盛胡丹筠

邮箱:hudanyun.sheng@outlook.com | 电话 13683132915 | 个人网站 https://danniesheng.github.io/

专业技能

- 机器学习与生成式 AI:监督学习(分类、排序、特征工程)、模型评估与误差分析、检索增强生成(RAG)、Prompt 设计、Embedding 检索、交叉编码器重排、多策略回退与不确定性控制
- 数据与系统工程:端到端 ML / GenAI 流水线设计、异构文档解析(PDF 抽取、OCR、表格结构识别)、人机协同审核系统(Human-in-the-loop)、异常处理与鲁棒性设计(编码问题、损坏文件、模型误判)
- 平台与工程能力:Databricks ML 工作流、Docker 容器化部署、Azure OpenAI / OpenAI API、Python、Pandas、scikit-learn、FAISS、Hugging Face

工作经历

强生 (Johnson & Johnson) - 数据科学家 (外包)

中国北京, 2024.04 至今

在合规与数据质量约束严格的环境下, 负责商业与医学场景的 ML / GenAI 系统设计与落地

新药上市 *HCP* 预测与排序模型

- 构建早期处方医生识别模型与医生优先级排序系统
- 基于历史销售与推广数据设计特征工程与评分流水线
- 使用 Kedro 搭建模块化数据管道, 提高数据校验与可复现性

医学内容合规检测(*RAG* 系统)

- 设计基于 RAG 的自动化检测流程, 对比日本本地 FAQ 与药品说明书
- 结合 embedding 检索、规则校验与 LLM 推理构建多策略回退机制
- 通过错误分析优化 prompt 与检索策略, 降低误报率

医生知识检索与问答平台

- 构建多语言医疗 RAG 系统(FAISS + E5 embedding + 交叉编码器重排)
- 实现日中姓名标准化与模糊匹配(NFKC、假名转罗马字、繁简统一)
- 搭建 Streamlit 可视化界面, 支持参数调优与证据溯源

医院排班 *PDF* 解析与审核系统

- 设计鲁棒的排班表抽取流水线, 支持数字 PDF、扫描件、损坏文件等多种格式
- 集成 Camelot、OCR、LLM 图像理解等多策略, 并构建质量门控机制
- 开发审核 UI, 支持批量标注与错误记录, 保障流程稳定性

Zenni Optical - 数据科学家

中国北京, 2023.06 – 2024.01

处方识别系统(*Rx Recognition API*)

- 构建基于 FastAPI 的处方识别服务, 抽取球镜、柱镜、轴位等结构化字段
- 结合 OCR 与规则后处理逻辑, 提升复杂版式识别准确率
- 适配日文处方格式, 支持多区域数据存储与日志分析(CloudSQL / BigQuery)

人机协同质量分析工具

- 开发内部 Streamlit 可视化工具, 用于 OCR 结果检查与问题定位
- 分析 API 调用日志, 诊断图像质量对识别结果的影响

德克萨斯西南医学中心 (UTSW) - 数据科学家

定量生物医学研究中心

美国, 德克萨斯州, 达拉斯, 2021.09 – 2023.05

CyTOF 空间影像分析系统

- 构建整合空间影像与单细胞表达数据的分析平台
- 使用 Flask + Docker 部署用户友好型界面

病理图像核分割与分类

- 基于 Mask R-CNN 实现细胞核分割与 6 类分类(检测率82.5%, 分类准确率82.0%)
- 针对弱标注数据定制损失函数, 恢复约 20% 可用训练样本

Donald Danforth Plant Science Center - 数据科学研究员 美国，密苏里州，圣路易斯 2020.02 – 2021.09

多模态植物图像分析管道开发

- 构建多模态植物表型数据处理与可视化流水线（RGB / 热成像 / 高光谱）
- 开发叶片实例分割与时序生长跟踪系统
- 参与 PlantCV 开源项目开发与社区支持
- 新增分割与特征提取模块，参与文档撰写与教学推广
- 支持跨团队图像分析任务，提升数据可用性与科研可视化质量

个人项目

AI News Agent – 生成式 AI 新闻问答系统

构建轻量级 *RAG* 管道，实现 *AI* 新闻的自动抓取、摘要生成与问答支持

- 实现定时抓取与分类机制，集成本地嵌入模型与 OpenAI LLM 生成每日新闻摘要
- 支持用户在本地语料库基础上进行新闻问答，结合混合检索与语义匹配
- 使用 Streamlit 构建交互式界面，部署于 Hugging Face Space
- 实现 Docker 化部署流程，支持嵌入模型热切换与模块化扩展

教育背景

佛罗里达大学 美国，佛罗里达

电子与计算机工程硕士 2018.01 - 2019.12

硕士论文：基于高光谱图像的柳枝草基因型分类 (Switchgrass Genotype Classification using Hyperspectral Imagery)

佛罗里达大学 美国，佛罗里达

工业系统工程硕士 2016.01 - 2017.12

同济大学 中国，上海

物理学学士 2011.09 - 2015.06

毕业论文：X射线多层膜KB成像的强度均匀性校正

学术研究经历

佛罗里达大学 机器学习及感知实验室 (Machine Learning and Sensing Lab) 美国，佛罗里达州，盖恩斯维尔

机器学习研究助理 2017.05-2019.12

- 基于多实例学习设计图像分割算法，实现微根管图像中植物根系的自动检测，提升数据标注效率
- 搭建高光谱与热成像图像处理流程，包括预处理、解混合与目标检测，应用于植物表型大数据自动化分析
- 研发基于孪生网络的降维算法，增强植物基因型分类的可分性，模型可迁移用于高维医学影像分类
- 利用基于纯正端元提取的高光谱解混合的算法，实现高光谱图像的目标检测和图像分割

出版文章

- Sheng, H., Wang S., et al. “MTIA: An open-source python package for systematic multiplexed tissue image analysis” (in preparation)
- Sheng, H., Gutierrez, J., Schuhl, H., Murphy, K. M., Acosta-Gamboa, L., Gehan, M., & Fahlgren, N. (2023). Increasing the Throughput of Annotation Tasks Across Scales of Plant Phenotyping Experiments. Authorea Preprints.
- Rong, R., Sheng, H., Jin, K.W., Wu, F., Luo, D., Wen, Z., Tang, C., Yang, D.M., Jia, L., Amgad, M. and Cooper, L.A., 2023. A deep learning approach for histology-based nucleus segmentation and tumor microenvironment characterization. Modern Pathology, 36(8), p.100196.
- Panda, K., Mohanasundaram, B., Gutierrez, J., McLain, L., Castillo, S. E., Sheng, H., ... & Slotkin, R. K. (2023). The plant response to high CO2 levels is heritable and orchestrated by DNA methylation. New Phytologist, 238(6), 2427-2439.
- Yu, G., Zare, A., Sheng, H., Matamala, R., Reyes-Cabrera, J., Fritschi, F.B. and Juenger, T.E., 2020. Root identification in minirhizotron imagery with multiple instance learning. Machine Vision and Applications, 31, pp.1-13.