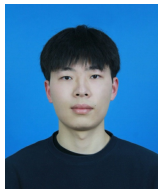
谢东霖

17760489273 | donglinxie11@gmail.com | 重庆
1998-06 | 男 | 中共党员



教育经历

四川大学 2016年09月 - 2020年06月
软件工程 本科 软件学院
2017-2018年四川大学优秀学生
2017-2018年四川大学综合一等奖学金
GPA：3.47/4.00

四川大学 2020年09月 - 2023年06月
电子信息 硕士 计算机学院
2021年四川大学优秀研究生
GPA：3.69/4.00

工作经历

华西医院循证医学中心 2023年07月 - 至今
算法工程师

1. 项目描述：多变量异步时序与结构化数据融合的产后出血预测。
职责描述：
（1）为了解决整个妊娠期的检查指标的时间序列异步问题，以及不同时间变量的不同长度，我们基于Piecewise Linear Encoding构建了一种多元异步时间序列嵌入模块，同时将具体的时间变量编码为位置向量；
（2）为了融合时序数据和结构化数据，我们基于transformer构建了双流的编码器模块，包括时序维度注意力模块，变量维度融合的注意力模块，以及双流并行的前馈网络模块；
（3）和该领域常用的预测算法相比，达到更优的预测效果（F1，AUPRC等指标上）；

2. 项目描述：基于多粒度的药品名标准化
职责描述：
（1）训练阶段：第一阶段使用bge预训练模型作为encoder构建双塔的模型，使用对比学习作为损失进行训练；第二阶段使用bge作为encoder构建单塔模型，将文本对以[cls]text1[sep]text2[sep]的方式进行构建，使用BCE作为损失进行训练；
（2）测试阶段：第一阶段使用bge输出的embedding计算语义相似度，同时使用bm25计算词汇相似度，将两者平均之后作为最终相似度，召回topk相似的药品名；将topk标准名和输入分别构建文本队，最后选择相似度最高的输出。

3. 项目描述：基于知识图谱的用药安全大模型
职责描述：
（1）本地知识库构建：step1. 收集相关医学文献，基于对文献内容解析进行切片，然后基于Qwen2.5-72b-instruct提取“疾病”，“症状”，“药物”实体，以及原始文献对该实体的描述信息；step2. 基于提取的实体，再利用大模型识别关联强的实体对，以及原始文献对该关系的描述信息；step3. 对实体名称有差异的同一实体进行合并，然后并整合来自不同文献对该实体的描述信息，最终形成具备文本属性的知识图谱；
（2）外部知识库构建：基于google的可编程搜索引擎服务构建定制化搜索api，纳入可信度高的在线网站作为外部检索链接（例如维基百科、who、fda、medlineplus等）；
（3）本地知识检索：step1. 利用大模型从输入问题中提取药物实体，然后从本地知识库中检索top k相关实体作为检索起点；step2. 基于beam search的策略从知识图谱中搜索知识路径，每次搜索包括了路径剪枝和路径判断两步：首先从当前节点出发的所有边中选择top k条路径，然后判断当前路径是否具备足够信息回答问题；
（4）外部知识检索：step1. 基于药物实体检索相关网页；step2. 解析网页中的文本信息，然后利用大模型提取去问题相关的信息作为在线证据；
（5）问题推理：step1. 采用plan-and-solve的策略，让Qwen2.5-72b-instruct首先理解问题并结合检索的知识指定相关的解题步骤，最后按照步骤进行推理并给出结论，得到初步回答。step2. 将QwQ-32b-preview作为评估器，对当前回答从多维度进行评估。step3. 基于评估结果，利用Qwen2.5-72b-instruct对于初步回答进行修正。

实习经历

快手 2021年11月 - 2022年09月
内容理解算法实习生
（1）负责视频商品识别优化工作：对视频帧数据和文本进行多模态的建模，进行多标签的识别，首先对比不同视觉backbone（cnn、vit、swin-transformer等）选择更有的视觉backbone来提取视觉特征，然后基于patch的细粒度多模态融合替换原始的全局融合，以及使用非对称加权损失（ASL loss）对难样本和正样本进行动态加权，最终recall提升3%;
（2）负责电商域 benchmark 的打榜： fashion-gen(多模态图文 商品分类)，采用和（1）中类似模型框架；muge(图文跨模态检索和图像字幕)任务，在跨模态检索部分使用粗粒度检索+细粒度召回的框架，在图像字幕任

务上使用gpt2+vit的框架进行跨模态文本生成，同时在600w的电商图文上进行预训练，在2-4月榜单第一。

（3）负责广告视频的场景分割和识别任务：首先训练一个连续帧判别模块，用于修正人工标签；然后在大量短视频帧上基于分类任务训练视觉特征提取器；最终采用两阶段框架，首先抽取视频帧的特征，然后利用多模态融合模块融合帧和文本（title、ocr、asr）之间的语义信息，对于融合后的视频多模态语义特征进行场景分割以及场景标签识别任务（部门自研项目）。

竞赛经历

2022 微信大数据挑战赛 rank9（全国一等奖） 2022年06月 - 2022年08月

项目描述：本赛题要求基于微信视频号短视频数据(视频、ocr、asr，标题)以及对应的分类标签标注，对短视频进行分类预测，提供了100w的无标注数据和10w的标注数据。

职责描述：

（1）模型结构采用非端到端的结构，首先利用开源的clip vit提取每一帧的特征；视觉部分使用三层的transformer encoder进行帧间的建模（同时添加[cls] token来学习全局特征），文本部分分别截取title、asr、ocr的前90长度的文本利用[sep]拼接之后利用开源bert base提取文本特征；最后利用两个6层的transformer decoder来进行模态融合。

（2）在100w的无标注数据进行预训练（mlm、mfm、vtc、vtm四个任务）；然后在10w标注数据上同时加入对比学习任务和动量蒸馏以及分类任务进行finetune。

2024 AI4S Cup- 大模型提取“基因-疾病-药物”知识图谱 rank7（三等奖） 2024年02月 - 2024年04月

项目描述：使用大型语言模型从海量生物医学文本数据中自动化提取结构化的知识图谱，包括(基因，调控类型，疾病)三元组抽取，(化合物, 疾病)关系抽取，(药物，药物，相互作用)三元组抽取。

职责描述：

（1）数据预处理：对于不同任务构造对应的prompt，然后将数据处理成LLM微调的形式；

（2）模型微调：将知识图谱抽取拆解为 实体抽取任务 + 二/三元组抽取任务，然后使用dora的微调方式对LLM（mistral-7b-instruct-v0.2）进行微调，首先在训练前期使用实体抽取任务数据对模型进行微调，然后在训练后期加入二/三元组抽取任务数据进行联合微调；

（3）在推理阶段，首先利用微调后的LLM抽取文本中的实体数据，然后将实体信息和原始文本信息进行prompt构造后利用LLM抽取二/三元组数据。

2024数字中国创新大赛-少样本条件下的社交平台话题识别 rank9（优胜奖） 2024年02月 - 2024年04月

项目描述：对于每一个task包含query set和support set，其中support set为少量样本，然后附带标签，任务即是从query set中找到和support set同类的话题文本。

职责描述：

（1）数据预处理，过滤文本中一些特殊标识或符号，例如超链接等信息；

（2）模型训练，由于训练集的标签和测试集标签不重叠，因此为了提升模型的泛化性，首先对模型（m3elarge）在给定社交域的话题数据进行继续预训练，然后基于话题数据进行分类的微调，同时使用对抗训练和EMA更新参数提升模型泛化性。

（3）在推理阶段使用 全局相似度+局部相似度进行support和query的匹配。同时由于该任务对阈值比较敏感，不同task阈值并不完全相同，因此根据每个task中support文本之间的相似度来计算阈值；

科研成果

发表国家发明专利两篇

谢东霖，罗崇军，魏骁勇，张栩禄，杨震群. 基于深度学习的多目标视频推荐方法、装置及存储介质：202111134439.X[P]. 2023-11-28.

魏骁勇，**谢东霖**，张栩禄，杨震群. 一种长短不一的文本在不同粒度下的文本匹配方法及装置：202111023691.3[P]. 2023-04-07.

技能/证书及其他

- **技能：** pytorch，深度学习，机器学习，ollama, metagpt
- **语言：** 英语（CET-6）