魏先生

男 | 15810312588 | chaowei2003@hotmail.com

14年工作经验 | 求职意向：算法工程师/系统架构师/技术总监

个人优势

1. 精通各类开源大模型的结构优化与微调技术，成功应用于 AR、数字员工等创新场景。
2. 深耕深度学习与计算机视觉领域，特别是多模态视频理解方向，具有丰富的研究经验。
3. 以第一作者身份在 ICRA、ICASSP 等国际顶级会议发表论文6篇，合作发表论文1篇。
4. 拥有10年 AI 领域科研经验、8年系统架构经验，并具备5年以上的团队管理经历，领导过70人以上团队。
5. 深度参与腾讯亿级用户系统的架构设计及运行维护，积累了大规模系统的开发与管理经验。
6. 熟悉电商、金融等领域的业务逻辑与系统架构，能够迅速理解和适应多领域业务需求。
7. 兼具 AI 模型构建与系统研发的综合能力，具备广泛的技术知识和强大的实践能力。

专业技能

1. 扎实的计算机理论基础，对机器学习，深度学习有深入理解；
2. 在深度学习与计算机视觉，大模型（LLM，VLLM）设计与应用，知识图谱等方面有较为深入的研究；
3. 擅长高并发微服务架构设计，对软件设计模式有较深入理解；
4. 熟练掌握C/C++/python/java及与之对应的生态框架；
5. 熟练掌握Pytorch，Tensorflow，PYG，DGL及CNN，Transformer，GNN等模型设计；
6. 熟悉Nginx, Dubbo, spring， ActiveMq等组件的部分源代码；
7. 熟悉Mysql、Nginx、Tomcat、Zookeeper、Redis等常用组基本原理、性能优化及维护；

教育经历

清华大学 博士 计算机科学与技术 2018-2024

中南大学 硕士 计算机科学与技术 2008-2011

湘潭大学（双一流） 本科 计算机科学与技术 2003-2007

工作经历

腾讯科技 高级后台开发 2011.05-2015.04

小牛资本 技术总监 2015.04-2018.08

图灵通诺 算法研究员(兼职) 2021.03-2023.03

浙江大学南昌研究院 大模型算法 2024.06-至今

项目经历

XR智能眼镜多模态场景理解　算法研究员　2024.06-至今

内容：

负责XR智能眼镜产品线AI算法的设计与实现，对标苹果Vision Pro，构建三层式大模型应用架构（硬件与驱动层、服务器模型层、端侧模型层），推动产品智能化进程：

1. 语音助手（增强版Siri）

- 设计端到端语音模型（speech-to-speech），采用Encoder-only语音编码器和Decoder-only语音解码器，并通过非自回归方式生成语音输出。
- 实现200毫秒以内的超低响应时间，速度较传统ASR + LLM + TTS流程快4倍，较端到端语音模型SpeechGPT快2倍，性能超越当前SOTA模型Moshi。

2. 场景翻译

- 语音翻译：基于Meta开源模型Seamless，采用LoRA微调优化汉英互译，首字延迟低至100毫秒以下。
- 视觉翻译：利用多模态大模型（VLLM）进行场景微调，支持XR视野内多语言内容实时翻译。
- 文本翻译：采用Google开源模型T5，确保高效、精准的文本转换。

3. 物体识别与交互

- 使用多模态大模型（VLLM）实现视觉问答（VQA）功能，支持用户在XR眼镜视野中对任意物体进行智能提问与实时解答。

业绩：

1. 实时性：响应时间最快可控制在200毫秒以内，这一速度至少比传统的ASR + LLM + TTS组合快四倍，相比其他端到端语音模型SpeechGPT快两倍，并且超过了当前的SOTA模型Moshi。

2. 准确性：基于LossPred的无监督算法，通过分析预训练自监督模型（例如HuBERT）在不同掩蔽跨度上的损失，揭示了未注释语音信号的嘈杂音节状分割。我们的模型在无监督音节分割、分类和低比特率单位到音频重合成方面达到了最先进的水平。

3. 轻量级：服务器端模型的规模为3B。通过量化和压缩等技术，模型体积可压缩超过16倍，使其能够在手机等终端设备上运行。与Moshi相比，该模型在规模上小了2.5倍，但性能相当。

4. 提升用户体验：我们支持打断和噪声识别。与其他端到端模型不同，我们部署了两组模型互为主备机制，其中一组处理用户任务，另一组实时监听指令和噪声识别。这种设计有效实现了打断和过滤噪声的功能，从而显著提升了用户体验。

5. 功能扩展：针对我们的应用场景，我们集成了AI智能体，以实现用户意图的识别功能，并通过多种工具调用扩展了大模型所不具备的能力，例如实时天气查询。

无人店大模型自动训练平台　算法研究员　2021.04-2023.04

内容：

负责智能货柜与无人店产品线模型自动化训练与发布体系的构建与实现，涵盖数据采样、模型训练与测评、模型打包与发布等全流程：

1. 数据仓库设计与实现

- 为满足大型模型对数据规模和质量的要求，构建集中化数据管理与处理系统。
- 实现数十亿多类型标注数据的结构化存储，并对回流数据进行实时清洗。
- 借助深度模型评估数据质量，为500TB训练样本建立标签索引，显著提高数据筛选效率，保障模型训练的稳定性和鲁棒性。

2. 大型模型优化

● 将基础模型从传统ResNet架构升级至Transformer架构，显著提升模型性能。

● 在无人店与智能货柜场景中，基于预训练ViT大型模型进行微调（fine-tuning），商品单次识别错误率较ResNet101降低11.6%。

3. 自动化训练与发布平台

● 开发自动化平台，根据模型检测与识别效果及数据仓库状态，自动触发训练与测评流程。

● 根据测评结果实施灰度发布并进行实时监控，最终全量推送最优模型，提升模型迭代效率与质量。

4. 基于对比学习优化ViT模型

● 设计新的对比损失函数与正负样本生成策略，提升基础模型的泛化能力。

● 简化检测、检索、识别等下游任务的训练难度，为多任务模型提供更强的通用性支持。

业绩:

1. 构建并实现模型训练的全自动化与标准化流程，大幅节省人力成本，释放两名模型训练工程师的工作量，同时显著加快模型更新与迭代速度。

2. 通过深度优化模型，成功将识别错误率从平均3%降低至1%以下，显著提升系统识别精度。

3. 推动数据仓库系统上线，实现训练数据的统一管理与高效利用，大幅提高数据处理效率与模型训练效果。

金融科技服务平台　团队负责人/技术总监　2015.03-2018.08

内容:

针对集团业务的现状，进行需求分析，将系统拆分为五大体系：

● 基础技术体系，包含基于SOA的分布式框架，消息中间件集群，数据库集群，持续集成环境等；

● 金融平台体系，包含统一支付结算中心，统一交易中心，统一账户中心，区块链平台等；

● 金融业务体系，包含小牛在线、小牛消费分期、小牛微贷、公募基金代销等20个业务产品线。

● 大数据体系，包含数据仓库，大数据风控，用户画像，精准营销平台等产品

企业IT体系，办公系统与办公网络的建设。

业绩:

● 降本增效：降低大量基础模块重复开发，同时减少运营的人力成本；

● 安全性与实时性：抽象出统一支付、交易、账户等核心系统，实现交易更安全，清结算更实时；

● 用户体验：登录与账户体系打通，交易更实时，用户体验更好；

● 集团内部表彰：年会部分获“优秀团队”，个人表彰为“十佳经理人”；

统一支付结算平台　架构师/核心模块开发　2012.11-2015.03

内容:

针对腾讯电商(ECC)业务发展，已有拍拍网，QQ商城，QQ网购，及收购的易迅网等多平台，需要将这些台核心功能打通，为用户提供一致的购物体，基中支付结算是核心之一：

● 统一支付平台的需求分析，架构设计与核心模块开发，相关技术难点问题解决。

● 结算系统，费率模块的设计开发

● 统一支付平台，费率模块运营监控系系统的设计与开发。

业绩:

针对腾讯电商(ECC)业务发展，已有拍拍网，QQ商城，QQ网购，及收购的易迅网等多平台，需要将这些台核心功能打通，为用户提供一致的购物体，基中支付结算是核心之一：

● 统一支付平台的需求分析，架构设计与核心模块开发，相关技术难点问题解决。

● 结算系统，费率模块的设计开发

● 统一支付平台，费率模块运营监控系系统的设计与开发。

● 技术架构更清晰，并发能力更强，大大降低开发上线时的风险，采用四级分布式架构：存储层（DB-Myqsql、分布式Cache-Cmem+TTC）、数据访问层(dal_set)、逻辑业务层(AppframeWork)、web访问层（Apache+Nginx+TGW）

QQ网购订单系统2.0　主要开发　2012.06-2012.12

内容:

QQ网购二期，将QQ商城商家迁移至QQ网购，相对于B商户（如一号店）商城卖家IT能力较弱，订单2.0项目需为商家提供一套完善的交易前端系统，同时重构个人中心及客服系统.主要完成的功能有:订单详情，订单列表，标记发货，退款，确认收到等功能:

● 与产品经理完成需求分需分，明确需实现的功能，完成概要设计。

● 技术难点分析、解决及接口规范定制

● 完成商户统系订单详情，订单列表功能

业绩:

● 支持C2C，B2C，B2B2C，货到付款等多种交易模式

● 对底层架构的重构，读出速度提升100倍，写入速度增加10倍，并发量提升100倍

● 支持更灵活多变的促销方式

科研项目

多模态开放集识别　2020.01-2024.03

● 项目描述：针对智能驾驶中的实际应用场景，综合 利用可见光（RGB）视频、骨架序列、测距（深度）视频、红外视频、文本等多种 模态的行为数据，从特征提取、跨模态协同学习与模态特征融合等三个角度，对多模态行为识别方法进行研究。针对开放集零样本跨模态行为识别，提出了一种基于场景图结构表征学习和预训练大模型特征对齐的方案。针对场景变化频繁且对鲁棒性与精度要求高的场景，提出了一种视频可变形注意力 多模模型征融合的方法。最后，将以上方法应用到智能驾驶场景中安全驾驶员的行为识别任务中。

● 项目业绩：1.在骨架单模态场景全监督场景在三个Benchmark上取得最最先进的结果；在开放集跨模态零样本文本提示识别场景，在NTU-60、NTU-120与Kinetics-400数据集上分别超出最陷阱的方法，27.%，19.7%与2.1%；在模态模型融合识别场景，在四个Benchmark上超过最先进的方法5%以上；已将部分成果发表示于ICRA、ICASSP等多个国际顶级会议上；

小样本/零样本动作识别　2020.01-2024.03

● 项目描述：主要研究半监督的支动识别方法，解决数据集不足或难以收集的场景。

● 研究内容：2.提出扩散模型的GCN实现方式 2.提出一种基于Transformer的全局Pooling方法；3.数据增广方法; 4.改进Meta-classifier;

● 项目业绩：1.在NTU-60、NTU-120与NW-UCLA三个数据上取得SOTA的结果: 1.在全监督场景下，分别超过当时DOTA方法为1.2%与1.6%与0.8%的性能; 2.在1-shot场景下，都过了SOTA方法5.0%以上的性能；已将部分成果发表示于ICIP、ICASSP等多个会议;

小样本图分类　2020.01-2024.03

● 项目描述：现实场景中存在着大量Graph结构的数据(社交网络，交通网路，分子结构等)，而在研究Graph的某些领域中，如药物筛选，蛋白质的性质等，可用样本数相对较少，因此需要借助小样本元学习方法。

● 项目业绩：1.在四个数据集上取得SOTA的结果，特别是1-shot/5-shot样本极少的情况下，效果比SOTA效果高出10%；已将部分成果发表示于IJCNN、ICIP等多个会议;

小样本语义分割 2019.04-2019.12

● 项目描述：语义分割在自动驾驶、机器人领域，医学图象处理是十分关键的技术，然而在标注图片像数的代价非常大， 有些研究场景样本极其有限，如医学领域，因此研究小样本语义分割具有重要的意义。

● 研究内容：1，跨域学习(domain adaptation). 3，探索使用该场景的Meta-learning方法

● 项目业绩：1，Pascal VOC 2012 and FSS-1000 超出SOTA结果2%以上；

资格证书

大学英语四级　大学英语六级　雅思6.5分　计算机三级　系统架构设计师

荣誉奖项

● 2019年9月，全国机器人大赛两项第一名；

● 2017年1月，集团年会年度表彰为“十佳经理人”；

● 2011年4月，被中南大学评为“优秀毕业生”；

已发表论文

● Chao Wei and Zhidong Deng. ” Incorporating Scene Graphs into Pre-trained Vision-Language Models for Multimodal Open-vocabulary Action Recognition”. 2024 IEEE International Conference on Robotics and Automation (ICRA 2024), Yokohama, Japan, 13-17 May 2024.（TH-CPL A，CCF B）

● Chao Wei and Zhidong Deng. ” Open-Vocabulary Skeleton Action Recognition with Diffusion Graph Convolutional Network and Pre-Trained Vision-Language Models”. 2024 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2024), Seoul, Korea, 14-19 April 2024. (TH-CPL B，CCF B)

● Chao Wei and Zhidong Deng. ” A Novel Contrastive Diffusion Graph Convolutional Network for Few-shot Skeleton-based Action Recognition”. 2024 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2024), Seoul, Korea, 14-19 April 2024. (TH-CPL B，CCF B)

● Chao Wei and Zhidong Deng. ” Accommodating Self-attentional Heterophily Topology into High- and Low-pass Graph Convolutional Network for Skeletonbased Action Recognition”. 2023 International Joint Conference on Neural Networks (IJCNN 2023), Queensland, Australia, 18-23 June 2023.（TH-CPL B，CCF C）

● Chao Wei and Zhidong Deng. ” Few-shot Graph Classification with Contrastive Loss and Meta-classifier”. 2022 International Joint Conference on Neural Networks (IJCNN 2022), Padova, Italy, 18-23 July 2022.（TH-CPL B，CCF C）

● *Hongchao Lu, *Chao Wei, and Zhidong Deng. ” Learning with Memory for Fewshot Semantic Segmentation”. 2021 IEEE International Conference on Image Processing (ICIP 2021), Anchorage, Alaska, USA., 19-22 September 2021.（TH-CPL B，CCF C）