张钧洋

研究方向：资源受限场景下大模型推理优化 | 导师：李向阳教授
政治面貌：中共党员 | 邮箱：zhangjunyang@mail.ustc.edu.cn
籍贯：安徽池州 | 出生年月：1998.1.31 | 个人主页：jzhang.tech

教育经历

中国科学技术大学，博士，计算机科学与技术学院，计算机科学与技术专业 2022年9月-2026年6月
一作论文：第一作者发表CCF A类论文3篇、B类1篇，其中包括SCI一区1篇、共同一作1篇。

中国科学技术大学，硕士，计算机科学与技术学院，计算机科学与技术专业 2020年9月-2022年6月
主修课程：算法设计与分析(96)、高级算法设计与分析(92)、机器学习与知识发现(92)等。**GPA: 4.09/4.3。**

重庆大学 (985)，本科，计算机学院/弘深学院电子类创新班，计算机科学与技术专业 2016年9月-2020年6月
主修课程：高等数学(99)，数据结构(95)，计算机网络(93) 等。**GPA: 3.90/4.0。综合绩点年级第一。**

科研成果

第一作者发表CCF A类论文3篇、B类1篇，其中包括SCI一区1篇、共同一作1篇。以下是部分代表性论文成果：

1. **A-VL: Adaptive Attention for Large Vision-Language Models** 发表于CCF-A类会议 AAAI。
2. **TensAllo: Adaptive Deployment of LLMs on Resource-Constrained Heterogeneous Edge Devices** 发表于CCF-A类会议 IEEE INFOCOM。
3. **Deploy Efficient Large Language Model Distributed Inference Pipeline for Heterogeneous GPUs** 发表于CCF-B类会议 IEEE/ACM IWQoS。
4. **WordWhisper: Exploiting Real-Time, Hardware-Dependent IoT Communication Against Eavesdropping** 发表于CCF-A类、SCI一区期刊 IEEE Transactions on Mobile Computing。

项目经历

蔚来汽车校企合作项目“基于大模型的智能座舱系统优化” **2024年1月 至 2024年12月**
- 背景：多模态大模型可增强智能座舱交互体验，但端侧计算资源受限，亟需降低推理成本。
- 行动：独立负责端侧推理优化子课题。分析模型内部计算模式，精准识别冗余，设计自适应计算模块。
- 产出：显著降低计算负载，减少49%推理时延，降低43% KV Cache显存占用。在蔚来智能车上真实部署。

百度校企合作项目“面向ESG专业领域问答系统” **2023年3月 至 2023年11月**
- 背景：ESG领域知识庞杂异构，亟需基于RAG架构的专业问答系统提升信息获取和决策的效率。
- 行动：参与研发知识库与行业大模型问答系统。重点优化RAG架构下搜索召回、文档解析等模块。
- 产出：构建超300G文档的行业知识库，交付并上线ESG智能问答系统，现已在百度ESG主页提供服务。

华为校企合作项目“端侧AI能效比提升” **2022年1月 至 2023年2月**
- 背景：当前算力需求急剧扩展，利用异构设备推理成为新需求，但大模型异构推理易发生不平衡和阻塞。
- 行动：独立负责异构设备推理优化子课题。深入分析大模型流水线理论建模，创新提出自动部署调度算法。
- 产出：大幅提高提高资源利用率，提升异构算力下的大语言模型推理吞吐达37.1%，峰值提升83.0%。

华米科技校企合作项目“基于可穿戴设备的用户行为识别” **2020年3月 至 2020年12月**
- 背景：可穿戴手环手表中需要高精度地检测用户当前运动状态，如走路、跑步、骑车、打球等行为。
- 行动：使用小型LSTM+CNN同时提取陀螺仪、加速度计的时空特征，采集运动数据集并设计模型。
- 产出：提高了用户运动行为的识别精度，尤其是有手部动作的行为因为联合多传感器空间特征更为精准。

实践经历

专业实践：负责中科大校园USTC DeepSeek的集群中模型推理服务的部署和调度，在多机多卡PD分离相关部署实践中有经验，动手能力较强，并且擅长解决调度类问题。

组织工作：担任中科大研究生智能物联党支部组织委员，推进超过20名同学入党流程。担任中科大研究生课程“计算机应用数学”的助教。担任重庆大学微软学生俱乐部社团社长。多次参与各类会议活动志愿者等。