郑桂能

LinkedIn: www.linkedin.com/in/guineng-zheng-339982a0/
Phone: (+86) 13661585811
Email: zgn7@hotmail.com

专业技能

自然语言处理：

1. 长期聚焦于自然语言处理，全面掌握大语言模型（LLM）全栈技术，具备从 LSTM 到早期 BERT 模型的应用经验，系统掌握 Transformer 架构从 BERT 到 GPT 的演进路线，深入理解 Decoder-Only 架构在生成式任务中的优势与与应用场景。具备从零构建大语言生成模型的实践经验，系统掌握 DeepSeek 技术栈及其实现机制，并能将领域知识深度集成至模型中，构建面向特定场景的大模型系统。

2. 拥有丰富的大模型本地化开发与训练实战经验，具备从零部署大规模 DeepSpeed 集群的能力，使用虚拟化技术搭建超 20 节点的训练环境；通过 MinIO 构建对象存储系统，完全模拟 OpenAI 内部全过程训练场景，完成端到端的大模型训练框架部署。

3. 熟悉大语言模型的部署流程，能够使用 ngrok 构建安全的公网访问通道，实现本地模型的远程推理接口；具备端到端推理服务部署能力，包括模型加载优化、API 接口设计、Web 访问集成，拥有完整的大模型应用工程链路经验。

4. 熟练掌握主流大模型微调技术，包括 LoRA、QLoRA、PEFT 等参数高效微调方法，并熟悉其在低资源条件下的高效适配流程，能够根据任务需求进行快速模型定制，显著提升特定领域的性能与部署效率。

5. 熟练掌握 RAG（Retrieval-Augmented Generation）技术，具备构建高质量向量知识库以增强大语言模型检索能力的实战经验。能够根据特定任务需求，设计并集成高效的检索模块与生成模块，实现大模型在开放领域问答、文档理解、知识增强等场景中的性能提升与响应准确性优化。具备从数据预处理、向量化建库、检索引擎接入，到与 LLM 融合推理的完整工程实现能力。

6. 熟悉链式思维（Chain-of-Thought）与模型自反思（Self-Reflection）等大语言模型推理增强策略，能有效提升模型在长文本理解、复杂问答与知识注入等场景中的稳定性与可靠性。具备将此类策略集成至自动推理系统与评估平台的工程实现能力。

7. 对于大规模非结构化文本的结构化数据抽取与分析具有一线经验，拥有丰富的知识图谱与知识库构建经验，曾服务于亚马逊 Product Graph 业务场景，广泛应用于个性化推荐系统与知识注入任务。深入掌握各种序列标注任务，包括词性标注（POS）、命名实体识别（NER）、句子分割（CHUNKING）等，积累了丰富的实际数据标注与模型训练经验，显著提升了基于知识图谱的下游任务效率与效果。值得指出的是，在当前大语言模型（LLM）时代，图谱模型已作为支撑 RAG 任务的核心知识结构被广泛采用，本人早在 LLM 普及之前即已深耕图谱构建与应用实践，具备扎实的图谱语义建模与系统实现能力，在该领域具有长期积累与专家级的理解深度。

8. 擅长文本结构化数据提取，具备从大规模原始日志中自动抽取模板的技术能力，支撑集群运维系统高效运行。首次提出针对日志模板分析的客观评估体系，并据此系统分析现有主流方法的局限，指出其在工业实践中的不足，明确提出了度量体系与算法革新的发展方向，为该领域奠定基础。

9. 在可解释性机器学习方向具备深入研究，聚焦于增强序列模型的可解释能力。提出可微二值分布方法（differentiable binary variable），并引入一致性稀疏性约束（consistency sparsity constraint），显著提升模型在工业应用中的稳定性与鲁棒性。相关方法已应用于文本模板挖掘与白盒可解释模型部署。

数据库及大数据：

1. 具备丰富的大数据平台构建经验，曾参与阿里巴巴云计算 5K 集群“登月”项目，该项目是当时国内规模最大的大数据处理系统，支撑了阿里金融的大规模风控业务及淘宝中台的运营体系。拥有完整的 SQL 引擎全链路开发经验，包括 ANTLR 语法定义、语法解析、SPJA 算子实现、Hadoop 作业构

建，以及计算任务的集群调度与优化。提出并实现了基于历史任务记录的集群资源优化系统，整体提升集群计算效能超过 35%。

2. 精通大数据应用的全栈开发，具备扎实的大数据业务实现能力，曾自主开发分布式 MapReduce 作业与 Spark 应用，能够结合业务需求完成端到端的数据处理流程。擅长基础算子优化与任务性能调优，曾参与并支持阿里巴巴 B2B 业务在阿里云大数据平台上的全链路建设与系统优化工作。

3. 精通数据仓库设计与 ETL 流程构建，曾主导招商银行信用卡中心大数据基础平台的开发与增强，负责与华为 FusionInsight 团队的深度对接。系统支撑招行内部风控的全链路 ETL 任务，结合客户分群进行深度数据分析，为业务运营提供稳定、可靠的数据服务。针对早期 FusionInsight 平台稳定性不足的问题，提出并落地了底层优化方案，确保关键业务系统的每日稳定运行。

4. 深入掌握大规模实时流式数据处理技术，曾在 NEC 普林斯顿研究院参与流式复杂事件处理（CEP）项目，基于 Spark Streaming 框架构建了稳定高效的分布式 CEP 系统，有效支撑 NEC 内部海量日志与实时流数据的处理需求。

5. 当前重点关注向量数据库在大语言模型（LLM）检索增强生成（RAG）任务中的应用实践，熟悉 FAISS、Milvus 等主流向量检索引擎的索引机制、数据建模与部署流程。具备从传统结构化数据向语义向量空间映射的建模能力，能够构建并优化支持 LLM 检索任务的高性能知识库系统，实现稳定可靠、低延迟的推理后端服务。

计算机视觉：
1. 熟悉计算机视觉领域的主流技术，具备实际的人脸识别项目经验。曾基于业界领先的 InsightFace 框架实现高精度人脸检测与识别系统，并成功部署于实际门禁管理系统中，满足了商业场景下对实时性与稳定性的双重要求。

2. 具备扎实的深度学习人体关键点检测、姿态估计与动作跟踪经验。曾在青岛海尔集团卡奥斯工业研究院主持相关项目，成功将其应用于家电生产线安全监测与员工操作规范监督的预研任务，有效提升了产线的安全管理能力与整体生产效率。

3. 在动作识别领域提出了新型商用级算法，基于 HRNet 骨骼提取网络，结合序列模型构建高效人体动作识别方案。所开发算法在识别精度方面显著优于传统的 DTW 相似度匹配方法，同时计算资源开销远低于现有 ST-GCN 架构，在实际部署中展现出极高的性能与性价比优势，具备大规模落地应用的潜力。

项目经验

Logflux 开源日志解析工具套件: `https://github.com/logflux/logflux`
独立开发的高性能开源工具套件，用于自动化日志解析，旨在推动日志解析领域研究成果的复现与工业落地。项目首次系统性指出日志模板挖掘领域长期缺乏客观、统一评价标准的问题，并重新实现与整合了包括 Drain、IPLoM、Nulog 在内的 16 种经典与前沿日志解析算法，构建了统一、公正的评测平台，用于系统评估各类算法的解析性能与实际适用性。

本项目由本人从零独立设计与开发，涵盖遗传算法、深度学习、聚类、动态规划等多种方法，所有核心代码均由本人 from scratch 编写，体现了本人长期坚持的一线开发实践能力。

其中，Nulog 模块基于 BERT 架构引入预训练语言模型，用于增强日志建模能力，展现了项目在大语言模型（LLM）技术栈方向的兼容性与前瞻性。同时，项目提出了高效的向量化日志匹配机制（基于 NumPy 与 CUDA 实现），显著提升了大规模日志数据处理的效率。

该项目已被多个研究团队及工业用户广泛采纳，显著提升了日志解析的可重复性与实际部署效率，现作为犹他大学 Flux 研究组的核心基础研究工具长期维护。相关成果已发表于国际会议 ACM REP’24（ACM Conference on Reproducibility and Replicability）。

工作经验

Amazon Product Graph 2017. 8 - 2017. 12
秋季实习

· 就职于亚马逊商品知识图谱研究组，提出采用 BIOE 标注策略从非结构化商品文本中提取结构化信息，结合 LSTM、自注意力机制和 CRF 模型，构建面向知识图谱的高质量信息抽取流程。设计并实现主动学习（Active Learning）策略，记录模型训练历史，显著降低人工标注成本并提升模型泛化能力。开发基于 t-SNE 的可视化工具，直观展示注意力权重分布，揭示注意力机制对模型性能的积极影响。项目中的 attention 机制为 BERT 时代的 QKV 自注意力机制奠定技术基础，具有重要的前瞻意义，对 Product Graph 顺利过渡至大模型时代发挥了关键作用。**技术栈：Python, Keras**

NEC Laboratories 2016. 5 - 2016. 8
暑期实习

· 在 NEC 普林斯顿研究院参与流数据处理平台开发，基于 NEC 内部框架构建稳定高效的分布式实时流处理系统，对标 Spark Streaming，支撑 NEC 内部安全审计与实时日志分析平台，支持每秒万级数据流处理，满足关键业务对实时性与系统鲁棒性的高要求。**技术栈：Java**

招商银行信用卡中心 2014. 11 - 2015. 7
大数据开发工程师

· 负责数据仓库 ETL 任务的全流程设计与性能优化，制定基于数据驱动的客户分析与客群细分策略，显著提升数据服务与业务联动效率。深度对接华为 FusionInsight 团队，针对华为初期大数据平台的不稳定性问题，切入华为驻场运维团队的全链路工作，设计并实施底层优化方案，全面保障系统的高可用性与每日运行稳定性。**技术栈：Java**

阿里云技术有限公司 2012.4 - 2014.10
数据平台开发工程师

· 深度参与阿里云 5K 节点大数据平台（ODPS）建设，负责 SQL 引擎核心模块开发及资源调度系统优化。熟练掌握工业级数据库实现细节，包括 ANTLR 文法解析、SPJA 算子重排与 MapReduce 作业生成。设计并实现了基于历史任务日志设计智能资源调度系统，使阿里云 ODPS 整体集群计算资源效率提升超过 35%，稳定支撑阿里金融与淘宝等关键业务的高频跑批需求。**技术栈：C++, Java, Python**

教育背景

犹他大学, 博士, Flux 研究项目组 *2019.3 - 2024.11*
研究方向: 数据挖掘 & 机器学习 & 大规模系统日志分析 & 可解释性机器学习
导师: Vivek Srikumar, Robert Ricci

犹他大学, 硕士, 数据库研究组 *2015.8 - 2018.8*
研究方向: 数据库实现，大数据平台实现
导师: 李飞飞

南京航空航天大学, 学士 & 硕士 *2005.9 - 2012.3*
研究方向: 关系数据库实现
导师: 秦小麟

编程语言

专精技能: Python, Java, C++, SQL

· **Python:** 拥有 10 年以上 Python 开发经验，广泛应用于机器学习、深度学习、自然语言处理、大数据处理与系统部署等方向。熟练掌握 PyTorch、TensorFlow、Scikit-learn、NumPy、Pandas、Matplotlib、Seaborn、NLTK、SpaCy、Transformers、HuggingFace PEFT、FAISS、FastAPI、Flask、SQLAlchemy、

SQLite、OpenAI SDK、LangChain、Milvus SDK、PySpark 与 CUDA 编程等核心工具和库。具备丰富的模型训练、算法设计、向量检索系统搭建及 LLM 推理服务部署经验，代码风格工程化程度高，能胜任高性能计算与工程应用的各类任务。

· **Java:** 大数据领域核心开发语言，精通 Hadoop、Hive、Spark 等生态体系，具备 Hive SQL 调优能力，包括监测任务性能指标并根据收集的数据进行算子重排与执行计划优化。具有丰富的分布式系统开发与优化经验，具备从大数据平台搭建到业务部署的完整链路经验，曾主导多项企业级数据平台优化项目。

· **C++:** 具备生产环境级别的 C++ 开发经验，曾于阿里云大数据平台 ODPS 中使用 C++ 实现基于 SQL 历史记录的资源调度优化系统，有效提升调度效率。具备针对大规模 ETL 中重复任务进行算子复用与调度建模的能力，熟悉高性能数据处理底层实现与调优。

· **SQL:** 精通 SQL 引擎的开发与优化，具备 SQL 解析器与执行引擎的全链路开发经验。擅长大规模 ETL 任务优化与 SQL 执行计划调优，曾在招商银行信用卡中心主导风控业务 SQL 优化与平台性能提升工作。

基础技能: Scala, Javascript

· **Scala:** 具备使用 Scala 语言开发 Spark 大数据分析应用经验，主要应用于数据驱动的 RDD 算子处理。

· **Javascript:** 使用 D3.js 实现 Radial tidy tree 开发数据可视化应用，用于解释 Log 挖掘度量论文中的最优路径问题。

发表论文

Logflux: A Collection Of Automated Log Parsing Tools.
ACM REP '24: Proceedings of the 2nd ACM Conference on Reproducibility and Replicability
Guineng Zheng, Robert Ricci and Vivek Srikumar
Data Mining & System Implementation

Opentag: Open attribute value extraction from product profiles.
ACM SIGKDD '18: International Conference on Knowledge Discovery & Data Mining,
Guineng Zheng, Subhabrata Mukherjee, Dong Xin Luna and Feifei Li
Natural Language Processing

Deeplog: Anomaly detection and diagnosis from system logs through deep learning.
ACM CCS '17: SIGSAC conference on computer and communications security.
Du, Min, Feifei Li, Guineng Zheng, and Vivek Srikumar.
System Security

No Ground-truth But Guidance: Elbow Metric for Log Parsing.
ACM SIGKDD, in review
Guineng Zheng, Vivek Srikumar and Robert Ricci
Natural Language Processing

Rethinking Evaluation for Log Parsing.
ACM TKDE, in review
Guineng Zheng, Vivek Srikumar and Robert Ricci
Natural Language Processing