张良

个人信息

性别： 男

年龄：30岁

联系方式

电话：15302659252

微信号：15302659252

邮箱：zljack2022@163.com

求职信息

工作时长：5年

求职意向：大模型算法

期望城市：深圳

个人优势

1 熟悉 LR、DT、SVM、K-means、GBDT、XGBoost 等机器学习算法；

2 熟悉深度学习任务中 RNN/LSTM/GRU 、Bert、Gpt 系列相关模型以及 Transformer 等深度学习模型；

3熟悉常用模型调优方法（知识蒸馏、模型量化、模型微调、数据增强、训练策略优化)；

4熟悉 cnn，resnet，fasterRCNN，GooleNet,yolov 等深度学习网络模型；

5 熟悉 ChatGLM、llama、baichuan 等大模型熟练使用 TensorRT,Onnx,Fastdeploy 进行推理优化

6 RAG，LangGraph ，Llamaindex，Ollama，Multi-Agent

工作经历

深圳软通动力信息技术有限公司 大模型算法工程师 2024.08-2025.02

1. 数据工程：

收集整理15万条历史客服对话数据，清洗重复、无效样本，标注2万条高质量数据；

设计数据增强策略，通过同义词替换、句子重组扩充数据集，规模扩大至3.5万条。

2. 模型开发与优化：

主导意图识别模型架构设计，实现BERT与BiLSTM的融合，初始准确率达82%；

调优超参数（学习率、批次大小），引入Focal Loss后，少数类意图识别准确率提升18%；

对比测试5种主流模型（如RoBERTa、ALBERT），最终选定BERT - WWm为基线模型。

德科信息有限公司广州分公司 大模型算法 2024.04-2024.06

项目描述:基于企业年报数据搭建本地知识库问答系统，项目主要分为向量数据库搭建、用户 query 解析、本地数据库召回、基于 LLM 文档

回答四个模块，其中用户问题分为统计型、计算型、开放型三个类别，Embedding 模型使用 m3e 模型，向量数据库使用 ES

1、数据预处理入库

2、构建用户 query 解析模块

3、针对统计型和计算型问题

4、针对开放型问题

深圳软通动力信息技术有限公司 大模型算法 2023.08-2024.01

1、用昇腾服务器适配 llama ，百川等模型调优迁移

2、对接生态适配模型

3、用 deepspeed，lora 等训练工具用分布式框架做模型推理，docker 部署模型

4、llama，baichuan，chatglm 3的 ptuing，RLHF，DPO，LoRA 训练模型

5、pandas 数据处理分析

6、 用 doccano 对数据进行 NER 标注

7、 命名实体识别 ，事件，关系抽取f-tuing 垂直领域数据

8、 微调 Chatglm、llama、qianwen进行关系抽取，事件抽取，对话问答

深圳市双银科技有限公司 自然语言处理算法 2023.04-2023.07

知识图谱，数据处理

深圳市博云慧科技有限公司 自然语言处理算法 2020.05-2023.01

负责不同业务场景下的命名实体识别、文本分类、文本摘要、知识图谱,图像分类,目标检测等内容工作,对模型进行设计和优化

项目经历

Agent智能客服系统 算法工程师 2024.08-2025.01

内容:

项目背景：针对传统客服系统意图识别准确率低、长尾问题处理能力不足的问题，为某互联网金融企业开发基于深度学习的意图识别智能客服系统，实现信贷咨询、理财推荐、账户管理等20+业务场景覆盖，目标将意图识别准确率从75%提升至90%以上。

技术实现：

核心模型：采用BERT - WWm预训练模型结合BiLSTM + CRF架构，在金融领域语料上进行微调；引入注意力机制优化上下文语义理解。

数据处理：使用Python的Pandas、Numpy进行数据清洗与预处理；通过bge模型生成词向量，构建金融领域专用词表。

优化策略：

采用Focal Loss解决类别不平衡问题，减少高频意图对模型的主导影响；结合DenseNet网络增强特征提取能力，提升长尾意图识别效果。

部署与集成：使用Flask搭建API服务，集成至企业原有客服系统；通过TensorRT进行模型加速，推理速度提升3倍

性能提升：意图识别准确率从75%提升至92.3%，召回率达89.7%，F1值提高15个百分点；

业绩:

个人职责

1. 数据工程：

收集整理15万条历史客服对话数据，清洗重复、无效样本，标注2万条高质量数据；

设计数据增强策略，通过同义词替换、句子重组扩充数据集，规模扩大至3.5万条。

2. 模型开发与优化：

主导意图识别模型架构设计，实现BERT与BiLSTM的融合，初始准确率达82%；

调优超参数（学习率、批次大小），引入Focal Loss后，少数类意图识别准确率提升18%；

对比测试5种主流模型（如RoBERTa、ALBERT），最终选定BERT - WWm为基线模型。

基于 LLM 构建本地知识库问答系统 nlp算法 2024.04-2024.06

内容:

基于企业年报数据搭建本地知识库问答系统，项目主要分为向量数据库搭建、用户 query 解析、本地数据库召回、基于 LLM 文档

回答四个模块，其中用户问题分为统计型、计算型、开放型三个类别，Embedding 模型使用 m3e 模型，向量数据库使用 ES。

主要工作：

1. 数据预处理入库：将 PDF 年报数据转成 HTML、TXT 格式，使用 pdfplumber、paddleocr 等第三库提取表格数据，文本数据
使用 Langchain 工具进行分块分级处理，两者合并统一提取融合导入到 ES 向量数据库；
2. 构建用户 query 解析模块： 用基于模板的正则抽取和基于 LLM 的关键词抽取方法提取 query 中的数据入库的关键词、别名以
及计算相关的指标，其中基于 LLM 的关键词抽取采用 few-shot 方法。将抽取的关键词与数据库做向量相似度匹配，根据相似度
分数与阈值的大小关系对 query 进行分类；
3. 针对统计型和计算型问题：大于阈值的划分为该类问题，直接生成检索语句，查库得出关键信息，交给 LLM 生产答案，值得注
意的是在 LLM 生成的答案外层需嵌套一层正则，将答案中的公式计算部分重新计算数据替换原答案的结果；
多轮对话系统迭代
4. 针对开放型问题：将用户问题做 embedding 召回和关键词 ES 召回，提取 Top-K 相关分块文本给 LLM的 prompt 模板中生成
对应的 prompt ，由 LLM 生成回答。

业绩:

项目优化：
1. 对文档数据进行分块入库时，先用正则表达式识别标题行，用序号区分类型（如几级标题等），存储标题分层递归关系，记录行
号，根据分层递归的标题层级对正文进行初步切割（使用 Langchain 中的 RecursiveCharacterTextSpliter 函数，以
chunk_size=500 的长度切分），并给分块文本添加层级标题；
2. 使用 few-shot 抽取关键词的效果不尽人意，这里结合 LLM 的 In-context learning 的能力，构造 history ，通过模拟多轮对话
的方式让模型输出更加稳定的 json 结果，对于异常的 json 调整 temperature=1 增大 LLM 生产的随机性，利用它本身的修正能
力多次生成，提取比较好的结果；
3. 结合入库的分块分级操作，做文本召回的时候不仅仅召回 query 文本，还加入了关键词召回，通过这个召回增强的操作提高输
出效果。

模型推理 推理测试工程师 2023.09-2024.01

内容:

内容:
1、用昇腾服务器适配llama,通义千问,百川,星火等大模型调优迁移
2、对接生态适配模型
3、用deepspeed,lora等训练工具用分布式框架做模型推理,docker部署模型
业绩:
处理大规模数据和分布式训练掌握GPU,TPU等硬件加速器的使用和优化;分析并解决在训练和推理中的性能瓶颈或故障,优化模型推理的资源利用率。

业绩:

处理大规模数据和分布式训练掌握GPU、TPU等硬件加速器的使用和优化;分析并解决在训练和推理中的性能瓶颈或故障，优化模型推理的资源利用率。

大模型算法 模型优化 2023.08-2024.01

1、用昇腾服务器适配llama,百川等模型调优迁移
2、对接生态适配模型
3、用deepspeed,lora等训练工具用分布式框架做模型推理,docker部署模型
4、llama,baichuan,chatglm3的ptuing,RLHF,DPO,LoRA训练模型
5、pandas数据处理分析
6、用doccano对数据进行NER标注
7、命名实体识别,事件,关系抽取f-tuing垂直领域数据
8、微调Chatglm、llama、qianwen进行关系抽取,事件抽取,对话问答

民生诉求 NLP算法 2023.04-2023.06

解决民生问题为政府提供更好的对接民生的诉求窗口
1、 pandas 数据处理分析
2、 用 doccano 对数据进行 NER 标注
3、 用 uie 模型进行 NER ， 事件， 关系抽取f-tuing 垂直领域模型
4、 微调 Chatglm-6b进行事件抽取问答

在线客服对话系统 算法工程师 2022.05-2023.01

商品的丰富性和多样性给顾客带来了不同的体验,但是也给顾客造成了些许的困扰,找不到商品或者忘记商品的品名,通过描述商品的属性特征,推荐给顾客想要的商品,一定程度上解决了顾客的需求,同时也督促业务部门采购缺失的商品,丰富商品结构,满足顾客需求,项目描述:
1、数据来源:数据部门给的合计12万条数据。主要职责是协助业务部门对清洗过的数据进行标注并审核标注结果,
2、命名实体识别:基于Bert+CRF,IDCNN+CRF算法搭建命名实体识别模型,使用10万条进行训练
通过badcase分析后再次训练,Bert+CRF模型的准确率由89%提升到91.2,IDCNN+CRF由93.3提升到
94.6最终选用IDCNN+CRF模型。
3、负责将数据写入到neo4j数据库
4、用GPT2做多轮对话任务

顾客投诉及建议文本摘要生成 算法工程师 2021.08-2022.03

随着电商平台的用户日益激增,顾客对商店的服务以及商品需求的日益提升,每天需要处理的投诉和建议在逐渐增加,为了节省人工成本,提高人员效率,快速高效处理顾客的投诉和建议,提升顾客购物体验,
1、从公司服务部门中拉取投诉和建议的数据,进行数据清洗,除去停用词、无效字符等,得到模型训练需求的数据(结合服务部门获取人工摘要)。总数据集18万条。
2、搭建基于GRU的Seg2Seg+Attention的生成式文本摘要V1.0模型
3、解决OOV问题,构建PGN+coverage的V2.0模型
4、基于beam-search优化V2.0模型,解决重复生成问题,完成PGN+coverage+beam-search的3.0模型
5、PGN优化版本在ROUGE-L评估时f指标由31.72%提升33.31%,p指标32.90上升到34.11%由29.26%提升至31.17%;通过模
型优PGN+coverage+beam-search对解决短语重复生成、0OV问题都有较好的效果

网上商城评论情感分类 算法工程师 2021.04-2021.07

项目背景:

公司主要是以线下销售为主,但是随着互联网发展,越来越多的顾客喜欢在网上购买,实现就近配送。为

了能快速了解顾客对配送商品的满意度,具体分析差评商品的原因,进行后期改善。

提高产品的质量合格度,减少员工的工作压力,减低错检率

项目职责:

1、数据分析:标签数量分析、句子长度分布分析,部分数据清洗、句子截断补齐等预处理工作;

2、Fasttext模型数据分类,准确率为88.9%,作为V1.0版模型

3、构建单层BiLSTM+Sigmoid模型V2.0,准确率为90.5%

4、使用gensim工具自训练词向量,Fasttext和BiLSTM准确率分别为93.5%;

涉及技术:Fasttext+Bilstm

教育经历

中南林业科技大学涉外学院 本科 土木工程 2014-2018