主页：https://cloud-peterjohn.github.io
GitHub：https://github.com/cloud-peterjohn
简历：https://cloud-peterjohn.github.io/resume-no-login.html
项目：https://cloud-peterjohn.github.io/projects.html

周 炳烨

教育背景

- 2022年9月~2026年6月：四川大学计算机学院（SCU），本科生
- 2025年2月~2025年6月：国立阳明交通大学资讯学院（NYCU），交换学生

学术表现

- 学习成绩
  - 平均学分绩点3.90，必修平均绩点3.89，成绩均分92.11，必修成绩均分91.74
  - 必修成绩排名6/327（前2%）
  - 四级565分，六级533分
  - 入选四川大学吴玉章学院树人班（本科生荣誉学士学位）
- 竞赛经历
  - 全国大学生数学建模竞赛全国二等奖
  - 全国大学生数学竞赛省级三等奖
  - 第十四届全国大学生电子商务创新、创意及创业挑战赛省级二等奖
- 获奖情况
  - 2023-2024学年度本专科生国家奖学金
  - 2023年度长江电力奖学金

科研经历

- 大创
  - 国家级大创《基于子空间学习的粒网络威胁诊断关键技术》
  - 大创简介
    当前的离群点检测难以应对数据中的低维不可分性、随机性与不确定性。因此，我们首先利用核方法进行维度转换，克服了离群点在一般方法下线性不可分的问题；再采用自表示学习方法，根据样本在高维空间的重构误差构建离群分数。同时，我们通过计算表示矩阵的余弦相似矩阵来得到模糊邻域离群分数，利用其刻画模糊相似关系来应对数据中的不确定性。最终，依据两个分数进行离群点的判断。
  - 个人工作

大创成员，负责实验的运行、基线方法的比较。

- 学术会议
  - 受学校资助，参加CNCC中国计算机大会（2024，横店）
  - 参与中国多智能体大会（2024，成都）的筹办
- 论文
  - 《Evolution Meets Diffusion: Efficient Neural Architecture Generation》
  - 论文简介

    我们将耗时的神经架构搜索问题，转化为神经架构生成问题，使用生成模型直接生成在下游任务中最优的神经网络架构。我们对扩散模型进行了改进，使其在迭代去噪的生成过程中无需网络（如UNet）来预测待消除的噪声，而是使用进化算法来引导模型的去噪，使得适应度值更高的前代个体x_t对后代样本x_{t-1}占据更主导的作用，从而逐步生成最终的最优样本架构x_0。
  - 实验结果

    在NAS-Bench-101, NAS-Bench-201, NAS-Bench-301, TransNASBench-101-Micro/Macro, MobileNetV3, DARTS, AutoFormer多个不同规模、不同任务的搜索空间中实现了SOTA精度，无需神经网络、也无需训练，并且生成（推理）速度提高了高达50倍。
  - 个人工作

    第一作者，主要负责方法框架的设计、神经预测器的设计、实验的设计、代码的编写、实验的运行、论文的撰写（除摘要外）。

科研兴趣

- 大模型的高效微调算法

  虽然LoRA，AdaLoRA，DoRA已经有较广泛的应用，能够大幅减低微调时的显存占用，但是微调仍然是基于服务器端进行，即使QLoRA进一步通过先量化再微调允许在个人PC上微调大模型，但是对于真正的边缘端（如手机、树莓派）还是难以实现，因此我对“如何进一步从微调算法层面，降低大模型微调的显存需求，提高微调算法的效率”比较感兴趣。
- LLM Agent

  Agent通常将LLM作为一个通用代理解决各类下游任务，通常使用Zero/Few-Shot CoT、ToT/GoT、Self-Consistency等Prompt工程方法。它是一个低成本但效果好的应用范式，因此我对相关开发很感兴趣。

  无论是上下文学习还是思维链技术，我认为都可以视作一个最优化的任务：如何优化Prompts以使得模型的生成内容更符合用户需求。因此，我认为思维链可以进一步结合进化算法、群体智能进行优化。
- 大模型的量化与推理加速

  模型剪枝（如基于泰勒一阶方法、基于Hessian二阶方法），和模型量化（如AutoGPTQ，SpinQuant，AutoAWQ）已经较为成熟且有广泛的落地应用。但是对于KV-Cache的加速（如KV-Cache的量化、压缩、低秩分解），不同的方法从不同的角度实现了加速，但是尚且缺乏框架/代码库将其统一起来，同时使用。
- 数据集蒸馏的泛化性

  数据集蒸馏主要包含基于梯度匹配、基于轨迹匹配、基于特征/激活匹配、基于分布匹配等方式，但是蒸馏后的数据集通常过拟合于某个架构，缺乏泛化性。并且数据集蒸馏计算成本很高，缺乏零成本评估方法。

实践经历

- 四川大学计算机学院主持人大赛二等奖，四川大学吴玉章学院2024年音乐会演出，四川大学校园路跑72/3000
- 四川大学学子家乡行社会实践活动优秀个人，第31届世界大学生夏季运动会城市志愿者
- 四川大学优秀学生，四川大学优秀学生干部，四川大学优秀共青团员，四川大学优秀共青团干部，四川大学优秀志愿者
- 四川大学校青年志愿者协会人力资源部副部长（2023年6月~2024年6月）、四川大学计算机学院行政一班学习委员（2024年6月~2025年6月）

项目经历

- NOAA Fisheries Steller Sea Lion Population Count
  - 项目简介
    我们使用YOLO模型计数图片中各类型海狮数量。本项目的困难点在于训练数据集没有Bounding Box / Mask的标注，每只海狮只有人为标记的一个颜色点；且图像中海狮分布差异巨大，海狮过于集中在小部分区域，海狮的重叠很严重。我们针对上述问题，提出了自适应的Bounding Box标注方法，并综合使用切片、多尺度训练等技巧。
  - 项目结果
    实现了Kaggle竞赛的第二名，计数的RMSE低于0.12。
  - 个人工作
    上述全部工作，其他队友尝试Faster-RCNN/UNet等检测、分割方法，并分析了模型架构对最终效果的影响（影响不显著）。
- Acceleration and Deployment of LLMs
  - 项目简介
    对Llama-3.2-3B-Instruct进行推理加速，并部署在边缘设备上。我们使用AutoGPTQ进行W4A16量化，并在推理时采用KV-Cache的8-Bit量化、使用Llama-3.2-1B-Instruct进行推测解码和Flash-Attention的推理加速。最后，我们借助MLC-LLM库将模型部署在边缘设备。
  - 项目结果
    在PPL不超过11.5的情况下，在NVIDIA T4 Tensor Core GPU上实现了80+的吞吐率，在NVIDIA GeForce RTX 3060上实现了25+的吞吐率，在AMD Ryzen 7840H上实现了11+的吞吐率。
  - 个人工作
    上述全部工作，队友使用AutoAWQ得到了类似的效果，另一队友使用Half-Quadratic Quantization + torch.compile()，但由于HQQ库有BUG，无法正确地保存和加载。
- SCU-ACM-OJ
  - 项目简介
    面向校级ACM竞赛而设计的Online Judge，为用户和管理员提供了自动压缩包解析等高效的操作，实现了比赛中实时榜单的功能，并设计了基于LLM的威胁代码检测+虚拟环境隔离的双重安全保障。
  - 个人工作
    用户端的前端界面开发，包含用户账号+密码/邮箱+验证码登录、查看所有比赛列表、查看比赛详情、浏览题目和代码提交、测评记录查看、个人信息修改等界面及其前后端交互，以及网页路由鉴权和操作鉴权。