

工作经历

- 2020.10 - 今

英国巴克莱银行，数据科学家, 数据科学团队

- 企业地址匹配原型构建。使用companies house，gleif等数据构建外部数据集，使用外部数据微调BERT。使用Pandas + PySpark构建了内外部融合数据集（75万对正向地址对），然后使用Spark + Deep Java Library从0构建推理job完成对内部数据进行地址匹配推理。
 - 交易数据图原型构建，使用网络分析对节点的交易情况和供应链进行分析和可视化。
 - 用户活动分析，尝试分析系统故障对用户的影响。
 - 在团队内维护Spark集群，使用Deep Java Library 和 PySpark UDF帮助团队构建分布式推理流程。
 - 参加了一个6周的云上DevOps培训，包括完整构建一个使用GitHub, DockerHub, Jenkins和AWS EKS集群完整的部署流程。构建过程使用Terraform创建基础架构，并用Ansible执行自动化部署安装。
- 2020.08 - 2020.09

快手 自然语言处理工程师，内容与风险管理

技术栈: Spring, Ceph, Dragonfly, Tensorflow, Faiss, Docker, Gitlab
 - 为一个内部新构建的模型管理模型批量迁移模型
 - 维护内部推理系统（基于Spring）
 - 使用百万级别的数据对基于Transformer的用户简介风险模型进行重训练

在快手的短暂经历是非常非常开心的，认识了很多朋友也学到了很多知识。主动离开主要是因为原公司（巴克莱）提供了一个数据科学的机会，反复思考下做出的艰难决定。
- 2019.08 - 2020.07

英国巴克莱银行，毕业生后端开发, 卡平台团队

技术栈: Jenkins, Jira, Confluence, BitBucket, Openshift, Docker, GridGain, Maven, Gradle, Wiremock, Mockito, Spring Boot, AWS, SonarQube, Karate, AppDynamics
 - 端到端功能开发，测试（单元，功能，性能），部署（CD）
 - 为API添加缓存层，以减少重复数据获取的延迟
 - 使用内部的Spring Boot模板迁移遗留代码，并对遗留代码进行重构以提高其可读性和性能
 - 从头构建内部使用的工具（比如 git hooks）和 脚本（python / bash）来自动化软件开发工程，并减少潜在的人类错误
 - 参与了NLP项目（比如基于BERT的地址命名实体识别），同时跟进NLP领域的最新进展，特别是在迁移学习，模型压缩（低资源推断）和多模态融合方向
 - 在新冠疫情期间远程工作的半年中，和团队一起仍然保持了较高的沟通和交付效率

教育背景

- 2018 - 2019

伦敦大学学院，网络科学和大数据分析（硕士），Distinction

核心课程: 概率图模型; 复杂网络; 情感计算; NLP; 信息检索; 多智能体AI; 应用机器学习; 深度学习导论
- 2016 - 2018

英国利物浦大学，互联网计算，一等荣誉学士

核心课程: 软件工程; 数据库开发; 网络原理 (OSI导论); 面向对象编程; 分布式系统原理; 软件开发工具（主要关于测试）; C语言和内存管理; iOS编程 (Swift); 知识表达和推理 (Modal Logic 和 Descriptive Logic); 多代理系统 (MARs); E-commerce (拍卖机制, RSA, DH密钥交换, 椭圆曲线加密)
- 2014 - 2016

西交利物浦大学，信息与计算科学

核心课程: 计算机系统; 数据库导论; Java编程导论; 算法基础和问题求解; 数据结构; 操作系统概念; 微积分; 离散数学导论

项目

- 2019.06 - 2019.09

项目实习 (研究生学位论文) @ Astroscreen

技术栈: Python, Keras, Tensorflow, MulticoreTSNE, Matplotlib

社交网络语言的来源识别 (推特和Gab) 完成了一个爬虫从Gab.com爬取语言 (帖子)，使用正则表达式预处理数据，fine turning BERT 和 XLNet来分类文本来源，并将输出使用 t-SNE可视化，使用准确率，F1 Score, 困惑矩阵, 马修斯相关系数评价模型。
- 2019.03 - 2019.04

信息检索课程项目

使用Fact Extraction and Verification (FEVER)数据集进行了多项练习。
包括词频统计并验证zipf's law;实现向量空间文档索引 (TF-IDF); 实现查询似然 (Query likelihood) 文档索引（并分别应用Laplace平滑, Jelinek-Mercer平滑和Dirichlet平滑），实现逻辑回归比较句子相似性; 实现Precious, Recall和F score函数; 使用神经网络检验文档Truthfulness。
- 2019.02 - 2019.03

Integrating BERT and Embeddings into CommonsenseQA Challenge

我们在CommonsenseQA 1.0数据集（3选项）上fine-turning了Google BERT并尝试整合了Conceptnet Numberbatch and ELMo 词嵌入来尝试提高模型性能。这个数据集包含一组需要人类常识来回答的多选题。
使用BERT+ELMo的组合我们在验证集上取得了68.79%的准确率（BERT: 67.47%; BERT + Numberbatch: 67.68%）。多次实验取最佳结果
- 2019.02 - 2019.03

通过CTR预测和拍卖函数选择以最大化广告点击数

在一个实时广告需求方平台拍卖的历史数据库中预测是否用户想要点击在线广告。预测结果作为一个拍卖竞价函数的输入去预测出价价格。同时总出价会有一个最大限制。数据集是非平衡数据集（超过300000条拍卖纪录中约有3000个正样本（点击数））。我们尝试了许多不同的模型（XGBoosting, 浅层nn, 逻辑回归）和一些竞价函数。我们也在这个项目应用了 undersampling 和重矫正技巧。我们与其它学生在leaderboard进行了比赛（共30组学生），取得了第3名。
（我们取得了185个点击，前两名均取得了186个点击）。
- 2018.12 - 2019.01

上传越多的up主越受欢迎吗？一个对bilibili基于网络的分析

技术栈: Python, networkx, graph-tool, MySQL

该项目检验了一些b站用户（up主）关注网络的属性（度分布和assortative系数），并通过网络可视化检查了是否上传视频数量（反映活跃度）和节点入度（反映受欢迎程度）是相关的。过程中写了一个爬虫来通过b站的RESTful API抓取数据。

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