BRIX LABS

lidian

教育背景

重庆大学，信息与通信工程，工学硕士　　2016.09 – 2019.06

科研方向：深度学习、信号处理、图像处理

西南大学，通信工程，工学学士　　2011.09 – 2015.06

工作经历

深信服科技南京研发中心　｜ 高级算法工程师（AI+安全）　　2022-5 – 至今

百度上海研发中心　｜ 高级算法工程师（搜索技术部门）　　2020.06 – 2022-5

华为成都研发中心　｜ 开发工程师　　2019.06 – 2020.06

项目经历

【深信服期间】

➢ 用户画像构建 风险行为监测 恶意网页识别

项目背景：**零信任**理念下的企业级员工访问控制系统，以“**数据流量身份化**”和“**动态自适应访问**”为核心，满足不同子场景下 **B** 端客户应用安全访问需求。

主要工作：项目核心成员。

- 完成基于**时序预测、聚类**的行为分析和异常检测算法，模型从个体身份、资产设备、访问载体、目标应用等数据进行特征挖掘，**基线偏离度**作为评价指标捕获风险事件和用户；
- 分析**访问行为**和**应用信息**，提取 Html 网页内容和标签信息，**Bert** 预训练模型进行 fine-tune，完成对潜在的恶意、钓鱼网页的识别。

效果：行为异常事件**召回率提升 15%** 恶意网页**识别准确率 84%->95%**

技术点：特征工程、预训练、**Bert**、**LSTM**、**Xgboost**

➢ 二进制代码 语义表征 相似性分析 离群评估

项目背景：二进制代码分析任务中，深度学习技术展现出明显优于传统方案的效果。通过预训练语言模型，表征通用程序指令，应用于版本间软件比较、恶意软件变体识别、抄袭侵权判定等。

主要工作：项目负责人。

- 数据集搜集和算法调研。基于 **Bert** 构建汇编指令表征模型，围绕程序指令的内部格式、控制流依赖的上下文信息、指令间数据依赖关系进行语料处理和特征挖掘；预训练（MLM+CWP）。
- 效果评估：操作码、操作数离群点检测；汇编语言基本块语义相似性检测。

效果：**precision：0.85，auc：0.9**

技术点：特征工程、**Bert**、预训练、逆向分析

- 终端安全 代码表征 Java 内存木马检测

项目背景：**无文件攻击**能绕过传统查杀软件，隐蔽性强、危害性大。**Java 内存木马**作为典型，希望解决**传统方案检出效率低、泛化能力差、特征规则容易被绕过**等问题。

主要工作：项目负责人。

- 设计基于**机器学习**技术的 Java 内存木马智能检测方案，模型围绕 Java Class 结构信息、Web 中间件属性、安全场景经验属性、综合统计属性等进行**特征挖掘**和**离线验证**。
- 编码开发，完成全链路工具在引擎侧集成上线；作为运营平台负责人进行 case 分析和效果迭代；作为**第一发明人申请专利**：《一种基于内存扫描和 AI 智能分析的 Java 内存马检测方法》。

效果：引擎侧**检出率 60%->96%**

技术点：特征工程、**Xgboost**、**Lightgbm**、**Web** 中间件

【百度期间】

- 语义向量索引构建（向量检索）：计算向量相似度，作为垂搜基础工具，提供技术支持
  - 基于海量语料，针对 Ernie 模型编码后的语义向量，以 MongoDB 进行存储，利用 Faiss 构建索引，以 bRPC 向外提供语义检索接口，可实现 TopK 向量快速查询
- Query 改写：帮助用户初步纠错，使 Query 表达更明确，提升下游检索精准度
  - 构建标准 Query 库（头部 PV、线上运营、规则、命名实体、热词等），针对 SimBERT 编码后的语义向量，利用 Faiss 构建离线索引。在线基于 SimBERT 对新来 Query 进行编码并检索获取 Top-1，引入 Lru-Cache 减少高频 Query 重复推理。
- 相关性策略构建：改善垂类应用搜索场景下，item 结果的相关性打分机制，提升用户体验
  - 针对百家号、好看视频等热门应用子场景检索准确率提升需求，从样本数据文本(title&content)特征和属性特征出发，利用 bm25、TF-IDF、ctr-cqr、term_offset 等算法对正排信息进行处理，获得不同特征下的相关性分数结果，检索准确率评估由 90%提升至 95%
- 排序性能优化：解决多路 BS 服务召回后，精排前置阶段处理的耗时成本和系统开销问题
  - 作为项目负责人完成整体设计和开发。首先通过新增控制逻辑自动解析和同步粗排侧回传的排序信号，并采用节点化改造的堆排序适配，对归并的时间复杂度进行优化（$\alpha O(m*n)+\beta O(mn(log(mn)))=>\alpha O(n)+\beta O(n(log(n)))$），最后完善数据反序列化逻辑，进一步避免冗余性能开销

专业技能

编程语言：C++ Python

框架工具：Pytorch、Pandas、Numpy、sklearn

模型工具：Transformer、Bert、LLaMA、BLIP-2、Xgboost、Lightgbm

算法业务：搜索及推荐算法、机器学习、自然语言处理、大语言模型

获奖经历

- 2022 年第二届深信服 AIFirst 算法挑战赛一等奖（冠军）
- 2021 年百度移动生态产品线(MEG)创新项目奖
- 2019 年华为技术有限公司网络产品线算法编程大赛 16 强
- 2019 年第五届华为软件精英挑战赛成渝赛区一等奖（季军）