흉부 X-ray 진단 및 검색을 위한 Med-LVLMs RAG 구현에 관한 연구

임재영, 윤수연*

국민대학교, *국민대학교 소프트웨어융합대학 소프트웨어학부

dlawowo01@kookmin.ac.kr, *1104py@kookmin.ac.kr

Research on Med-LVLMs RAG implementation for chest X-ray diagnosis and retrieval

Jae Young Im, Soo Yeon Yoon*

Kookmin Univ., *Kookmin Univ.

요 약

본 연구는 흉부 X-ray 영상 기반의 폐질환 진단을 보다 정확하고 설명 가능한 의료 보조 시스템으로, Med-LVLMs 기반 멀티모달 RAG(Retrieval-Augmented Generation) 시스템을 제안한다. 영상 기반 진단에 의미 기반 검색과 응답 생성을 결합함으로써, 의료 전문가의 진단을 보완하고 설명 가능한 AI 기반 의료 지원 시스템의 가능성을 제시하고자 한다. 이를 위해 LLaVA-Med 1.5 모델과 CLIP 기반 Vector DB 검색을 연계하고, MIMIC-CXR 데이터셋을 활용하여 파인튜닝 및 VQA 성능평가를 수행하였다. VQA 성능 평가 결과 RAG를 적용한 4 Epoch 모델은 BLEU는 0.0909, ROUGE-L은 0.2548, BERTScore-F1은 0.3954로 기존 모델 대비 유의미한 성능 향상을 보였다. 특히 1~3 Epoch 구간에서도 높은 문장 정확도와 구조적 일관성을 확보하였으며, RAG 적용 시 외부 지식 기반 검색을 통해 응답 생성 품질이 더욱 향상됨을 확인할 수 있었다. 이는 의료 영역에서 설명 가능한 질의응답 시스템으로의 확장 가능성을 시사하며, 영상·텍스트 통합형 AI의 실질적 임상 활용 가능성을 보여주었고, 이를 통해 의료 인력이 부족한 환경에서도 임상적 의사결정을 지원할 수 있는 실용적인 진단 보조 도구로서의 활용 가능성을 확인하고, 벡터 기반 지식 검색과 응답 생성을 통합한 RAG 시스템이 의료 영상 기반의 다양한 응용 연구와 실증 사례에 폭넓게 활용될 수 있을 것으로 기대된다.

Ⅰ. 서 론

1, 연구배경 및 필요성

기존 흉부 X-ray를 이용한 폐질환 진단은 의료진의 경험과 숙련도에 크게 의존하고 있다. 이러한 문제를 해결하기 위해 딥러닝 기술을 활용한 폐질환 진단 보조 기술이 연구가 되어왔다. 그러나 CNN 기반 접근법은 (정다현, 2023)[1]과 같이 영상에서 병변을 이진 또는 다중 클래스 분류만 수행할 뿐, 판독의 근거를 설명하지 못하는 '블랙박스(Black Box)' 문제가 존재하였다. 이러한 문제점을 해결하기 위해, 최근에는 이미지와 텍스트를 동시에 이해할 수 있는 LMM 기반의 Med-LVLMs(Medical Large Vision Language Models) 연구가 활발하게 진행되고 있다. 그리고 Med-LVLMs의 RAG 시스템을 결합하여 외부 지식 베이스에서 유사한 병변 사례를 검색할 수 있는 진단 보조 시스템 설계 방안도 제안되고 있다[2]. 이에 본 연구에서는 Med-LVLMs와 RAG가 결합된 흉부 X-ray 진단 보조 시스템 구현을 목표로 LLaVA-Med 모델에 MIMIC-CXR 데이터셋을 LoRA 방식으로 Fine-Tuning을 수행하였다. 그리고 CLIP 모델을 활용하여 흉부 X-ray 이미지와 판독 보고서를 벡터화하여 Qdrant Vector DB에 저장하고, LangChain을 통해 질의 시 유사 사례를 검색하고 이를 LMM 입력에 통합함으로써 의미 정합성 높은 진단 응답을 생성할 수 RAG 파이프라인을 구성하였다. 이를 통해 단순 분류를 넘어, 병변 유사도 기반 질의응답과 자동 보고서 생성 등 다양한 임상적 활용 가능성을 제시한다.

Ⅱ. 실험

1, RAG 시스템 개요

본 연구에서는 흉부 X-ray 영상과 진단 텍스트를 기반으로 유사도 검색 및 응답 생성을 수행할 수 있는 Med-LVLMs와 RAG를 결합한 시스템을 설계하였다. [그림 1]은 제안된 시스템의 전체 구조를 나타낸다. 본 시스템은 Retrieval Phase와 Generation Phase의 두 단계로 구성되며, Med-LVLMs 계열 모델인 LLaVA-Med 1.5[3]를 중심으로 구현하고자 하였다.

Retrieval Phase에서는 CLIP 모델을 통해 의료 영상과 진단 보고서를 동일한 임베딩 공간에 매핑하여 Vector DB에 저장한다. 이후 사용자가 입력한 흉부 X-ray 영상도 CLIP 모델을 통해 벡터화되어 Vector DB에서 의미적으로 유사한 Top-K 영상 및 보고서가 검색되고 이 과정을 통해 질의 영상과 임상적으로 유사한 사례들을 효과적으로 수집할 수 있다.

Generation Phase에서는 검색된 Top-K 문서를 원 질문 및 영상과 함께 통합하여 Prompt를 구성하고, 이를 기반으로 LLaVA-Med 1.5에 입력하여 최종 진단 응답을 생성한다. 이와 같은 문서 기반 증강(Augmentation) 기법은 단순 이미지 분석을 넘어서, 외부 지식과 유사 사례에 기반한 설명 가능한 응답 생성을 가능하게 한다.

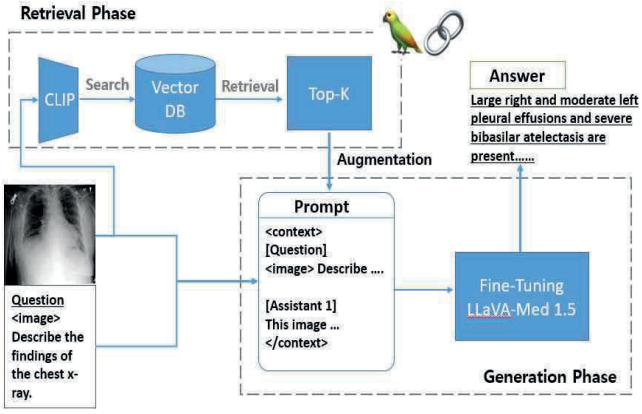

그림 1. 흉부 X-ray 기반 LMM 멀티모달 RAG 시스템 아키텍처

2, 실험 환경 및 데이터셋

본 실험은 고성능 연산 자원이 요구되는 멀티모달 학습을 위해 NVIDIA H100 SXM (80GB) GPU를 활용하였으며, Python 3.10.12 및 PyTorch 2.6.0+ cu118 환경에서 진행되었다. RAG 시스템 구현에는 Qdrant 벡터 데이터베이스와 Langchain 프레임워크를 사용하여 검색 기반 증강 모듈을 구성하였다.

LMM에서 흉부 X-ray에 대한 전문적인 진단 답변을 생성하기 위해 MIMIC-CXR[4] 데이터를 학습 데이터셋으로 활용하였다. MIMIC-CXR은 미국 MIT와 BIDMC가 공동 구축한 대규모 공개 흉부 X-ray 데이터셋으로, 총 65,379명의 환자에 대한 377,110장의 흉부 X-ray 이미지와 227,835건의 방사선 판독 보고서를 포함하고 있다. 또한, 모델 학습 효율성과 멀티모달 입력 구성을 위해 DICOM 영상을 JPEG로 변환한 MIMIC-CXR-JPG[5]와, 판독 보고서를 기반으로 GPT-4로 생성한 질의응답 형식의 JSON 데이터를 포함하는 LLaVA-Rad MIMIC-CXR Annotation 데이터셋을 함께 활용하였다.

3, 실험 및 결과분석

LLaVA-Med 1.5 모델은 Epoch 수에 따라 1부터 4까지의 단계로 Fine-Tuning을 수행하였다. 이후, 4 Epoch까지 학습된 모델을 대상으로 RAG 기법을 적용하여 검색 기반 증강 시스템을 구성하였다. RAG 시스템은 다음의 절차로 구성된다

① Vectorization : 사용자의 흉부 X-ray 질의 이미지를 CLIP 모델을 이용해 벡터화한다.

② 유사도 검색 : 벡터화된 쿼리를 기반으로 Vector DB에서 의미적으로 유사한 Top-K 진단 보고서를 검색한다.

③ 프롬프트 증강 : 검색된 문서를 기존 질의와 함께 프롬프트에 통합(Augmentation)하여 LLaVA-Med 1.5에 입력한다.

④ 응답 생성 : 최종적으로 Med-LVLM이 문맥 보강된 프롬프트를 바탕으로 진단 응답을 생성한다.

실험은 MIMIC-CXR 데이터셋을 1에서 4 Epoch 까지 Fine-Tuning 진행한 모델을 기준으로 VQA 성능 평가를 진행하였다. VQA 성능 평가는 BLEU, ROUGE-L, BERTScore_F1 세 가지 정량 지표를 활용하였다. BLEU는 문장 단위 정확도, ROUGE-L은 문장의 구조적 유사성, BERTScore_F1은 의미적 일치도를 각각 측정한다. [표 1]의 Epoch 별 실험 결과, 1 Epoch에서 BLEU, ROUGE-L이 각 0.0456와 0.2468으로 높은 성능을 보였으며, 문장의 유사도를 기반으로 평가하는 BERTScore_F1에서는 3 Epoch 한 모델이 0.1230으로 제일 높은 성능을 보였다. 그리고 RAG 시스템 성능 평가를 위해서 4 Epoch 모델에 RAG를 적용한 결과는 BLEU, ROUGE-L, BERTScore_F1가 각각 0.0909, 0.2548, 0.3954를 기록하였다.

표 1. 1~4 Epoch + RAG VQA 성능 평가 결과

Epoch + RAG	BLEU	ROUGE-L	BERTScore_F1
1 Epoch	0.0456	0.2468	0.1182
2 Epoch	0.0371	0.2277	0.1130
3 Epoch	0.0352	0.2369	0.1230
4 Epoch	0.0309	0.2229	0.0654
4 Epoch + RAG	**0.0909**	**0.2548**	**0.3954**

Fine-Tuning 초기인 1~3 Epoch 구간에서 세 지표 모두 상대적으로 높은 성능을 기록하였으나, 학습이 계속될수록 성능이 감소하는 경향을 보였다. 이는 과도한 학습으로 인한 일반화 성능 저하와 문장 구조 및 의미 보존 능력 저하를 시사한다.

또한, 4 Epoch 모델에 RAG 시스템을 적용한 경우 RAG 적용 전 대비 성능이 향상되는 것을 확인할 수 있었다. 이는 의료 영상 기반 질의응답의 성능을 높이기 위해 RAG를 통한 외부 지식 기반 검색을 통한 응답 생성 데이터 품질을 실질적으로 높일 수 있음을 확인할 수 있었으며, LMM 기반의 학습 모델 개선에도 도움이 될 것으로 여겨진다. 이를 통해 도메인에 특화된 데이터셋을 기반으로 사전학습 및 파인튜닝을 진행하여 높은 수준의 신뢰성과 정확도를 확보하는데 기여할 것으로 여겨진다.

Ⅱ. 결론

본 연구는 흉부 X-ray 영상 기반의 진단 질문에 대해 의미 기반 검색과 설명형 응답 생성을 통합한 Med-LVLMs 기반 멀티모달 RAG 시스템을 제안하고, 특히 CLIP 기반 Vector DB 검색과 LLaVA-Med 1.5 모델의 Fine-Tuning을 결합하여, 단순 분류를 넘어 실제 임상에서 활용 가능한 질의응답 시스템으로 확장 가능성을 보였다.

MIMIC-CXR 데이터셋을 활용하여 성능 평가를 진행한 결과, RAG 시스템을 적용한 4 Epoch 모델에서 BLEU는 0.0909, ROUGE-L은 0.2548을 나타냈고 BERTScore-F1은 0.3954 의 성능을 보이면서 기존 모델 대비 의미 있는 성능 향상을 기록하였다.

본 연구는 의료 영상 질의응답(VQA)을 위해 RAG 기반의 지식 검색과 생성 통합 구조를 제안하였다. 이러한 구조는 향후 의료뿐 아니라, 법률, 금융 등 설명 가능성과 고신뢰성을 요구하는 도메인 특화 LLM 응용 분야에서 핵심 기술로 작용할 수 있을 것이다. 특히 RAG 기반 LMM 구조는 정보 접근성을 높이고, 응급 상황 대응, 의료 인력 부족 지역 지원, 진료 정보 표준화 등 다양한 실무 환경에서의 적용 가능성을 확인하였다. 영상 기반 데이터와 자연어 질의 간의 의미 정합성 향상을 통해 환자 중심의 디지털 헬스케어 구현에도 기여할 수 있으며, 이는 인공지능이 단순 보조 수준을 넘어 설명 가능성과 정확성을 갖춘 의사결정 지원 시스템으로 발전할 수 있음을 시사한다.

참 고 문 헌

[1] 정다현. "소아 흉부 엑스레이에서의 약지도학습 기반 폐렴 위치 판별 방법." 국내석사학위논문 서강대학교 정보통신대학원, 2024.

[2] XIA, Peng. “Mmed-rag: Versatile multimodal rag system for medical vision language models“. arXiv preprint arXiv:2410.13085, 2024.

[3] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, and J. Gao, "LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day," Advances in Neural Information Processing Systems, vol. 36, pp. 28541-28564, 2023.

[4] A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C. Deng, R. G. Mark, and S. Horng, "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports," Scientific Data, vol. 6, no. 1, p. 317, 2019.

[5] A. E. W. Johnson, T. J. Pollard, N. R. Greenbaum, M. P. Lungren, C. Deng, Y. Peng, Z. Lu, R. G. Mark, S. J. Berkowitz, and S. Horng, "MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs," arXiv preprint arXiv:1901.07042, 2019.