

智能汽车信息物理系统知识图谱模块化设计

丁 梓 温 浩 张 晗
中国科学院大学 北京 101400

摘 要：我国汽车保有量超3.66亿辆，购车、用车、养车咨询需求激增，但现有客服模式存在知识获取效率低、交互门槛高、成本高三大痛点。本文设计并实现一套基于知识图谱与多模态技术的汽车智能问答系统，支持本地化部署。系统整合多源异构知识，构建含22460节点、43180关系的细粒度图谱；ViT-Large车型识别准确率92.85%；LoRA微调DeepSeek-R1:8B仅训练0.27%参数量；GraphRAG召回率从65.2%提升至92.4%。基于Django与FastAPI开发，端到端响应 ≤ 1.78 秒，问答准确率超92%，可有效替代基础人工咨询。

关键词：RAG；大语言模型；ViT识别；知识图谱；汽车问答系统

引言

当前汽车服务主要依赖人工或纯文本系统，存在查询繁琐、车型信息难以输入、中小车企部署困难等问题，且现有智能问答系统对微观实体关系、多模态协同和本地化支持不足。本文结合知识图谱与ViT网络，整合手册与配置表，构建细粒度图谱，并引入视觉识别作为检索约束，采用自适应GraphRAG与轻量化大模型，实现低成本本地部署。创新点包括：专用细粒度图谱支持高效多跳查询；视觉识别与检索深度融合；全链路GraphRAG优化缓解幻觉。系统验证有效。

1 系统总体架构设计

本系统采用四层分层架构，自下而上为数据库层、算法层、业务层与前端UI层，各层通过标准化接口交互，实现低耦合高内聚。

1.1 数据库层

采用多模协同存储：MySQL存储用户账户、会话信息、问答历史及操作日志等结构化数据；Neo4j图数据库持久化存储汽车客服场景专用知识图谱，依托原生图存储与免索引邻接特性，显著提升多跳查询与子图匹配效率；FAISS向量库基于BGE-M3嵌入模型对非结构化文本进行稠密向量化，构建HNSW索引支撑语义检索^[1]。三类数据库通过统一数据接入接口屏蔽底层差异，为上层提供一致数据视图。

1.2 算法层

ViT-Large图像识别模块对用户上传的车辆图像进行细粒度车型分类，输出品牌、车系、年份及置信度，为检索提供精确实体约束；GraphRAG检索增强模块深度融合向量语义检索与图谱结构化查询，内置查询规范化、自适应路由、图感知重排与Cross-Encoder精排等全流程优化；LLM生成模块采用经LoRA领域适配的DeepSeek-

R1:8B模型生成专业合规回答。

1.3 业务层及前端UI层

业务层基于FastAPI搭建RESTful接口，封装用户认证、用户管理、知识库交互、多模态问答、历史记录管理等服务。前端UI层基于Vue构建响应式单页应用，使用Echarts实现数据可视化。全链路权限管理构建两级权限体系，普通用户仅访问个人数据，管理员具备全部运维权限。系统支持完全本地化部署，GPU承担ViT推理与大模型生成，CPU处理业务逻辑与数据流转。

2 核心算法设计与实现

2.1 ViT-Large车型识别算法

2.1.1 模型架构与训练策略

本系统采用VisionTransformerLarge (ViT-L/16) 作为车型识别的骨干网络，并使用ImageNet-21k预训练权重进行初始化。所有输入图像统一处理为384×384像素，在训练阶段采用随机水平翻转、色彩抖动及随机擦除等数据增强策略以提升泛化能力。为平衡迁移学习效果与过拟合风险，模型冻结了主干网络前12层Transformer编码器，仅微调后12层及分类头^[2]。损失函数选用交叉熵损失，优化器采用AdamW，初始学习率设为 $2e-5$ ，配合余弦退火调度与梯度裁剪进行优化。模型在NVIDIA RTX 4090上以批量大小32训练60个epoch，第47轮触发早停。推理阶段将模型导出为ONNX格式并使用ONNX Runtime加速，单图平均处理时间控制在185毫秒以内。

2.1.2 数据集与实验结果

训练数据集从公开来源经严格筛选，剔除了模糊、严重遮挡及非整车视角的样本，最终形成包含17286张高质量图像的专用数据集，涵盖丰田旗下34个车型，每车型平均500张图像并覆盖多视角场景。所有图像由专业人员精细化标注了品牌、车系及年份信息。实验结果

表明, ViT-Large的Top-1准确率达到92.85%, 宏平均F1分数为0.8975, 在训练样本少于200张的8个冷门车型上仍取得88.6%的准确率, 显著优于ResNet-50的86.40%和EfficientNet-B3的85.04%。单样本识别对比进一步显示, ViT-Large对主流车型、非典型视角及冷门车型三类场景均能准确识别, 而CNN模型在年份区分和相似车型鉴别上存在明显的错配现象。

2.2 GraphRAG检索增强优化

2.2.1 查询规范化与自适应路由

针对传统RAG结构化推理能力不足的问题, 本文提出了增强版GraphRAG框架。该框架首先利用DeepSeek-R1:8B将用户查询重写为规范化指令, 并将ViT识别出的车型信息自动融入作为检索的硬性约束条件^[3]。在此基础上, 设计了一个基于Sentence-BERT加MLP的轻量级路由分类器(参数量小于1M), 该分类器根据查询语义动态判断问题类型: 对于结构化强关联类问题, 触发Neo4j多跳查询; 对于语义开放类问题, 执行FAISS向量检索; 对于复合类问题, 则启用双路并行召回策略, 从而实现检索路径的自适应选择。

2.2.2 多路召回融合与精排

对于多路召回的结果, 系统采用联合评分机制进行融合。FAISS提供语义相似度得分 s_{vec} , Neo4j基于节点关联强度与路径长度提供结构合理性得分 s_{graph} , 最终综合得分通过公式 $s = \alpha \cdot s_{vec} + (1 - \alpha) \cdot s_{graph}$ 计算得出, 其中 α 经网格搜索确定为0.65。经过图感知重排后, 选取Top-10候选输入Cross-Encoder进行精排, 最终筛选出Top5上下文输入大模型生成回答。实验表明, 经过上述优化后, Recall@5从65.2%提升至92.4%, RAGAs各维度评估指标均达到0.88以上。以保养类标准问题为基准测试, 增强版GraphRAG能够完整复现权威答案中关于更换周期、机油规格、加注量等全部事实性信息。

2.3 DeepSeek-R1:8BLoRA领域适配

本系统采用LoRA方法对DeepSeek-R1:8B进行参数高效微调。该模型基于Llama架构优化, 原生支持64K上下文窗口, 非常适合处理长文本手册片段。微调时, 在所有注意力层的Query和Value投影矩阵中加入低秩适配器, 设定秩 $r = 64$ 、缩放因子128、Dropout率0.1, 由此可训练的参数量仅为21.3M, 占模型总量的0.27%。训练基于HuggingFaceTransformers与PEFT库, 使用2张NVIDIA A10GPU, 单设备批量大小为8并配合2步梯度累积(等效批量32), 采用AdamW优化器, 学习率 $3e-4$, 经过3个epoch后收敛。微调后的模型在专业术语准确性、操作步骤完整性和安全合规性方面显著优于原始模型,

同时较好地保留通用对话能力。推理阶段使用vLLM配合PagedAttention加速, 可高效处理长达8192tokens的上下文输入。

2.4 双链路检索与知识图谱构建

系统以34款目标车型的官方手册为数据源, 使用pdfplumber和PyMuPDF提取文本内容, 扫描页面则通过PaddleOCR进行识别。清洗数据后, 并行构建两种知识表示: 一方面, 将28742个语义片段经BGE-M3嵌入为1024维向量, 构建FAISSHNSW索引; 另一方面, 通过提示工程引导大语言模型抽取三元组, 经规则校验与人工审核(准确率超过93%), 按照预定义Schema导入Neo4j, 最终形成包含22460个节点、43180条关系的领域知识图谱。知识校验采用自动化全量校验辅以10%人工抽样的方式^[4]。在双链路检索中, FAISS负责语义泛化, Neo4j负责结构化推理, 二者通过路由分类器协同工作, 多跳查询的平均耗时被控制在200毫秒以内。

3 系统功能模块实现

3.1 用户登录与注册子系统

该子系统实现了账号注册、登录、密码重置及个人信息管理等基础功能。注册环节对用户名、邮箱及密码进行实时格式校验与唯一性检查, 密码均加密存储; 登录环节支持用户名或邮箱两种方式, 提供"记住我"自动填充功能, 连续失败登录会触发临时锁定机制以保障账户安全; 系统采用JWT令牌维护无状态会话, 确保前后端交互的身份认证安全可靠。

3.2 多模态智能问答子系统

多模态智能问答子系统采用左右分栏布局设计。左侧识别区支持jpg、png、webp等常见格式(文件大小不超过5MB)的图片上传, 调用ViT-Large推理后实时回传并显示品牌、车型、年份、置信度及处理时间。右侧问答区采用气泡式对话布局, 支持自动填充识别出的车型信息作为检索约束, 经GraphRAG检索与LoRA微调大模型协同处理后生成专业回答。全量问答数据均关联用户ID并持久化至MySQL, 用户可随时查询和删除历史记录, 形成完整的交互闭环。

3.3 知识库交互子系统

该子系统基于Neo4j实现知识图谱的可视化展示, 采用力导向图布局, 以不同颜色区分车型、配件、保养项等十余种实体类型。用户可通过关键词检索与类型筛选快速定位目标节点, 点击任一节点即可高亮其一阶及二阶关联关系, 并展示节点的完整属性信息。同时, 系统支持缩放、平移、拖拽等图形化交互操作, 便于非技术用户直观理解汽车知识之间的复杂关联。

3.4 后台管理子系统

后台管理子系统提供了全面的运维支撑能力,包括用户全生命周期管理(增删改查及权限分配)、全流程操作日志审计以及运营数据统计与可视化展示。系统能够实时统计日活用户、问答总量、满意度等核心运营指标,并支持将数据报表导出为Excel文件。所有运维管理操作均全程留痕,确保系统运营的合规性与可追溯性。

4 系统测试与性能评估

4.1 功能性测试

功能性测试围绕四大子系统设计31项用例,覆盖正常流程与异常场景。纯文本问答、车型识别、图文结合问答、图谱可视化、后台管理等核心功能均正常运行,针对模糊图片无明确提示的问题完成优化。整体通过率96.77%,核心业务流程完整闭环。

4.2 非功能性测试

非功能性测试从六方面验证:可靠性方面支持7×24小时稳定运行;性能方面端到端平均响应时间 ≤ 1.78 秒(ViT推理185ms、检索410ms、生成1185ms);安全性方面两级权限与JWT有效拦截越权访问;兼容性方面Windows/macOS下Chrome、Edge、Firefox均正常;稳定性方面限流防护可应对高频请求;易用性方面非技术用户无需培训即可操作。所有指标符合设计要求。

5 局限与未来展望

当前系统知识库主要集中在丰田品牌燃油车型,未来将扩展至主流合资、自主与新势力品牌,覆盖纯电、混动、增程等动力类型,补充电池管理、智能驾驶等知识,建立动态更新机制。多模态方面将构建仪表盘告警灯、OBD故障码等数据集,融合目标检测与OCR实现故障图像解析,引入语音交互支持驾驶场景。架构方面

采用微服务重构,引入分布式缓存与消息队列,探索边缘计算部署。算法方面研究混合检索与模型量化蒸馏技术。商业化方面开发多租户架构,提供定制化知识库构建与微调服务。本研究为中小车企提供了低成本、可落地的智能客服技术路径,在细粒度知识图谱构建、多模态协同及轻量化部署方面的探索,为垂直领域智能问答系统开发提供了实践参考。

6 结语

本研究针对汽车领域智能问答落地需求,设计了一套融合知识图谱、ViT视觉识别与GraphRAG的多模态架构。系统构建含2.2万节点的知识图谱,结合ViT-Large车型识别(准确率92.85%)及LoRA微调的DeepSeek-R1:8B模型,通过自适应GraphRAG将召回率提升至92.4%,实现端到端响应 ≤ 1.78 秒、准确率超92%的性能。基于Django与FastAPI开发的四大模块经测试验证了工程可行性。该方案为中小车企提供低成本、可本地化的智能客服解决路径,其细粒度建模、多模态协同及轻量化适配技术,亦可为装备维修等垂直领域提供参考。

参考文献

- [1]徐春,孙恩威,汪晓洁.基于知识和数据双驱动的DRG医疗问答研究[J].计算机工程与科学,2025,47(6):1121-1132.
- [2]李涓子,侯磊.知识图谱研究综述[J].计算机学报,2022,45(1):1-25.
- [3]孙茂松,郭志芄,刘知远,等.大规模语言模型研究进展与挑战[J].中国科学:信息科学,2023,53(6):1031-1056.
- [4]李克强,张新钰,王建强,等.智能网联汽车信息物理系统关键技术[J].汽车工程,2021,43(1):1-12.