

施沫寒 意向：机器学习，自然语言处理

手机：13720037514

籍贯：北京

邮箱：1282575394@qq.com

学历：硕士



项目经历

- | 时间 | 项目名称 | 职位 |
|--|------------|---------------|
| 2019.06 至今 | 星图平台 | 百度搜索策略部/算法工程师 |
| <ul style="list-style-type: none">星图是百度垂搜需求识别策略中台。负责为各垂类召回 query，并解析 query 中的成分以便检索。针对业务需求，优化音乐，教育，菜谱，股票等十余个垂类。召回手段包括基于用户反馈训练 LR 模型，基于 query 语义训练 NN 模型，挖掘词表等。解析手段包括训练 RNN，挖掘词表等。优化后大部分垂类准召达到 90% 以上。召回模型升级 BERT。引入基于 query-相关 title 语义特征的 BERT 离线模型。语义特征较用 one-hot 特征代表相关资源的方式粒度更细，BERT 模型能力更强，优化了小说，影视，音乐，股票等垂类的效果。建设时效性通路，将时效性延迟从天级别降低到小时级别。建设了覆盖特征抽取，特征灌库，离线在线模型计算并灌库结果的整套流程。 | | |
| 2019.06 至今 | 度晓晓意图识别 | 百度搜索策略部/算法工程师 |
| <ul style="list-style-type: none">度晓晓是百度新推出的人工智能手机助理。有内容满足，闲聊陪伴，事务管理等功能。意图识别和内容满足有策略需求。从零到一搭建意图识别系统。业务需要识别搜索，闲聊和提醒唱歌等垂类意图。其中最重要的是搜索闲聊意图识别。选用公开闲聊数据和从公司业务中挖掘的搜索闲聊数据（大量且无需标注）对 BERT 模型做了进一步预训练，再抽取产品真实交互数据人工标注后进行 finetune，得到带上文的 pair BERT 分类模型，准确率达到 92%。垂类意图暂用模板/词表识别，已接入十余个垂类。度晓晓语音交互场景下，唯一答案能通过播报，动画等方式满足用户需求，体验更好。Query 访问垂搜系统后返回一批资源，策略基于搜索点击行为、展现位置、规则模板等特征来判断当前结果需求的集中程度，若集中则出唯一答案。准确率 86.3%，召回率 70%，比比请求完整搜索快 500ms。 | | |
| 2017.11 至 2018.05 | 华夏银行单据智能识别 | 北京交通大学/算法工程师 |
| <ul style="list-style-type: none">银行用户手写单据录入工作繁杂，用 OCR 技术将其自动化，主要任务是识别手写公司名。对字符进行过度分割后，启发式地搜索合并路径空间。使用卷积神经网络进行字符图像的识别，使用循环神经网络构建的语言模型进行对搜索路径进行评分。基于 Tensorflow 和 CNN 训练汉字图像识别组件。数据集有 8,000,000 余张图片，包含 7000 余类。使用数十万公司名构成的字符串数据集训练多层 LSTM 语言模型。准确率达到 90% 以上。 | | |

科研经历

- | 时间 | 项目名称 | 角色 |
|---|----------------------------------|------|
| 2016.11 至 2017.11 | Random Pairwise Shapelets Forest | 第一作者 |
| <ul style="list-style-type: none">该论文发表于 PAKDD'2018 (CCF-C 推荐国际会议)提出了一种针对时间序列分类问题的可解释性较强的随机森林。相较于已有的方法，分类准确率，训练速度，可解释性均有改进。 | | |
| 2018.06 至 2019.06 | 一种基于时间序列特征的可解释步态识别方法 | 第一作者 |
| <ul style="list-style-type: none">该论文发表于《中国科学：信息科学》(CCF-A 推荐期刊)将上文提出的方法应用于步态识别问题，在较小数据集上能在准确率媲美深度神经网络的同时提供较好可解释性。 | | |

工作经历

- | 时间 | 公司 | 部门 | 职位 |
|------------|----|-------|-------|
| 2019.07 至今 | 百度 | 搜索策略部 | 算法工程师 |

教育背景

- | 时间 | 学校 | 专业 | 学位 |
|-------------------|--------|----------|----|
| 2016.09 至 2019.06 | 北京交通大学 | 计算机科学与技术 | 硕士 |
| 2012.09 至 2016.06 | 首都师范大学 | 计算机科学与技术 | 学士 |