王昊轩

+86 18622468042

billweasley20092@gmail.com

Github: https://github.com/billweasley

领英: https://www.linkedin.com/in/horace-haoxuan-wang

工作经历

- 2022.01 - 现在 **Garena | Sea Group** 机器学习工程师

  远程入职一周，发现实际工作内容和招聘时的工作描述不太一致，这个岗位是给Booyah!直播构建推荐后台，可能叫后台开发更合适一点。

- 2020.10 - 2021.12 **英国巴克莱银行**，数据科学家, 数据科学团队

  - 在无GPU和标签数据的情况下，进行企业地址和实体匹配。使用主动学习方法，从基于外部数据构建一个小数据集开始，训练一个XGBoost树模型，然后标注一些边界样本，并且迭代微调BERT tiny模型。 从0开始，在内部CPU集群上借助Deep Java Library 和Spark完成了pipeline, 并对600万内部地址对完成推理。在3万对的内部有噪音验证集上，该模型取得了94%的 F1 score (对比之前的CNN+BiLSTM模型为89%且泛化性能很差). 模型离线定时运行，600万对内部地址的运行时间大概在1个小时以内（集群有80块CPU)。
  - 使用历史平均值和Informer模型预测聚合后用户的交易数据（交易量和交易数值）。Informer模型是一个基于Attention机制为长序列优化的模型。取得预测值以后构建反事实，然后基于反事实去计算银行每次系统故障会产生多少损失并分析系统关键时间。
  - 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。和同事合作，我们写出了一个包可以在4行代码内启动一个Spark Session, 帮助没有分布式经验的同事使用容易的使用这一工具。
  - 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。

- 2019.08 - 2020.07 **英国巴克莱银行**，毕业生后端开发, 卡平台团队

  - 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

- 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

  是在Barclays休假的时候入职的，当时面临职业选择，快手给了我一个offer然后Barclays也同样给了换岗的机会。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。

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; 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

- 2021.09 - 2021.10 **微信历史记录分析**

  是给某个朋友的相识周年的礼物，是一个端到端的项目。我收集了我们所有的聊天记录（微信不提供公开方式导出数据），并对如下方面进行了分析：单句情感分析（在一个标签分布均匀的6类微博情感数据集(开心，中性，生气，害怕，焦虑，兴奋)上fine-turning了一个Roberta，取得了80%的accuracy [因为是个人项目且标签均匀只简单看了下accuracy]), 词云生成, 微信emoji统计, 和聊天实际统计. 最终产出是一个用 wechat-h5-boilerplate写出的包含所有信息的html5展示页.

- 2019.06 - 2019.09 **项目实习 (研究生学位论文) @ Astroscreen**

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

  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 Chanllenge**

  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

- 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