任广凯

算法工程师

邮箱：kevin.r5009@gmail.com 电话：+86 184 0124 5270 地点：北京市

个人简介

八年算法研究与工程经验，专注于模型算法的应用与优化。善于快速学习并复现 NeurIPS、ICLR、ACL 等顶会最新研究成果，具备从理论到工程实现的全链路能力。对各类大模型及经典中小模型有深入理解，能够精准分析其优势与局限，并结合业务需求进行高效优化与落地，推动算法在实际场景中的创新应用。

专业技能

Python | Java | JS/TS | PyTorch | LLM/VLM | 多模态理解与生成 | 大模型预训练 | Agent | 信息检索/RAG | 高效微调PEFT | 对比学习 | 强化学习 | 蒸馏/量化 | 异常检测 | 模型工程化部署 | 分布式/微服务 | DevOps/MLOps

工作经验

北京区块畅想数字科技有限公司

2023年6月 - 至今

资深算法工程师

- 负责 AI agent、大模型应用及智能搜索系统的研发，在区块链、加密交易、信息检索等领域推动 AI 技术落地。
- 主导多个 AI 项目的架构设计和技术实现，包括 LLM-baseed agent、智能合约交互及自动化交易系统。
- 参与区块链生态系统开发，构建 AI 驱动的 DeFi 交互、跨链资产管理及自动化智能合约执行方案。
- 负责 AI 搜索引擎的设计与优化，提升复杂查询的解析能力，实现 multi-agent 协作的信息检索系统。
- 在 AI 量化交易决策领域结合 LLM 及强化学习，优化交易策略，提高市场分析的准确性。

中国知网技术有限公司

2016年11月 - 2023年6月

研发工程师 & 高级项目经理

- 负责搜索引擎和推荐系统的设计与开发，包括全文检索、多路召回及个性化推荐算法优化。
- 主导文本纠错系统研发，结合NLP技术与多模态特征融合，提升复杂场景下的纠错能力。
- 设计并实现AI增强检索(RAG)系统，优化向量检索和语义理解能力，提高信息获取效率。
- 负责工业领域计算机视觉系统开发，专注于异常检测与缺陷识别的零样本学习方案。
- 参与AI模型训练与优化，涵盖BERT、Transformer、CLIP等大模型技术在多领域的应用。
- 构建大数据处理平台与AI基础设施，支持海量数据的离线批处理和实时流处理需求。

北京国美在线电子商务有限公司

2014年12月 - 2016年8月

研发工程师

- 负责基于 Elasticsearch 构建商品搜索引擎，优化索引结构和分词策略，显著提升搜索响应速度和结果精准度。
- 主导开发个性化推荐系统，运用协同过滤和矩阵分解算法精准匹配用户与商品，有效提高转化率。
- 参与搭建分布式计算架构，利用 Hadoop 和 Spark 进行高效数据处理，结合 Redis 加速热门商品推荐，显著提升系统性能和效率。
- 负责引入 Kafka 构建实时数据流处理框架，根据用户实时行为动态调整推荐结果，增强用户体验。
- 主导开发机器学习模型，使用 scikit-learn 和 MLlib 实现逻辑回归和随机森林推荐模型，显著提升推荐系统的泛化能力和稳定性。

北京曙光无限技术有限公司 2014年9月 - 2014年12月

研发工程师

- 负责基于肤色检测和双边滤波算法实现美肤功能，通过调整肤色区域的色调和平滑皮肤纹理，显著提升人像照片的美观度。
- 主导开发瘦脸和大眼特效，运用液化算法结合dlib提取人脸特征点，精准调整面部比例，使用户获得更理想的自拍效果。
- 参与实现滤镜功能，采用卷积神经网络 CNN 和神经风格迁移算法，为用户提供多样化且具有艺术感的照片风格转换效果。
- 主导设计并实现照片修复功能，运用深度学习算法对低质量照片进行画质提升，有效改善模糊、噪点等问题，增强照片的整体视觉效果。

项目经验

Avalanche(雪崩) Agent MCP framework

2025年1月 - NOW

项目概述：负责开发并维护Avalanche区块链网络上的MCP框架，该框架通过 AI 代理实现对 Avalanche 区块链协议的自动化操作，涵盖多种链上操作，包括代币交易、NFT 管理、DeFi 交互、跨链桥接等。该项目旨在简化开发者与 Avalanche 生态的交互，并推动 AI agent在区块链领域的应用。

- 开发了Avalanche Agent MCP framework，为开发者提供了便捷的 Avalanche 智能合约交互工具。
- 实现了几十种 Avalanche 链上操作的自动化，显著提升了开发效率。
- 集成了 LangChain 和 Vercel AI SDK，构建了支持自然语言交互和自主决策的 AI 代理。
- 实现了与多个主流Avalanche DeFi协议的集成，丰富了工具包的功能。
- 实现了与Wormhole和deBridge的集成，打通了Avalanche与其他链的资产转移通道。
- 开发了基于LangGraph的多Agent系统，提升了工具包的复杂任务处理能力。

LLM-based 加密资产量化交易系统 CryptoTradeAgent

2024年6月 - 2024年12月

项目概述：负责设计并实现了一个基于大语言模型（LLM）的加密货币交易代理，CryptoTradeAgent。该代理能够综合分析链上数据（如交易记录、市场数据）和链下信息（如金融新闻）来制定交易策略。通过引入反思机制，代理能够回顾过去的交易表现，并据此优化未来的决策，从而在加密货币市场中实现更高的回报。

- 设计并实现了CryptoTradeAgent的整体框架，包括数据收集模块、市场分析模块、新闻分析模块、交易决策模块和反思模块。
- 开发了数据收集模块，负责从CoinMarketCap和Dune Database等平台获取链上数据，以及从crypto金融新闻源获取链下信息。

- 利用LLM实现了市场分析agent和新闻分析agent，分别负责分析市场趋势和解读新闻事件对市场的影响。
- 设计了交易代理的决策逻辑，使其能够根据市场分析和新闻分析的结果，生成买入、卖出或持有的交易建议，并控制交易量。
- 引入了基于强化学习的反思代理，使其能够回顾过去的交易记录，分析成功和失败的经验，并为未来的交易决策提供反馈。
- 使用deepseek-r1和GPT-o1对CryptoTradeAgent进行了实验评估，并将其性能与传统的交易策略和时间序列模型进行了比较。

结合多子问题并行查询加多级检索策略的AI搜索引擎

2023年7月 - 2024年6月

项目概述：主导开发了Thought-like AISearch，一个基于大语言模型（LLM）的多智能体Web信息检索与集成系统。该系统模仿人类思维过程，通过规划和搜索两个核心模块协同工作，实现了对复杂用户查询的深度解析和高效信息检索。

- 开发了搜索规划 agent 模块，负责将复杂的用户查询分解为多个原子子问题，并构建动态的推理图，指导信息检索过程。
- 开发了搜索执行 agent 模型，负责执行规划 agent 分配的子问题，通过分层检索策略，从海量网页中提取有价值的信息。
- 引入了多智能体协作机制，实现了规划 agent 和 搜索 agent 之间的信息交互和任务分配，提升了系统的整体效率和准确性。
- 设计并实现了LLM上下文管理策略，有效处理长文本输入和多轮对话，避免信息丢失和干扰。

基于多模态大模型 OpenCLIP 的零样本工业产品异常检测系统

2022年10月 - 2023年6月

项目概述：本项目构建了一套零样本异常检测系统，用于工业缺陷检测。基于多模态大模型 CLIP 方法，我们优化了 CLIP 在异常检测任务中的泛化能力，实现了无需目标领域训练数据即可精准检测异常。

- 设计并实现了 OpenCLIP 异常检测系统的整体架构，包括对象无关的文本提示学习、全局+局部异常特征优化机制。
- 开发对象无关的文本提示学习机制，从泛化角度优化 CLIP 识别能力，使其聚焦异常区域而非前景物体类别。
- 实现全局+局部异常学习（Glocal + Local）优化机制，结合全局特征和细粒度异常特征，提高模型在多场景下的检测精度。
- 优化 CLIP 的文本与视觉空间，采用可学习文本嵌入（Prompt Tuning），增强局部异常识别能力。

基于多路召回和 Learning-to-Rank 的文献增强检索系统

2023年2月 - 2023年6月

项目概述：负责开发基于 AI 技术的增强检索系统，旨在提升学术文献、企业知识库等海量信息检索的精准度与效率，满足用户复杂多样的查询需求。

- 参与了AI增强检索系统的设计与开发，包括系统架构设计、功能模块开发和性能优化。
- 设计并实现了融合传统检索与向量检索优势的多路召回检索模块，提升检索的准确性和全面性。
- 开发了快速直达文献原文段落的功能，使用户能够跳过文献层面，直接定位到原文片段，提高文献调研的效率。
- 设计并实现了自由灵活的检索输入功能，支持自然语言输入和语音输入，并具备智能识别检索意图、智能纠错和规范引导检索的能力。
- 参与了系统的数据处理工作，包括文献数据清洗、向量化加工和全维度向量数据总库的构建。

基于多模态融合的文本纠错 MMSpellCheck 系统开发与优化 2022年2月 - 2023年12月

项目概述：该项目突破传统文本纠错方法在复杂场景下的局限性，利用多模态信息的互补优势，实现更精准、更自然的文本纠错。

- 参与了 MultiModalSpellCheck 系统的设计与开发，包括多模态特征提取、融合策略设计和纠错模型训练。
- 负责了多模态特征融合模块的开发，实现了文本、语音和视觉信息的有效融合，提升了模型对复杂错误的识别能力。
- 参与了纠错模型的训练和优化工作，包括损失函数设计、超参数调整和模型性能评估。
- 设计并实现了基于规则的后处理模块，对模型输出进行修正，进一步提升了纠错的准确性。

一个端到端的基于 BERT 和 Transformer 的文本纠错框架 2016年11月 - 2020年12月

项目概述：设计一个端到端的框架能够同时利用拼写错误字符的视觉和语音特征，并最大限度地减少它们对上下文的误导性影响从而提升纠错的召回率和准确率。

- 使用轻量级 Transformer 检测拼写错误字符的位置，在校正器中使用 BERT 来捕获原始句子中每个字符的视觉和语音特征，纠正器直接使用原始句子作为输入，以保留拼写错误字符的视觉和语音特征。
- 采用了一种后期融合策略，将校正器的隐藏状态与检测器的隐藏状态融合，这种融合策略旨在消除拼写错误字符的误导性影响。
- 通过端到端的联合训练来实现检测和校正任务的同步执行，该框架同时执行校正和检测任务，并通过联合训练来优化性能。

CNKI 搜索与推荐系统优化 2014年12月 - 2016年8月

项目概述：项目采用 Elasticsearch 作为全文检索引擎，并结合 Spark 与 Flink 进行大规模数据处理；同时引入深度学习模型（如 DNN、Transformer）以及知识图谱构建技术，提升系统对学术资源语义理解与跨领域关联推荐的能力。

- 参与基于 Elasticsearch/ELK Stack 的全文搜索服务建设，支持复杂查询和实时聚合分析,探索向量搜索技术（FAISS、Milvus）在推荐场景中的应用
- 主导深度学习推荐模型从早期 Wide&Deep、DeepFM 到后期 DIN、DIEN、序列推荐（BERT4Rec）及多任务学习（MMOE）的研发应用，有效提升了推荐相关性和用户转化率。
- 构建基于 Hadoop、Spark、Flink 的离线批处理平台和 Spark Streaming/Flink/Kafka Streams 的流处理系统，实现海量数据的实时采集、特征提取与计算。
- 主导实现在线学习（FTRL、在线更新框架）和实时推荐服务，通过 Lambda 架构、微服务、容器化部署（Docker、Kubernetes）构建多级推荐架构（召回→粗排→精排→重排）。
- 建立以 Prometheus、Grafana为核心的实时监控系统，持续跟踪模型性能（AUC、NDCG、Precision、Recall、CTR、CVR等指标）。
- 推动实验平台（A/B测试、TensorBoard、MLflow）的建设，实现模型迭代和线上调试自动化管理。

国美电商平台搜索推荐系统 2014年12月 - 2016年8月

项目概述：负责某电商平台的搜索与推荐系统架构设计和优化，提升用户搜索体验与个性化推荐能力。项目采用基于Lucene的Elasticsearch作为搜索引擎，并结合Hadoop生态进行大规模数据处理，同时引入协同过滤和矩阵分解等推荐算法，增强商品推荐的精准度。

- 搜索引擎优化，基于 Elasticsearch 构建商品搜索引擎，优化索引结构、分词策略及搜索召回，提高搜索查询响应速度。
- 推荐系统构建，采用协同过滤（CF）和矩阵分解（MF）算法构建个性化推荐系统，优化用户-商品匹配，提高转化率。
- 分布式计算架构，利用 Hadoop、Spark 进行数据预处理和批量计算，提升大规模数据处理效率，并结合 Redis 加速热门商品推荐。
- 实时推荐优化，引入 Kafka 构建实时数据流处理框架，基于用户行为（点击、收藏、购买）动态调整推荐结果。
- 使用 scikit-learn 与 MLlib 实现基于逻辑回归和随机森林的推荐模型，提高推荐系统的泛化能力。

教育背景

特莱恩大学（the U.S.） 2024年5月 - 2026年5月

Master in Information System

北京科技大学 2020年9月 - 2023年1月

软件工程 硕士

洛阳师范学院 2009年9月 - 2013年7月

计算机科学与技术 学士

证书与荣誉

中级软件设计师（软考）

高级项目管理师（软考）

大学英语六级，467

校优秀毕业生（大雾），2023年