法的开发与性能提升，推动计算机视觉、机器学习算法在众多实际应用领域的性能优化和落地

2. 在工业场景中，运用图像检测和分割技术,对目标进行识别抓取，跟进领域前沿算法,保持算法在工业界的领先性；

3. 负责对项目中的算法进行优化创新,并进行专利,软件著作权，进行论文的书写，关注计算机视觉、人工智能前沿技术,实现论文代码的复现。

项目经历

哈尔滨移动万卡项目 大模型部署工程师 2025.02-2025.05

（1）基于客户昇腾设备，部署deepseek-qwen-32B，deepseek-llama-70B，deepseek-R1，deepseek-V3,QWQ-32B等模型，使用benchmark工具进行精度与性能测试；

（2）完成DeepSeek系列模型、QwQ模型、Qwen等模型的W8A8,W8A16,W4A16的量化，采用离群值抑制方法去除异常值，规避量化后精度下降的问题；

（3）基于昇腾设备搭建dify工具，与已部署的模型进行搭配使用，完成RAG的搭建，基于客户的需求搭建工作流（例如：图生文，pdf识别，联网搜索等）；

（4）基于客户的数据集与业务场景，完成deepseek-llama-70B的微调，适应客户的业务场景，当前多部门已投入使用。

中华人民共和国审计署项目 现场实施工程师（FAE） 2025.02-2025.02

内容:

（1）基于客户昇腾设备，部署deepseek-llama-70B模型，通过后台调用openai接口，接入客户chatbox软件，完成智能对话功能；

（2）基于部署deepseek-llama-70B模型，完成性能/精度/并发的测试，输出文档交付给客户；

（3）现场部署deepseek-qwen-32B,deepseek-qwen-14B,deepseek-671B，使用benchmark工具进行精度测试，向客户展示每个模型的智能问题的准确度

业绩:

北京智谱项目 大模型部署工程师 2024.08-2025.01

（1）在北京智谱现场开展 Cogvlm2 - 40B 模型从 GPU 至昇腾框架的迁移适配工作，全方位测试其精度与性能状况，对多种不亲和算子进行替换与调整，精心优化性能与精度参数，保障模型于昇腾框架下能够稳定且高效地运行。

（2）于北京智谱现场实施 ChatGLM - 32B、ChatGLM - 130B 模型的部署工作，协同进行性能与精度的深度优化，严谨有序地开展精度和性能测试流程，最终输出全面且专业的交付报告，为项目成果的成功交付给予坚实有力的支撑。

（3）支撑客户 ChatGLM - 32B、ChatGLM - 130B、ChatGLM - 230B在迁移后与量化后的存在精度问题，且使用性能工具定位性能瓶颈；

（4）在北京智谱现场与客户展开深入的沟通交流，精确洞悉客户需求的关键要点，迅速精准定位并妥善圆满地解决客户在项目推进过程中所面临的各类难题，切实有效地提升客户满意度并推动项目高效向前推进

深圳微众银行 大模型部署工程师 2024.05-2024.07

（1）现场开展 gte 模型从 GPU 至昇腾框架的迁移适配工作，全面测试其精度与性能表现，针对多类不亲和算子予以替换及修改，精细调优性能与精度指标，确保模型在昇腾框架下稳定高效运行。

（2）现场实施 Longformer、bigbird 模型的部署作业，协同推进性能与精度的深度调优，严谨开展精度和性能测试流程，最终输出详尽且专业的交付报告，为项目成果交付提供有力支撑。

（3）现场高效完成 Qwen1.5 - 72B、Llama3 - 8B 于华为云的部署任务，并及时输出规范的交付报告，有力保障云服务的顺利上线与稳定运营。

（4）现场与客户深度沟通交流，精准把握客户诉求核心，迅速定位并妥善解决客户在项目进程中遭遇的各类问题，有效提升客户满意度与项目推进效率。

上海天翼云 大模型部署工程师 2024.01-2024.05

（1）现场协助 llama2、qwen、sd 等模型向昇腾体系的迁移适配进程，凭借对昇腾架构的精准洞察，确保迁移流程高效流畅，达成模型性能的卓越优化。同时，能够即时响应并妥善处理客户在昇腾应用中遭遇的各类问题，为业务在昇腾平台上的稳固推进保驾护航。

（2）现场协助 chatglm 模型在 1024 卡与 512 卡配置下的测试工作，参与设计并提出针对大规模并行计算性能评估优化的可行

性方案。

（3）对项目中的模型交付件展开全面且细致的系统梳理，运用清晰严谨的逻辑架构以及标准化的格式规范，精心打造出一套高质量的交付文档体系，为项目的顺利交付提供不可或缺的助力，有力保障项目全流程的完整性与专业性。

基于扩散模型的医学图像生成

2023.06-2024.01

1.使用数学模型来模拟疾病过程或医学图像的生成过程。基于物理或生物过程，并根据这些模型的特点进行参数调整和计算,生成合成的医学图像。

2.在项目开发中,进行专利,软件著作权,论文的书写与投递，与现有的图像生成网络模型实现代码复现,进行对比实验

基于双目相机的水果智能采摘机器人

2022.02-2023.06

1. 与高校展开合作,使用智能小车搭配机械臂对水果进行采摘,负责视觉算法的编写,控制机械臂抓取;

2. 在项目开发中,进行专利,软件著作权,论文的书写与投递。

基于双足热图的糖尿病足患病区域检测

2021.09-2022.06

1. 与医院展开合作,采集数据,后使用深度学习算法进行脚部识别,获取脚步轮廓,后进行不对称分析,获取患病区域,与医生进行交流沟通,探索AI 与糖尿病足结合创新点。

基于X光小牙片填充状态智能化检测

2021.09-2022.01

内容:

1. 与医院展开合作,采集数据,后使用深度学习算法进行识别,对牙齿中的小根管的填充进行判断,得出小根管是否过满、欠满,并且与医生进行交流沟通,探索AI 与口腔医学结合的方法。

业绩:

教育经历

济宁医学院 本科 计算机科学与技术 2016-2020

重庆师范大学 硕士 计算机应用技术 2020-2023