王子涵

15001397081 | zw2782@columbia.edu | 北京市海淀区
22岁 | 女
算法工程师

教育经历

清华大学 2017年08月 - 2021年06月
计算机科学与技术 本科
- **GPA：** 3.72/4.0; **荣誉奖项：**社会实践优秀奖学金（2018）、学业优秀奖学金/宜信奖学金（2019）

哥伦比亚大学 2021年09月 - 2023年03月
计算机科学 硕士
- **GPA:** 3.8/4.0

实习经历

摩根大通（纽约） 2022年04月 - 至今
算法研究员 AI Research部 - 语音组 纽约
- 研发在小语种上的语音情感分类模型。使用Allosaurus和MFCC序列特征、以CNN+LSTM+self-attention作为特征提取器，并对GE2E和Wav2vec预训练模型进行微调。
- 创造性地加入基于batch内聚类簇间距离的对比学习辅助loss；并将推荐中的多目标优化模型MMoE与多语言语音意图识别相结合，解决多个域之间的”跷跷板问题“。
- 该模型在德语数据集EMODB和波斯语数据集SheMo上表现均超过现有SOTA，现在准备论文撰写工作。*(Mentor: Akshat Gupta)*

快手科技有限公司 2021年06月 - 2021年09月
搜索算法工程师 研发线 - 搜索技术部 北京
- 开发用户搜索界面排序算法的可解释性。实现了基于交叉特征检测方法NID和模型无关局部可解释性算法LIME的全局交叉特征检测算法GLIDER, 并在现有模型上加入重要的二阶交叉特征，在线上实现了0.4%的AUC提升。
- 为了增强模型自身的可解释性，实现分层注意力可解释CTR预估算法InterHAt和基于SE-layer的FiBiNET。
- 在用户搜索排序算法的用户侧增加用户关注列表的序列信息，尝试GRU4Rec、DIN、DIEN、Transformer等模型对用户行为序列建模，比两层MLP的baseline模型实现了2.8%的AUC提升。
- 实现用华为提出的PAL方法和浅层网络建模位置信息的方法消除position bias，使得长尾物品得到更多曝光。
- 实现华为提出的端到端数值型特征自动离散化和Embedding方法AutoDis，以及无需查表的类别型特征Embedding方法Deep Hash Embedding(DHE)，在AUC效果不降的条件下大大降低模型参数量。 *(Mentor: 李宣平)*

京东 2020年06月 - 2021年04月
算法工程师 京东零售 - 数据共享平台部 北京
- 对京东泰国站CTR预估模块(Embedding+MLP)进行改进，实现xDeepFM, DCN-mix, AutoInt等模型进行高阶特征交互。
- 实现MMoE、PLE、ESMM等多目标预估模型，对京东泰国站商品的点击、加购物车、购买进行预估，同时加入PCGrad进行梯度的裁剪和投影，来减少任务之间的冲突。
- 参与供应链节点健康度评估的PaaS化项目。将Amazon的仓储绩效指标(IPI)的设计思想与京东的实际情况结合，创造性地提出"二级指标选取-HiveSQL构建数据模型-评分算法计算得分"的“三步走”设计流程，使得搁浅两个月的项目走出困境，并推动其在泰国的试点落地。
- 训练泰-英机器翻译模型。构建200余万条泰-英平行语料库，并使用seq2seq+attention的算法训练机器翻译模型，训练得到BLEU评估得分0.59的离线翻译模型。目前数据库中的“商品英文名称”字段使用此翻译模型做泰-英翻译，大大降低了后续NLP算法的难度。
- 参与印尼MKT标签开发。选择topk个信息增益最大的token作为性别预测的特征，并使用GBDT+xgboost集成学习训练性别预测模型，将性别预测算法准确率从78.2%提升到83.7%。(*Mentor: 吴凯*)

研究经历

基于向量迟交互的多语言检索召回 2021年09月 - 至今
哥伦比亚大学 - Database Research Group
- 与Yelp搜索合作，旨在为目前基于lexical matching的检索召回阶段增加一路必要补充。该项目在Github上已获得100+star。
- 在基于上下文迟交互的检索模型ColBERT的基础上，增加两个针对检索设计的预训练任务。共使用Relevance Ranking、Query-Language Modeling、Representative wOrds Prediction 三个预训练任务，在mBERT checkpoint上继续训练。
- 使用15种语言进行预训练，支持15种语言的多语检索。在其他语种上使用小样本微调后也可支持其他语种的搜索。
- 使用MaxSim打分进行召回。在MSMARCO数据集（英语）上Recall@1000达到95.7%
- 使用大规模向量相似度检索库FAISS建立索引；支持批量召回和重排。*(Mentor: Luis Gravano)*

基于人工智能的GRAPES体系预报产品特征挖掘与融合 2019年12月 - 2021年05月
清华大学高性能所
参与“基于人工智能的GRAPES体系预报产品特征挖掘与融合”国家重点实验项目，为2022北京冬奥会开发出“多分辨率、多模式、多要素”的气象预报算法。使用中国气象局提供的MESO-3km和GFS-10km气象预报模式以及超高精度观测数据，训练U-Net+单格点回归层模型，对两种模式进行偏差订正和融合；同时使用SENET赋予气象模型以可解释性。最终实现两米温度误差达到1.5摄氏度以下、10米u风和10米v风误差达到1m/s以下，达到冬奥会标准。该工作被《焦点访谈》报道。*(Mentor: 赵颖)*

竞赛经历

链想家计算科技大赛 人工智能前沿组 - 论文人名冷启动消歧挑战赛第四名 2020年05月 - 2020年06月
队长
使用Biendata同名消歧数据集，利用同名作者论文的信息，如标题、作者机构、摘要、关键词等，将论文分配到正确作者的档案中。使用网络嵌入(network embedding)中的deepwalk模型，用随机游走和word2vec的得到节点关系向量；用word2vec模型得到论文语义表征向量。使用DBSCAN和层次聚类相结合的方法进行论文聚类，最终得到Macro Pairwise-F1度量为0.91267的聚类结果，排名4/139

专业技能

- **编程语言：**熟悉python, HiveSQL, C++, Matlab, R, Java, shell；深度学习框架惯用pytorch和tensorflow
- **英语水平：**GRE V: 162/170 Q: 170/170 (全球top5%)；托福114/120(全球top1%), 听说读写全部为advance
- **其他：**热衷技术分享，知乎有3k+粉丝（ID: 魔法学院的Chilia）