

曾梓健

求职意向：算法工程师

学历：硕士

工作年限：1 年

籍贯：四川·成都

联系方式：18302862864(微信同号)

电子邮箱：zzj18302862864@163.com



简介

参与 2 项省级重点及多项科研/商业项目，在深度学习领域具备扎实的模型开发与工程实践经验。熟悉大模型的参数高效微调（PEFT）与性能优化，并具备从模型训练、微调到高性能部署（如模型量化、推理加速）的完整技术闭环能力。

教育背景

重庆交通大学 | 硕士 | 市政工程 | 研究方向：智能决策与优化算法(智慧城市内涝仿真应用) | 2021.09 - 2024.06

- 荣誉：校一二三等奖学金、第三届水科学数值模拟创新大赛 | GPA: 3.6/4.0 (专业 Top 5%)
- 学术：发表 SCI 一区 1 篇，核心论文 2 篇，申请专利 1 项，省重项目 2 项

西华大学 | 本科 | 水利水电工程 | 研究方向：水动力学 | 2017.09 - 2021.06

- 荣誉：院一二三等奖学金 | GPA: 3.5/4.0 (专业 Top 10%)

技术栈

- 编程与系统：Python, Linux, Docker, Git
- AI 框架与平台：PyTorch, llama-factory, Huggingface, TensorRT, Ollama, Dify, Coze
- 大模型：
 - 架构与平台：Agent (LangChain/LangGraph) | RAG (Chroma/Milvus) | Prompt
 - 模型：PEFT (LoRA/QLoRA/Xtuner) | 部署 (Ollama/vllm/LmDeploy) | Transformer (Bert/GPT/T5)
 - NLP：文本分类、命名实体识别、情感分析等经典 NLP 任务
- 计算机视觉：
 - 经典网络：CNN (VGG/Resnet) | RNN/LSTM | GNN (GCN)
 - 模型：YOLO 系列(目标追踪/实例分割/姿态检测)
- 训练/部署：
 - 分布式训练 | 蒸馏 | 量化 | 剪枝 | TensorRT 加速 | Docker 容器化 | API 封装

工作与项目经历

北京京能清洁能源西南分公司 | 算法工程师

2023.11 - 2024.12

➤ 职责:

- 主导能源场站（水电、风电）无人化运维的 AI 视觉算法研发与落地；
- 协作跨领域专家与工程团队进行系统开发；
- 辅助系统部署，完成从算法到产品的闭环交付。

➤ 项目一：水电站无人值守智能防钓鱼告警系统

2024.02 - 2024.04

- **目标：**利用公司现有的 YOLOv8s 人员检测模型作为预训练权重，构建适用于防钓鱼场景的基线模型，精准识别水电站禁区内的非法钓鱼及异常滞留等高风险行为。

- **技术方案与贡献：**

专用数据集构建与模型优化：利用公开及自采数据集约 1800 张专用数据集进行迁移学习，系统性地补充困难负样本（如水面倒影、栏杆误识别为鱼竿）与多样化正样本（不同角度、光照下的人竿组合），将模型在新场景下的 mAP 提升至 0.95。

核心算法调优：实现从“静态检测”到“动态行为分析”的跨越，借助 yolov8s 识别与追踪，设计并实现了“事件+行为”双重告警逻辑；使用 TensorRT 对模型进行 INT8 量化，原始模型 mAP 从 96.1% 降至 95.2%。

边缘端应用开发与部署：辅助部署于边缘端(Orin Nano 32GB)的告警应用程序。程序通过多线程管理视频流拉取（RTSP 测试，SDK+GStreamer 部署），实现秒级告警响应。

告警事件接口：与后端工程师协作，定义基于 JSON 格式的 RESTful API 告警接口。将告警事件及 Base64 编码的证据快照封装后，通过 HTTP POST 异步上报至云端，并设计队列缓存机制。

- **成果：**系统综合告警准确率提升至 92% 以上，使项目满足业务上线的严格标准，实现从 0 到 1 的产品化；禁区的现场巡检频率降低了约 80%，同时告警响应时间从原先的数小时（事后发现）缩短至秒级。

➤ 项目二：智慧能源企业级知识与数据 Agent 平台

2024.07 - 2024.11

- **目标：**构建企业级的“认知中枢”智能平台，通过融合大语言模型（LLM），打通非结构化文档（知识）与核心生产数据（数据）的壁垒，实现统一入口下的智能问答与实时分析，解决企业内部信息“孤岛问题”，驱动业务运营的智能化转型。

- **技术方案与贡献：**

Agent 混合查询架构设计：主导设计并实现了基于“工具调用”（Function Calling）的 Agent 架构。将复杂的业务需求解耦为 RAG 知识工具与 SQL 数据工具，由 Agent 作为决策核心，自主规划并执行“数据查询+知识检索”等复合型任务。

Text-to-SQL 能力攻坚：动态检索并注入业务规则上下文（如实体别名），引导模型自主生成含 JOIN、GROUP BY 的复杂查询，实现自然语言到 SQL 的精准转换。

核心模型选型与性能调优：负责核心模型的选型与评估，确定采用 Qwen2-72B 作为推理核心，BGE-M3 作为嵌入模型。通过对向量数据库的 HNSW 索引参数进行调优，在企业级文档(十万级知识块)下实现

了亚秒级的检索响应。

分布式部署:为保障集控中心业务的 7x24 小时稳定运行,使用 LMdeploy 将核心推理服务在 4 张 A100 上进行集群化部署, 实现负载均衡与热备份。

- 成果:** 将跨系统、跨领域的复杂信息查询的平均耗时从 30 分钟以上缩短至 10 秒以内, 显著提升了一线运维及管理人员的工作效率; 打通了生产数据与办公应用的壁垒, 使非技术人员也能通过自然语言进行数据分析, 为精细化运营和数据驱动决策提供了强大的工具支持。

某考研机构 | 兼职

2025.02 – 2025.06

➤ 项目三: 考研政治多智能体协作应答系统

2025.03 - 2025.05

- 目标:** 设计并实现一个基于 langgraph 框架的自主协作多智能体系统, 使其能够通过动态任务规划与 Agent 间的协同, 深度融合结构化知识与实时信息, 为复杂的用户问题提供专家级解答。

- 技术方案与贡献:**

知识资产构建与自动化处理: 主导设计并实现了自动化知识处理流水线, 利用 DeepSeek-v2-lite 与 Qwen3-30B 模型针对非结构化教材中抽取向量化元数据与知识图谱三元组, 构建支撑深度 RAG 的 ChromaDB 与 Neo4j 双模知识库。

核心 Agent 开发: 构建事件驱动的智能体协作图谱, 开发了学科知识、网络搜索、知识库检索等多个职责单一的专家智能体节点。

Agent 规划与动态路由: 采用轻量级的 DeepSeek-v2-lite 模型构建了系统的“规划器”, 通过 Prompt 工程与 Pydantic 模型约束, 使其能精准解析用户意图, 并动态调用不同工具 (知识引擎、网络搜索、真题库)。

响应合成: 选用旗舰级的 Qwen3-30B 模型作为“合成器”, 负责对各并行工具返回的异构信息进行交叉验证、提炼与汇总, 生成逻辑严谨、内容丰富的最终答案。

- 成果:** 多 Agent 架构实现了对复杂问题的动态任务分解与多工具协同调用; 通过动态融合静态知识库 (Graph RAG) 与实时网络搜索, 攻克传统 RAG 无法处理时事性问题的核心局限。

➤ 项目四: 基于人工智能的城市内涝快速预报与风险预警研究 | 重庆市技术创新与应用发展川渝科技创新合作计划

2024.05 - 2024.06

- 目标:** 利用图神经网络 (GNN) 技术, 为高耗时的水动力物理模型构建一个高精度、高速度的代理模型, 并将其与 CCMO 优化算法深度集成, 实现一个能够在短时间完成全局寻优的智能决策框架。

- 技术方案与贡献:**

GNN 模型开发与调优: 基于 PyTorch Geometric (PyG), 自主设计并实现了一个包含多层图卷积网络 (GCN) 的代理模型。通过对网络深度、学习率等超参数进行系统性调优, 并引入 Dropout 层防止过拟合, 最终使模型在独立测试集上的 R^2 (决定系数) 从基线的 0.75 提升至 0.92, 证明了其对物理过程的高度拟合能力。

性能量化与验证: 通过基准测试验证, 集成 GNN 代理模型后, 单次种群(100)评估的时间由原先的 1 分钟级锐减至 0.7 秒, 使优化器在同等时间内可探索的解决方案数量提升近两个数量级 (>100 倍)。

决策结果接口与可视化: 与团队成员协作, 设计了用于输出帕累托最优解集的 JSON 接口。利用

Matplotlib 和 PlatEmo 平台对最优 LID 布局方案进行可视化渲染。

成果: 成功构建了一套从 0 到 1 的 AI 智能决策系统, 实现了对小规模城市内涝防治方案的全局、快速寻优, 其费效比全面优于任何通过传统人工试错法得到的方案, 让决策周期从“周”级降至“小时”级。

➤ **项目五: 共享打磨车间无人值守智能管理系统 | 自接项目** **2023.05 - 2023.06**

- 目标:** 在资源受限的 NVIDIA Jetson Nano 边缘设备上, 部署高精度、鲁棒的 YOLOv8 车辆检测系统, 驱动无人值守计费逻辑 (核心解决“违规多车识别”问题)

- 技术方案与贡献:**

专用数据集构建与模型优化: 基于 YOLOv8s 模型, 利用公开及自采数据集进行迁移学习, 使其在复杂车间场景 (如光照变化、车辆部分遮挡、多角度拍摄) 下, 车辆检测的准确率和召回率均达到商业条件。

边缘设备部署: 辅助完成 CV 模型在 NVIDIA Jetson Nano 边缘计算平台的部署。利用 NVIDIA TensorRT 对模型从 float16 转 INT8 量化和引擎加速优化, 原始模型 mAP 从 95.8% 降至 95.1%, 将推理速度提升了近 3 倍。

数据传输: 使用 Python 和 OpenCV 开发部署于 Jetson Nano 上的核心应用程序。通过 OpenCV 直接获取 RTSP 视频流, 30 秒为周期, 从每个摄像头拉取一帧静态图像用于后续处理。

API 封装: 设计并实现了边缘端与云端后台的通信机制。程序将检测结果 (如车间 ID、车辆数、时间戳) 封装成 JSON 格式, 通过 RESTful API (HTTP POST) 异步 (子线程阻塞方式) 上报给后端服务器。

- 成果:** 实现全自动化计费: 成功实现了基于视觉计数的动态计费系统, 有效支撑了后端“一车一价”和“违规多车惩罚性计费”的核心业务模式, 预计每年可减少约 95% 的费用流失, 将人工巡查和现场管理的成本降低了近 100%。

证书与技能

语言: 英语六级 (CET-6) | **其他:** C1 驾照