王昊轩

+86 18622468042

billweasley20092@gmail.com

Github: https://github.com/billweasley

领英: https://www.linkedin.com/in/horace-haoxuan-wang

工作经历

2025.03 - 现在 • **高级(senior)数据科学家, 诚信和顾客体验** @ Grab

- 初始项目是计划产生一个手机上用的关键词识别模型，以结合多模态大模型来保护的士司乘安全。

- 2022.03 - 2025.02 **机器学习工程师，语音识别与语言技术** @ Zoom

  - 与其它2名同事合作，设计, 实现和完成了Zoom的离线转录服务的升级和改造，使得该服务由一个单语种服务变成了多语种（36种）支持多业务（5+）的服务。架构基于AsyncMQ/Kafka，使用k8s和istio进行部署, 基于cpu用量进行动态扩缩容策略，服务全球Zoom客户。
  - 维护 ASR 的推理仓库, 并解决线上的转录质量相关的问题。
  - 对ASR模型和服务进行压测，以确定最优的推理参数，内存用量，cpu用量等参数。
  - 在LAS/seq2seq模型上实现基于多头注意力（MHA）的时间对齐，从而提供良好的单词级时间戳，以满足多语言转录的业务需求。
  - 独立完成**Whisper**推理支持的实现，优化和性能评估（WER, RTF/latency/throughput），使用内部in-house VAD（人声检测）模型和开源的WhisperX，相比OpenAI的实现取得了更高的吞吐，并在多数测试集上实现了更低的wer。
  - 使用闭源LLM进行ASR error correction后处理，从Zipformer-Transducer模型中导出N-best list，结合biasing word list撰写prompt发送到Claude进行name entity修正，离线实验在medical数据集上**Rare word WER从37.8%变为17.5%**。
  - 独立开展ASR/LLM的结合进行语音+文本模态supervised finetuning的实验，尝试提高ASR解码结果的一致性。目前在全格式的WER（词错率）和正规化的RWER（稀有词词错率）均取得优于产线模型相当的水平。

- 2019.08 - 2021.12 **软件工程师和数据科学家** @ Barclays

  - 入职时作为Java工程师，在Barclayscard做Java后端开发，支持Springboot的缓存业务。
  - 公司地址匹配和实体匹配，没有内部GPU和可用的标签数据。使用主动学习方法解决。在CPU集群上使用我自己从头开始构建的基于DJL的管道完成了600万内部配对样本的推断。该模型在嘈杂的测试数据集上实现了**94%的F1得分（起初为89%）**。模型在我们的Spark集群上以分布式方式进行离线推断。对于600万配对样本，运行时间不到1小时（在拥有80个CPU的集群上）。

教育背景

2018 - 2019 硕士，网络科学与大数据分析 @ **伦敦大学学院，** 优异
2016 - 2018 学士，互联网计算 @ **利物浦大学*，** 一等荣誉
2014 - 2016 学士，信息与计算科学 @ **西交利物浦大学***

***注:**2+2模式（前两年在中国苏州，后两年在英国利物浦），双学位。

个人项目

- 2024.06 - **关于病历数据在LLM上的微调与评估**

（进行中）使用十几万条中文病例数据，在科室分类，病历总结，出院证明等任务上对不同的LLM foundation模型（Llama3-instruct, Llama3中文-chat，Qwen2）进行全量微调，在域内测试数据集上，微调后的数据集在问诊总结/出院总结等场景下的BLEU/ROUGH，和科室分类的多分类accuracy上均取得了巨大的提升（BLEU 0% ~ 30% -> 49% ~ 55%， ROUGE-L 20% ~ 30% -> 60% ~ 64%, Accauracy 0% ~ 36% -> 69% ~ 71%）。我们计划后续将数据开源。

技术分享

- **"Accelerating Deep Learning on the JVM with Apache Spark and NVIDIA GPUs"**

作者: Haoxuan Wang, Qin Lan [AWS], Carol McDonald [Nvidia]; 链接: https://www.infoq.com/articles/deep-learning-apache-spark-nvidia-gpu/?itm_source=articles_about_ai-ml-data-eng&itm_medium=link&itm_campaign=ai-ml-data-eng

早期项目

- 2019.06 - 2019.09 **项目实习（硕士学位论文）** @ Astroscreen

社交媒体帖子语言来源识别（推文和Gab帖子）项目。 完成了Gab.com的语言（帖子）数据收集爬虫，使用正则表达式进行数据预处理，通过微调BERT和XLNet构建模型来分类这些数据的来源， 使用t-SNE可视化结果，进行了"leave-one-hashtag-out"交叉验证，并使用一些常见指标（准确率、F1分数、混淆矩阵、马修斯相关系数）评估模型。 微调后，XLNet在标签均衡的测试数据集上显示了86%的F1分数，而在随机均衡的测试数据集上为97%。结果显示了使用模型进行来源检查的潜力，也表明了避免数据泄漏的重要性。

- 2025.02 - 2025.02 **Work trail** @ Finalround.ai

单周项目，目标是ASR 中间结果的问题意图检测。独立实现了整个检测流程，在验证会议数据上取得了87%的F1结果，其中一半是早于asr final utterance提前检出的，并取得了offer。整个流程包括：1）使用规则处理问候句 2）使用segment-any-text, 句法分析和llama 1B模型的困惑度判断句子是否完整 3）使用sentence bert检测是否是确认性质的问题（e.g. "can you hear me?"）4）使用3B SLM判断最终提问意图。此外，实现了简历抽取的prompt，实践中可以大幅提高llm答案的个性化。