

이미지를 포함한 비정형 문서의 이해와 응답을 위한 지능형 시스템 설계 및 구현

하정헌¹, 유현창²

¹고려대학교 SW·AI 융합대학원 석사과정

²고려대학교 정보대학 컴퓨터학과 교수

habre@korea.ac.kr, yuhc@korea.ac.kr

Design and Implementation of an Intelligent System for Understanding and Responding to Unstructured Documents with Embedded Images

Junghun Ha¹, Heonchang Yu²

¹Dept. of Big Data Convergence, Graduate School of SW·AI Convergence, Korea University

²Dept. of Computer Science & Engineering, Korea University

요 약

본 연구는 이미지가 포함된 비정형 PDF 문서로부터 유의미한 정보를 추출하고, 사용자 질의에 대해 텍스트와 이미지 기반 응답을 생성하는 멀티모달 Retrieval-Augmented Generation (RAG) 시스템을 제안한다. 본 시스템은 OCR 기반 텍스트, 이미지 추출, 이미지 캡션 생성, 멀티모달 임베딩, 하이브리드 검색을 통합하여 설계 되었으며, Reciprocal Rank Fusion 과 Cross-Encoder 재랭킹을 통해 검색 품질을 향상시켰다. 실험 결과, 기존 단일 모달 방식보다 더 높은 응답 품질과 사용자 만족도를 보였으며, 답변의 출처를 제공함으로써 응답의 신뢰성과 투명성을 강화하였다. 본 연구는 멀티모달 AI 및 비정형 문서 처리 기술의 실용적 확장 가능성을 제시한다.

1. 서론

비정형 문서에서 정보를 추출하고 활용하는 기술은 최근 인공지능(AI) 연구와 산업 응용에서 중요한 주제로 부상하고 있다. 비정형 문서는 텍스트, 이미지, 표와 같은 다양한 데이터 형식을 포함하며, 이들의 구조적 복잡성과 비 일관성은 분석과 처리에 있어 큰 도전 과제를 제시한다[1]. 특히, PDF 와 같은 비정형 문서는 기업과 기관에서 빈번히 사용되며, 해당 문서로부터 유의미한 정보를 자동으로 추출하는 것은 업무 효율성을 크게 향상시킬 수 있다.

최근 Retrieval-Augmented Generation 기술은 대규모 언어 모델(LLM)의 한계를 보완하며, 외부 데이터베이스에서 동적으로 정보를 검색하여 보다 정확하고 맥락에 맞는 응답을 생성하는 데 성공적으로 활용되고 있다[2]. 이러한 RAG 기술이 텍스트 중심에서 멀티모달(multimodal)로 확장되면서, 텍스트와 이미지를 포함한 문서의 이해와 응답 생성이 가능해졌다. 멀티모달 RAG 는 텍스트와 이미지 간의 의미적 정렬(cross-modal alignment) 및 통합된 정보 처리를 통해 기존의 단일 모달 접근법 대비 더 풍부한 정보를

제공할 수 있는 잠재력을 가지고 있다[2].

본 연구에서는 이미지를 포함한 비정형 문서를 효과적으로 이해하고 사용자 질의에 적합한 응답을 생성하는 지능형 시스템을 설계한다. 이를 위해 PDF 문서 내 텍스트와 이미지를 분리 및 처리하고, 이를 벡터화 하여 검색 가능한 데이터베이스를 구축하며, 멀티모달 LLM 을 활용해 사용자 질의에 대한 텍스트 및 이미지 기반 응답을 생성한다. 이러한 시스템은 정보 검색 및 생성 과정에서 멀티모달 데이터를 통합적으로 활용함으로써 기존 시스템 대비 높은 정확도와 사용자 만족도를 제공할 것으로 기대된다.

2. 관련 연구

2.1. Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation 은 대규모 언어 모델의 한계를 보완하기 위해 정보 검색과 생성 과정을 결합한 기술로, 외부 데이터베이스에서 동적으로 관련 정보를 검색하여 응답의 정확성과 신뢰성을 높이는 데 성공적이다. 기존 연구에서는 RAG 의 개념을 정리하며, 이를 AI 생성 콘텐츠의 품질을 개선하는

데 활용할 수 있음을 보여주었다[3]. 특히, RAG 는 텍스트 기반 질의응답 시스템에서 높은 성능을 기록하며, 다양한 응용 분야에서 활용되고 있다.

2.2. 멀티모달: 텍스트와 이미지 통합 기술

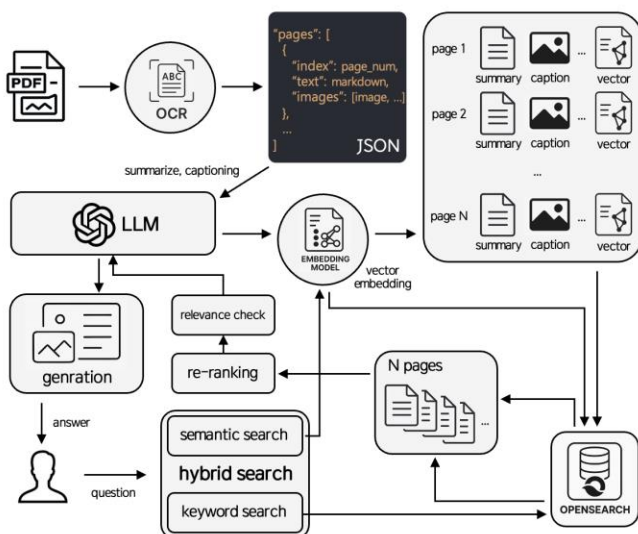
멀티모달 RAG 는 텍스트와 이미지 간의 의미적 정렬을 통해 정보를 통합적으로 처리하며, 기존 단일 모달 접근법 대비 우수한 성능을 보여준다. 기존 연구에서는 산업 응용 분야에서 멀티모달 RAG 를 적용하여 텍스트와 이미지를 함께 처리하면 응답의 품질이 향상될 수 있음을 입증하였다[4]. 다른 연구에서는 PubMed 데이터를 활용하여 알츠하이머병 연구에 특화된 멀티모달 RAG 시스템을 개발하였으며, 도메인 특화된 멀티모달 접근법이 생 의학 연구에서 중요한 역할을 할 수 있음을 보여주었다[5].

2.3. 비정형 문서 처리 기술: OCR, 이미지 분석, 텍스트 추출

PDF 와 같은 비정형 문서 처리 기술은 텍스트와 이미지를 효과적으로 추출하는 데 초점을 맞추고 있다. Optical Character Recognition(OCR) 기술은 문서나 이미지에서 텍스트를 추출하는 데 중요한 역할을 하며, 디지털 PDF 와 스캔 된 PDF 모두에 적용 가능하다. 또한, 이미지 분석 기술은 문서 내 시각적 요소(테이블, 차트 등)를 캡처하거나 의미를 파악하는 데 사용된다. 이러한 기술들은 비정형 문서의 복잡한 구조를 이해하고 데이터를 정리하여 검색 및 응답 생성 과정에서 활용할 수 있도록 한다.

3. 시스템 설계 및 구현

3.1. 시스템 아키텍처



(그림 1) 시스템 아키텍처

본 연구에서 제안하는 시스템은 이미지를 포함한 비정형 문서를 효과적으로 처리하고, 사용자 질의에 대해 텍스트와 이미지 기반 응답을 생성하기 위한 멀티모달 RAG 로 설계되었다. 그림 1 은 전체 시스템의 구성과 데이터의 흐름을 보여주며, 크게 4 단계로 구성된다. 3.2. 시스템 구현에서 해당 단계에 대해 자세히 설명한다.

3.2. 시스템 구현

3.2.1. 데이터 처리

데이터 처리 단계에서는 PDF 문서 내의 텍스트와 이미지를 추출하기 위해 Mistral 기반의 OCR 모델을 활용한 파이프라인을 구축하였다. 추출된 텍스트와 이미지는 각 페이지 단위로 분리되며, 이를 기반으로 본문 텍스트와 이미지 경로를 포함한 마크다운 형식의 JSON 구조가 생성된다. 이 과정에서는 문서의 원본 구조를 유지한 채 본문, 이미지, 이미지 위치 등의 메타데이터를 함께 기록함으로써, 후속 임베딩 및 검색 단계에서의 정합성과 원문 재구성 가능성을 확보한다.

3.2.2. 벡터 임베딩 및 적재

문서는 마크다운 형식으로 변환된 후, 페이지 단위 요약과 이미지 캡션 생성을 통해 임베딩 입력을 구성한다. 텍스트 요약과 캡션은 text-embedding-3-small 모델을 통해 1536 차원 벡터로 변환되며, 동일한 임베딩 공간에 매핑 된다. 각 벡터는 문서 ID, 페이지 번호, 이미지 경로, 캡션 등의 메타데이터와 함께 OpenSearch 에 저장된다. 이와 같은 전체 절차는 Algorithm 1 에 요약되어 있다.

Algorithm 1: Vector Embedding and Storage Procedure

Input: PDF document d

Output: Vectors and metadata stored in OpenSearch

```

1  Convert document  $d$  to Markdown format.
2  FOR each page  $p$  in document  $d$  DO
3    FOR each image  $i$  in page  $p$  DO
4      GENERATE image caption for  $i$  using GPT-4o.
5    END FOR
6  PERFORM summarization on page body and generated image captions.
7  COLLECT page summary and image captions as embedding candidates.
8  FOR each text item  $t$  in the collected set DO
9    COMPUTE 1536-dim vector using text-embedding-3-small model.
10   CONSTRUCT metadata including:
11     { document ID, page number, image path, generated caption }
12   STORE embedding vector and metadata in OpenSearch.
13 END FOR

```

3.2.3. 하이브리드 검색

본 시스템은 시맨틱 검색과 키워드 검색을 결합한

하이브리드 전략을 통해 질의에 대한 정밀 검색을 수행한다. 시맨틱 검색에서 질의는 벡터로 임베딩 되어 코사인 유사도 기반 상위 후보를 추출하고, 키워드 검색 결과와 함께 RRF 방식으로 통합된다. 이후 Cross-Encoder 를 활용해 의미 기반 재랭킹을 수행하며, 상위 5 개 결과를 최종 출력한다. 이 전체 과정은 Algorithm 2에 요약되어 있다.

Algorithm 2: Hybrid Retrieval Procedure

Input: User query q

Output: Top-5 ranked documents for response generation

```

1  EMBED query  $q$  into vector  $v_q$  using text-embedding-3-small.
2  PERFORM semantic search:
    - COMPUTE cosine similarity between  $v_q$  and each document vector in OpenSearch.
    - SELECT top 20 documents  $D_{sem}$  based on similarity scores.
3  PERFORM keyword search with query  $q$  to retrieve top 20 documents  $D_{kw}$ .
4  APPLY Reciprocal Rank Fusion (RRF) to merge  $D_{sem}$  and  $D_{kw}$  into  $D_{merged}$ .
5  FOR each document  $d$  in  $D_{merged}$  DO
6    COMPUTE relevance score between  $q$  and  $d$  using Cross-Encoder.
7  END FOR
8  SORT  $D_{merged}$  by Cross-Encoder scores.
9  RETURN top 5 documents from sorted  $D_{merged}$ .

```

3.2.4. 답변 생성

최종적으로 선정된 검색 결과를 바탕으로, 시스템은 사용자 질의에 대한 응답 생성을 수행한다. 이 과정에서는 프롬프트 엔지니어링 기법을 적용하여, 검색된 문서의 내용을 명시적으로 인용하고 맥락 정보를 반영한 응답이 생성 되도록 한다.

응답 생성에는 OpenSearch 로부터 반환된 메타데이터가 함께 활용되며, 각 응답에는 해당 문서의 출처 정보(문서 ID, 페이지 번호, 텍스트 본문, 포함 이미지 등)가 마크다운 형식으로 첨부된다. 이로써 사용자는 생성된 응답의 신뢰성을 문서 원문을 통해 직접 확인할 수 있으며, 모델의 추론 근거에 대한 투명성과 검증 가능성이 확보된다.

결과적으로, 본 시스템은 단순한 자연어 응답 생성에 그치지 않고, 관련 텍스트 및 이미지 데이터를 포함한 근거 중심 응답을 제공함으로써 사용자 경험의 신뢰성과 해석력을 동시에 향상시킨다.

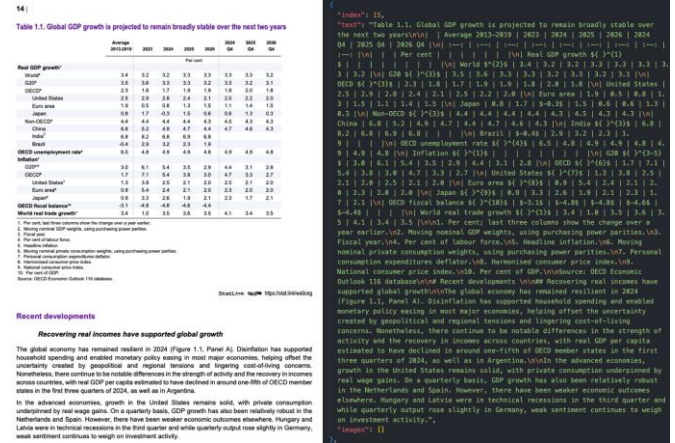
3.3. 실험

3.3.1. 데이터셋

제한하는 시스템의 효과성을 검증하기 위해, OECD에서 발간한 Economic Outlook 2024 보고서를 데이터셋으로 활용하였다. 해당 문서는 268 페이지 분량의 PDF 형식으로 구성되어 있으며, 다양한 본문 서술과 함께 수많은 표, 차트, 그래프 등의 시각 자료를 포

함하고 있다. 주요 내용으로는 경제지표 예측, 국가별 성장률 전망, 정책 제언 등이 포함되어 있으며, 문서 내 다양한 레이아웃, 언어 표현, 시각 요소가 혼합된 복합적 구조를 시스템이 얼마나 효과적으로 처리하고 응답 생성에 활용할 수 있는지를 측정하기 위한 실험적 기반으로 활용되었다.

3.3.2. OCR 실행 결과



(그림 2) 표 데이터 포함 페이지(좌)와 OCR 실행 결과(우)



(그림 3) 차트 포함 페이지(좌)와 마크다운 결과(우)

PDF 문서에 대해 OCR 을 수행한 결과, 페이지 단위로 텍스트와 이미지가 성공적으로 분리, 추출 되었다. 특히 수치 데이터가 포함된 표나 차트와 같은 이미지가 텍스트와 함께 정확히 인식되어, 페이지 마크다운 JSON 구조에 함께 저장되었다. 실험 데이터셋의 14 페이지에 있는 테이블 데이터와 15 페이지의 차트 이미지가 해당 페이지에 있는 텍스트와 함께 성공적으로 추출한 것을 그림 2 와 3 을 통해 보여준다.

3.3.3. 문서 요약에 위한 프롬프트 처리

OCR 결과로부터 추출된 각 페이지 데이터에 대해 이미지 캡션을 위한 프롬프트를 그림 4 와 같이 구성하였다.

이미지 캡션 추출 프롬프트
다음은 경제 보고서에 포함된 차트 이미지와 그에 따른 설명(캡션)입니다. 이 이미지가 전달하고자 하는 핵심 메시지를 요약해 주세요:
1. 차트 제목, 범례, 축 정보 등을 바탕으로 주요 **경제 지표**, **비교 대상 국가 또는 기간**, **숫자 추세**를 파악해 주세요.
2. **캡션**이 포함되어 있을 경우 해당 내용을 요약하고 보완해 주세요.
3. 시각적 내용만으로 알 수 없는 부분은 문맥에서 추론하지 말고 '추가 정보 필요'로 명시해 주세요.
최종 요약은 한 문장 또는 두 문장으로 간결하게 정리해 주세요.

(그림 4) 이미지 캡션 프롬프트 설계

그림 3의 데이터셋 15 페이지와 같이 이미지가 포함된 경우, 그림 4의 프롬프트와 GPT-4o 모델을 통해 그림 5와 같이 시각적 요소에 대한 캡션을 생성하였으며, 이를 함께 요약 정보에 포함시켜 의미 기반 검색의 정확도를 높였다.

이미지 캡션 추출 결과
1. Global GDP Growth (좌측 차트) 지표: 세계 GDP 성장률 (연간 및 분기별), OECD 및 비 OECD 국가 연간 성장 비교 대상: OECD, Non-OECD, 전 세계 (22년 Q1 ~ 24년 Q3) 추세 요약: - 세계 연간 성장률(빨간선): 23년 초부터 3% 내외로 완만한 하락 후 안정 - OECD 국가 성장률(파란선): 2% 아래에서 정체 - 비 OECD 국가 성장률(주황선): 높은 성장률(4~6%)을 유지, 23년 상승세 - 세계 분기별 성장률(녹색 막대): 22년 대비 다소 둔화, 23년 이후 안정
2. Global Output PMI (우측 차트) 지표: 글로벌 제조업 및 서비스업 PMI (구매관리지수) 비교 대상: 제조업 vs. 서비스업 (2022년 1월 ~ 2024년 10월) 추세 요약: - 50이 기준 값 (50 이상: 경기 확장, 50 이하: 경기 수축) - 서비스업(빨간선)은 대체로 50 이상 유지하며 확장세 - 제조업(파란선)은 대부분의 기간 동안 50 이하로 수축세 유지

(그림 5) 그림 3의 차트 이미지 캡션 추출 결과

3.3.4. 실험 결과

1. 차트 해석 질문 - 이미지 판독 능력을 평가
Q) 세계 GDP 성장률은 2022년부터 2024년까지 어떻게 변화하고 있나요?
A) Figure 1.1 Panel A에 따르면, 세계 GDP 연간 성장률은 2022년에는 약 3.4% 수준, 2023년에는 약 **3.2%**로 다소 둔화되었으며, 2024년에도 3.2% 수준으로 유지되고 있습니다. 전체적으로 완만한 둔화 후 안정적인 수준에서 유지되고 있으며, 보고서 본문에서도 "global growth has been stable"이라고 언급하고 있습니다.
2. 차트+텍스트 결합 질문 - 이미지와 텍스트를 함께 이해 하는지 평가
Q) 제조업 활동 둔화가 언급된 국가와 시기는 무엇인가요?
A) 본문에서는 다음과 같이 언급됩니다: "The pace of activity also remains more modest in manufacturing than in services sectors." 그리고 이어서: "Industrial production has continued to stagnate in the advanced economies..." 즉, 제조업 활동 둔화는 선진국들(advanced economies)에서 주로 나타났으며, 최근 몇 분기(2023년 전후) 동안 지속되고 있습니다. Figure 1.1 Panel B에서도 제조업 PMI(빨간선)은 서비스업보다 낮은 수준에서 등락을 반복하며 2023년 중반 이후 50 이하로 수축 국면을 나타내고 있습니다.
3. 이미지 기반 추론 능력 질문 - 정확한 판독과 해석 능력을 평가
Q) 현재 세계 경제에서 서비스업과 제조업의 비중은 어떤 방향으로 변화하고 있다고 볼 수 있나요?
A) 보고서와 그래프(Figure 1.1 Panel B)에 따르면, 서비스업이 제조업을 압도하며 경제 성장을 견인하는 추세로 자리 잡고 있는 추세입니다. - 서비스업 PMI는 2023년 내내 50 이상을 유지하며 확장세 지속. - 반면 제조업 PMI는 여러 시점에서 50 이하로 수축 국면을 보임. 보고서 본문에서도 다음과 같이 명시합니다: "The pace of activity also remains more modest in manufacturing than in services sectors." 따라서, 글로벌 경제의 중심축은 제조업에서 서비스업으로 이동하고 있으며, 이는 정책, 고용, 소비 구조에도 장기적 영향을 줄 수 있습니다.

(그림 6) 유형별 질의에 대한 시스템 답변 결과

응답 생성 단계에서는 검색된 페이지의 원문과 실제 이미지 정보를 기반으로 질의에 대한 자연어 응답과 함께 그림 3의 오른쪽과 같이 원문을 마크다운 형식으로 출력하여 제공한다. 그림 6은 데이터셋 15 페이지에 대해 3가지 유형별 질의에 대한 시스템의 답변 내용이다.

이처럼 논문에서 제안하는 시스템은 텍스트와 이미지와 같은 데이터를 효과적으로 처리하고, 사용자 질의에 대해 신뢰도 높은 응답과 출처 정보를 통합적으로 제공할 수 있음을 입증하였다.

4. 결론

본 연구는 이미지가 포함된 비정형 PDF 문서를 효과적으로 처리하고, 사용자 질의에 대해 텍스트와 이미지를 함께 활용한 응답을 생성하는 멀티모달 RAG 시스템을 제안하였다. OCR, 벡터 임베딩, 하이브리드 검색, 멀티모달 LLM을 통합한 프레임워크를 통해 문서의 복잡한 구조와 멀티모달 정보 간 의미 연관성을 해석할 수 있도록 설계하였다.

Mistral OCR 기반의 데이터 구조화, 시맨틱-키워드 결합형 검색 전략, GPT-4o 기반의 응답 생성 방식을 통해 텍스트와 이미지 정보를 함께 활용한 질의응답이 가능함을 입증하였다. 실험 결과, 본 시스템은 다양한 유형의 질의에 대해 높은 정합성과 문서 출처 정보를 제공함으로써 단일 모달 대비 사용자 만족도를 향상시켰다.

다만 고해상도 이미지 처리 비용 및 복잡한 표, 다이어그램 해석 등은 향후 보완이 필요한 과제로 남아 있다. 본 연구는 비정형 지식 문서의 이해와 활용에 있어 학문적, 산업적 가치가 있는 시도로 기대된다.

참고문헌

- [1] Engelbach et al., "Combining Deep Learning and Reasoning for Address Detection in Unstructured Text Documents", arXiv:2202.03103, 2022.
- [2] Abootorabi et al., "A Comprehensive Survey on Multimodal Retrieval-Augmented Generation", arXiv:2502.08826v2, 2024.
- [3] Zhao et al., "Retrieval-Augmented Generation for AI-Generated Content: A Survey" arXiv:2402.19473, 2024.
- [4] Riedler et al., "Beyond Text: Optimizing RAG with Multimodal Inputs for Industrial Applications", arXiv:2410.21943, 2024.
- [5] Lahiri et al., "AlzheimerRAG: Multimodal Retrieval Augmented Generation for PubMed articles", arXiv:2412.16701, 2024.