## 教育经历

**2018 年 09 月-2022 年 06 月　北京邮电大学| 计算机科学与技术 | 本科　绩点 3.82 排名:7/229 前 3% cet4-570 cet6-524**

**2022 年 09 月-2025 年 06 月　中国人民大学| 人工智能 | 学术型硕士　绩点 3.91**

## 证书荣誉

2020-2021 icpc 区域赛济南赛区银奖(全国二等奖)　2021-2022 ccpc 总决赛银奖(全国二等奖)

2020-2021 icpc ecfinal 总决赛银奖(全国二等奖)　2019-2020 国家奖学金

## 实习经历

**2025.05-至今　宽德投资　职位：机器学习工程师**

1. 负责优化大模型单机/多机推理与训练速度
2. 基于 vllm 的 disaggregate decoding 方法编写基于 NCCL 的多对多 PD 分离方法
3. 将 Cuda Graph 与 torch.compile 方法融入 DDP 训练中，提升多机训练时的训练速度
4. 基于 nginx 编写大语言模型的注册/分发网络，支持心跳检测与负载均衡，并支持推理集群动态扩容与缩容
5. 基于 Nvidia DCGMI 编写集群状态监测机制，可以使用 Grafana 进行查看
6. 基于 triton 算子融合优化大模型推理，推理速度速度提升 150%

**2024.09-2025.02　微软亚洲研究院 （MSRA）　职位：算法工程师**

1. 使用大模型不断纠错作为代码生成的新方法，构建新的强化学习训练方法
2. 使用 mcts 进行多轮 cot 数据收集并使用 llama_factory 对模型进行 offline 优化
3. 使用修改后的 verl 框架对模型进行 online 优化
4. 为了提升推理速度，对 vllm 调优 ，包括参数与模式选择还有异常处理
5. 为了提升评测速度，私用 python 自带 api 和 ray 框架构建代码测试沙箱 ，实现并行测试 、资源与权限与输入输出控制
6. 撰写学术论文，并以共一作者投稿至 acl2025

**2024.04-2024.09　北京通用人工智能研究院　职位：算法工程师**

1. 使用两种方法提升机器人模型推理速度：基于模型的修改和基于剪枝与缓存的提升速度方法

1.1 搭建多模态模型：使用 mamba 模型作为主干构建多模态大模型训练

1.2. 构建支持 deepspeed/FSDP 的多卡训练框架和基于 ray 的多卡并行推理模块

1.3. 通过参数/模型修改在 pope/gqa/vqav2 等多个模型达成 2.8b 模型的 sota

1.4. 使用模拟器数据对 vlm 模型进行训练 ，训练 15 分钟即可在抓取等任务上超过同等模型。

1.5 推理速度达到相同量级 transformer 模型的 8.7 倍

2.1. 搭建模型，使用 qwen 2-vl 模型对输入机器人的图片/文本进行编码，合并多帧信息输入到基于 transformer 的扩散模型进行去噪输出动作序列

2.2. 构建支持 deepspeed/FSDP 的多卡训练框架和构建基于 flask 和多进程的基于 rlbench 的测试框架

2.3. 在合并多帧信息模块中使用 cache 对之前输出进行存储 ，加快推理速度

2.4. 使用基于层的剪枝，将后训练后的结果推理过程中与上一层的隐藏层信息过于相似的层转换为恒等变换并再次训练

2.5. 在保证任务精度（完成成功率损失不超过 5%）在速度上超过了相同大小的模型，相比原模型提升了 11.7 倍的速度

**2022.03-2022.09　微软算法开发部门（STCA）　职位：算法工程师**

1. 使用基于 GNN 的图模型进行分层图理解，并对时序异常节点进行多节点归因分析
2. 清洗数据，设计测试用的数据集（包括出错次数多的和节点数目多的），完善测试 baseline
3. 通过 GNN 去恢复被添加了 mask 的原图像判断去理解图节点之间的关系
4. 修改搜索逻辑，使用基于蒙特卡洛树搜索和遗传算法的启发式搜索方式，可以避免节点数过多带来的运行效率下降
5. 相比 baseline 在真实数据集上 F1 值平均高出 10%，在复杂人造数据集上高出 20%，已经上线 azure 云平台

# 项目经历

**2024.10-2025.4** **货物期货价格预测**

1. 负责编写代码进行大白菜/猪肉等期货价格预测。
2. 使用 xgboost/线性模型/GNN 时序模型对期货价格进行预测。
3. 整体误差相比 R 语言基线模型减少 50%。
4. 通过 python/c++编写订单簿使用过往用户信息和期货价格预测结果进行成交预测和订单价格预测。
5. 根据上交所规则，分为启动/集合竞价/连续竞价/结束不同时段对大盘价格预测。
6. 优化整体复杂度，可以通过 O(N)时间计算价格（N 为当天订单簿长度）。

# 论文情况

一作/共一文章

***(ICLR 2024)

***(**nature machine intelligence)**

***(Nature Computational Science)

***(AAAI2026)

合作文章

***(ACMMM2024 Oral)

***(WSC 2022)