刘子祥

北京 • 后端研发工程师 • liuzixiang@zxlumen.cn • +86 156 1563 6435

基础信息

- 概况：6年工作经验 | 男 | 29岁
- 教育：山东理工大学 | 本科 | 计算机科学与技术 | 2014.09 - 2018.07

专业技能

- 精通：Python, MLOps, LLMOps, Ray, Django, FastAPI, Redis, Docker, Langchain, Langflow
- 熟悉：C/C++, Mysql, MQTT, Nats, Kafka, Gunicorn, Websocket, K8s, Grafana, Promethues
- 了解：Java, Golang, Deep Learning, Machine Learning, Data Science, Kaldi, MFCC, i-vector, GMM-UBM

工作经历

腾讯 **2025.04 – 2025.10**

工作职位：后端研发工程师

百度 **2018.07 – 2023.10**

工作职位：后端研发工程师

工作内容:

后端研发工程师/腾讯数据平台部（TEG） **2025.04 – 2025.10**

• Keywords: Python, Ray, MLOps, Model Scheduler, MQProxy

负责音频预处理管线一部分算子的调研与工程落地工作，以及消息队列、存储等中间件的链路优化改造工作。后续负责音频主管线的编排工作，以及基于Ray框架的管线框架开发工作。目前已完成音频17个算子、4条主管线的开发，并已投产使用。

后端研发工程师/百度智能云事业群组（ACG） **组织架构调整 + 转岗 2023.04 – 2023.10**

• Keywords: Python, LLMOps, MLOps, Model Scheduler, Ray, Ray Serve, FastAPI, Websocket, ChatAPI, K8s, Grafana, Prometheus, Langchain, Langflow, MQTT, Nats, Streamz, Simulation Optimization Engine

负责工业仿真优化系统和大模型应用系统的设计与核心研发，应用在了SOE仿真优化平台、PHM安全质检、安全生产智能助手（度安安）、开物数智化平台。均于2023年9月在百度云智大会发布。其中SOE、PHM已应用在中海油、隆鑫发动机检测等项目，度安安已在水务、易度等项目部署应用。

算法 + 后端研发工程师/百度知识图谱部（KG） **组织架构调整 2022.01 – 2023.04**

• Keywords: Python, MLOps, Django Service, Gunicorn, Docker, Kafka, Mysql, SQL Server, WSL, Cross-platform Model Train, Industry Model, Machine Learning, RL

负责工业场景的模型调研、设计、应用及优化，后续沉淀研发了MPO-Tools（Core Lib）+ MPO-Service（模型服务）+ MPO-Plant（模拟产线）的生产过程优化套件套件（iFactory Suites）。负责整体设计，以及核心算子库、模型服务框架的研发工作。结合业务场景进一步实现了跨产线结构的服务与模型迁移适配、以及模型自学习能力。上述套件已在轧钢、水冷、冲压、镀锌、制丝等多个工业领域实际投产使用。

数据挖掘工程师/百度大数据部（BDG） **转岗 2020.05 – 2022.01**

• Keywords: Python, Social Words, Social Network, Social Network Quality Metrics, Paddle, PGL, GraphLearning, MapReduce, Metapath2vec, Embedding, Dataflow Relationship Diagram, Language Model, Clustering, Feature Engineering

负责互动关系网络的模型优化，以及用户互动领域体系建设。提升了现有关系网络的效果，应用在了手百、好看、贴吧、金融等场景。同时搭建了数据流关系图谱和效果评估平台，降低了模型评估难度。同时将模型训练自动化，提升了模型迭代效率。

算法研发工程师/百度度秘事业部（DuerOS） **入职 2018.07 – 2020.05**

• Keywords: C/C++, Acoustic Echo Canceller (AEC), Automatic Speech Recognition (ASR) Accuracy Optimization, Speaker Recognition, Kaldi, Performance Optimization, Clustering, MFCC, i-vector, GMM-UBM

负责声学算法的调研、研发、落地工作，以及C++到C的代码迁移工作。包括基于Kaldi的语音识别（ASR）解码器迁移，回声消除（AEC）、声纹注册等算法的研发与落地，MFCC、i-vector等声学特征的提取优化，以及灌测工具的研发与优化。以上产出应用在了蓝牙耳机、车载支架、小度音箱等多个硬件产品。

后端研发工程师/百度金融 **实习 2017.07 – 2017.12**

• Keywords: JAVA, SpringBoot, Mysql, Redis, AMQ, Monitoring

负责股市通项目的维护与迭代工作，包括转码器、后端服务、监控的研发与维护、以及数据信号传输中间件BasicInfoCollect模块的代码重构工作。

重点项目经历

* 数据平台部（TEG）

音频管线开发 **2025.04 – 2025.10**

• Keywords: Python, Ray, MLOps, Model Scheduler, MQProxy

负责音频预处理管线一部分算子的调研与工程落地工作。

- 调研并落地了基于wav2vec2的音色识别算子。以及结合说话人识别模型与语音识别模型，实现了说话人字粒度时间戳预测算子。均已落地并支持需求。
- 优化管线架构：
  - 优化之前，原管线通过spark任务进行数据送标与产出交付，算子订阅kafka获取送标数据并处理，通过挂载ceph进行中间数据的存取。存在ceph带宽限流、高优需求的数据不支持插队等问题。
  - 优化后通过mq_proxy对送标数据进行分发，转换为通用的HTTP协议，方便不同模块调用。同时引入了不同优先级的Topic，结合算子上游的数据分发服务以及七彩石配置管理，实现高优任务插队的能力，使用cos进行中间数据存储，替代了挂载ceph的依赖，解决了跨地域带宽不足、ceph容量有限等问题。
  - 另外基于multiprocessing对算子进行了改造，基于平均资源利用率，对并发量进行微调，同样实例数量的情况下，不同算子日产能均有2~4倍的提升。

后期负责音频主管线的设计开发工作。

- 与算法同学紧密配合，设计搭建了语音主管线框架。基于现有的17个语音算子，通过参数化的方式实现了中英日韩四条管线。
- 借助ray框架的Dataset模块，取代了原本的spark投产与交付任务。借助cos实现了实时拉取、加载算子代码与模型文件的能力。借助Actor模块支持了不同算子的多集群节点混部、算力资源共用与动态调整等能力。
- 开发了前置的送标数据分桶任务，结合小时级管线调度，实现了几乎无需人工干预的自动投产处理的管线。

* 智能云事业群组（ACG）

Windmill_Promptflow - 工业大模型服务 **2023.07 – 2023.10**

• Keywords: Python, LLMOps, Ray Serve, FastAPI, Websocket, ChatAPI, K8s, Grafana, Prometheus, Langchain, Langflow

这个项目用于将工业场景中的大模型应用服务化。

基于此框架的安全生产智能助手（度安安）和开物数智化平台已于2023年9月在百度云智大会发布。目前已在水务、易度等多个项目部署应用。

我在此项目中负责服务框架的设计和核心研发。以及与前端、算法同学的代码适配和联调工作。

- 此项目基于RayServe + FastAPI + Websocket提供大模型问答服务。
- 基于langchain框架调用大模型。基于langflow框架做大模型链（应用）的定制化构建。
- 实现了APP与Host分离式部署，动态识别加载大模型应用。算法同学基于此服务开发发布了20+个大模型APP。提供了智能助手、API智能调用、文档检索、BI取数、图表绘制等能力。
- 支持了低代码的大模型应用开发，即大模型思维链的创建。已支持200+个定制化APP。

SOE - 工业仿真优化系统 **2023.04 – 2023.07**

• Keywords: Python, MLOps, Ray, Model Scheduler, K8s, MQTT, Nats, Streamz, Simulation Optimization Engine

该项目是为了解决工艺优化仿真问题，研发的一套通用的工艺优化仿真子系统。

基于该套系统研发的PHM安全生产平台、SOE平台均于2023年7月发版，已应用于成都加药、中海油、隆鑫发动机检测等项目。其中SOE平台于2023年9月在百度云智大会发布。

该系统由数采系统、模型推理系统、控制系统组成。我主要负责模型推理系统的设计和核心研发。

- 此系统基于Ray + Nats + Streamz实现。
- 设计了SchedulerActor、WorkerActor、Task组件，通过负载均衡，解决了Ray单个任务占用资源过多的问题。
- 基于Nats +Streamz实现了通用的模型推理数据传输组件，支持按时间/点位数据调度，离散、连续型数据，以及低速、高速数据。实现了通用的模型数据流、预处理、推理调度、后处理以及多模型的数据共享。

*** 知识图谱部（KG）**

iFactory_Suites - 生产过程优化套件 **2022.01 – 2023.04**

• Keywords: Python, MLOps, Django Service, Gunicorn, Docker, Kafka, Mysql, SQL Server, WSL, Cross-platform Model Train, Industry Model, Machine Learning, RL

这是一套通用的工业场景的生产过程优化套件，通过工业产线的环境参数优化机器的控制参数，进而提升产品质量、减少生产成本。

已于轧钢、水冷、冲压、镀锌、制丝等多个工业领域实际投产使用。

我在其中负责整体设计，核心算子库、模型服务框架的研发工作。以及在不同业务中的适配与落地工作。

- 该套件分为MPO-Tools（Core Lib）, MPO-Service（模型服务）, MPO-Plant（模拟产线）三部分。
- MPO-Tools基于scikit-learn、torch、paddle算法库实现，提供了核心的模型通用算子。
- MPO-Service是基于gunicorn + Django开发的模型服务，提供了通用化的产线、模型管理、模型调度、跨产线自学习等能力，已在多个工业场景验证和应用。
- MPO-Plant是离线模拟产线的框架，负责将以上两者串联，进行离线的产线上模型效果验证工作，减轻了算法同学在客户侧POC的人力成本。

*** 大数据部（BDG）**

互动关系网络模型优化 **2020.05 – 2022.01**

• Keywords: Python, Social Network, Social Network Quality Metrics, Paddle, PGL, GraphLearning, Metapath2vec, Embedding, MapReduce, Dataflow Relation Diagram

这是基于用户画像构建的用户互动关系网络，用来在用户之间做特征共享、群体分析，以及新用户冷启动。

联合ACG合力支持北京顶象、北银消费金融、上海墨非、度小满、中信百信银行、南京银行等多家金融客户。与MEG侧在手百、好看、贴吧侧DAU均有显著提升，并已使用。

我主要负责模型效果与性能优化。

- 引入Metapath2vec模型，整体auc从78.58% -> 80.78% (+2.2%)
- 与PGL团队合作，升级Graph结构，同时引入分布式游走算法。结合MR任务，使其能够支持10亿+用户、600亿+关系。
- 引入图文、直播、金融等领域的u-item关系，由同构网络升级为异构网络，在各领域得到的很好的效果。
- 建立了关系网络数据流关联图谱，覆盖了30+任务模块，便于查看数据流向，提升定位、解决问题的效率。
- 设计研发了关系网络质量评估任务，可视化观察网络优劣。
- 将MR -> paddle训练 -> 后处理流程自动化，迭代效率上由2人天降低到0.5人天。

互动领域体系建设 **2020.05 – 2020.12**

• Keywords: Python, Social Words, Language Model, Clustering, Feature Engineering

这是基于用户互动关系网络构建的用户互动领域体系。挖掘用户互动领域，提升用户画像效果。

该套体系已推广在婚姻模型、FSU金融业务侧使用。支持FSU金融特征，AUC +9.7%，KS +7.0%。加入到婚姻模型中，准确率84.01% -> 84.89% (+0.88%)，覆盖率91.85% -> 93.33% (+1.48%)。

整套体系由我带头搭建，以及推动评估和应用。

- 基于互动关系网络，结合NLP语言模型，得到用户核心社交词。设计了互动关系->社交词->互动领域的聚类算法，得到用户的互动领域语义化表征。进而构建了互动领域体系，整合了1w+的互动类别。设计了互动领域相似度算法，整合为1048个互动领域。
- 设计研发了评估工具和评估平台，通过BDID检索用户的互动领域，用户打分并在后台整理，用于厂内评估与反馈。

*** 度秘事业部（DuerOS）**

蓝牙耳机及车载支架，前端信号处理算法的研发与落地 **2019.02 – 2020.05**

• Keywords: C/C++, Acoustic Echo Canceller (AEC), Accuracy Optimization

此项目中我负责回声消除（AEC）算法的调研、研发、优化及落地工作。

- 基于PBFDAF算法，将Matlab版转为C版本。
- 设计研发了基于音频互相关的在线时延估计算法，优化掉了一部分延时，减轻了AEC运算负荷，设备CPU占用率降低8%（24MIPS），同时改善了AEC效果。
- 调研了单麦、双麦降噪算法，集成了基于噪声估计+谱减的单麦降噪算法，80db噪声下，可以做到基本与手机通话效果一致。唤醒率、识别率与友商打平，通话效果超过友商，已在集成到了耳机与支架产品中。

语音识别（ASR）解码器，优化以及从C++到C的平台迁移工作 **2019.02 – 2020.05**

• Keywords: C/C++, Automatic Speech Recognition (ASR), Kaldi, Performance Optimization, LCS, Knapsack dp

此项目中我负责ASR的优化及迁移工作。

- 基于Kaldi（ASR开源工具），完成了解码器及Nnet3从C++到C的迁移及性能优化。识别时间缩短30%，误识别由247次/36h -> 31次/36h
- 这期间自研了一套灌测工具，基于LCS + 01背包dp，单次匹配时长由几分钟减少到几秒钟。已完全取代人工匹配，单次灌测有上万条，预计每1000句节省1人力1小时的工作量。抽样检测，与人工匹配结果100%一致

无感知的声纹（说话人）注册，研发与优化 **2018.07 – 2019.02**

• Keywords: C/C++, Speaker Recognition, Kaldi, Clustering, MFCC, i-vector, GMM-UBM

此项目中我负责算法调研、研发工作。

- 基于query提取MFCC、i-vector特征，基于GMM-UBM训练说话人模型，采用层次聚类，实现了【无感知】用户注册。
- 在保证精度>=99%的同时，召回率可达80%。已部署使用。

*** 百度金融（实习）** **2017.07 – 2017.12**

• Keywords: JAVA, SpringBoot, Mysql, Redis, AMQ, Monitoring

我在实习期间，主要负责股市通项目的维护与迭代工作。

- 协助研发了股票/基金投顾管理平台，负责转码器模块，对沪深数据解码及包装，并存于Redis，同时通过AMQ进行各类股市信号的接收与传递。
- 协助维护了部分后端代码，监控各服务器状态，以及股市信号是否正常发送。同时负责 值班，定位 、解决股市盘前、中 、后可能出现的服务异常问题。
- 协助重构了用于数据及信号传输的中间件BasicInfoCollect模块，基于SpringBoot框架 ，进行MySQL的相关数据操作，包装后存于Redis供下游使用。

其他

校园荣誉

- 大学期间任集训队2014级队长，组织队内训练以及带队参赛。
- 2017年获得国家奖学金。

比赛荣誉

ACM-ICPC国际大学生程序设计竞赛亚洲区域赛	1枚	金牌	2016.11
CCPC中国大学生程序设计竞赛	2枚	银牌	2016.09, 2016.10
ACM山东省省赛	2枚	银牌	2015.06, 2016.06
CCCC中国高校计算机大赛团体天梯赛	1次	一等奖	2016.07
另获 4枚ICPC铜奖, 1枚CCPC铜奖	5枚	铜奖	

工作荣誉

大黄蜂车载支架项目	百度小赞	项目奖	2019.09
用户互动关系网络	百度小赞	项目奖	2020.10
工艺优化项目	百度小赞	项目奖	2021.10
个人	季度之星	个人奖	2023.01

SOE仿真优化平台	在百度云智大会发布。	2023.09
安全生产智能助手（度安安）、开物数智化平台	在百度云智大会发布	2023.09
度安安	入选【2023年人工智能融合发展与安全应用典型案例】领航型	2023.09

专利

度秘事业部	**2**项
一种自动计算语音识别字准句准的方法	2019.06
一种计算信号时延的在线算法	2019.07
大数据部	**8**项
互动领域的标签挖掘方法	2020.09
一种突发公共事件相关人群识别方法	2020.09
基于用户线上互动的关系网络建设与应用	2020.09
基于互动关系网络挖掘的用户表征集建设及联合应用	2020.09
一种基于用户交互网络图向量的推荐系统召回方法	2020.09
基于用户交互网络的画像属性预测方法	2020.11
一种用户时序关系网络建设方法	2021.07
基于用户交互网络图向量的金融风控打分支持方式	2021.10
知识图谱部	**3**项
一种基于进化算法及预测模型的钢铁镀锌工艺参数优化方法	2021.10
一种基于遗传算法的冲压工艺控制方法	2021.11
基于聚类进化的冲压工艺优化方法	2022.02

附

- 希望今后能参与有建设性的工作，与大家一起做出创造性的成果！