비전-언어 모델과 임베딩 변환기 기반의 의료 영상 이상 탐지 기법

노승서, 김은빈, 심종화, 황인준*
고려대학교 전기전자공학과
{seungseor, gichanac, indexlibrorum3822, ehwang04}@korea.ac.kr

Medical Anomaly Detection Scheme using Vision-Language Model and Embedding Converter

Seungseo Roh, Eunbeen Kim, Jonghwa Shim, Eenjun Hwang*
School of Electrical Engineering, Korea University

요 약

인체의 내부 구조와 상태를 촬영한 의료 영상은 질병 진단과 치료 계획 수립에 중요한 역할을 하지만, 형태와 패턴이 다양하고 복잡하여 진단에 많은 시간이 소요된다. 이에 대한 대안인 의료 이상 탐지 기술은 주어진 의료 영상에서 뇌종양과 같은 이상을 식별하는 것을 목표로 한다. 기존 의료 이상 탐지 방법은 재구축 모델을 활용해 입력 영상의 복원 오차를 측정하고 오차에 따라 이상 여부를 판단했으나, 별도의 임곗값을 설정해야 하며, 식별된 이상 부위를 표현할 수 없었다. 최근 등장한 대형 비전-언어 모델(LVLM)은 자연어 질의를 바탕으로 입력 영상을 이해하고 추론한 내용을 자연어로 대답할 수 있어 의료 이상 탐지를 위한 잠재적인 대안이 될 수 있지만, 일반적인 이미지를 중심으로 사전 학습하여 의료 영상을 이해하는 데 한계가 있다. 이를 해결하기 위해, 본 논문에서는 이미지를 LLM에 입력할 수 있도록 변환하는 임베딩 변환기와 이상 위치 감지를 위한 디코더를 LVLM에 결합하여, 자연어 질의 기반의 이상 판별과 이상 부위 시각화를 동시에 수행하는 의료 이상 탐지 기법을 제안한다. 의료 영상 데이터셋에 대한 비교 실험을 통해 제안된 기법이 기존 재구축 기반 의료 이상 탐지 방법보다 우수한 성능을 보임을 입증한다.

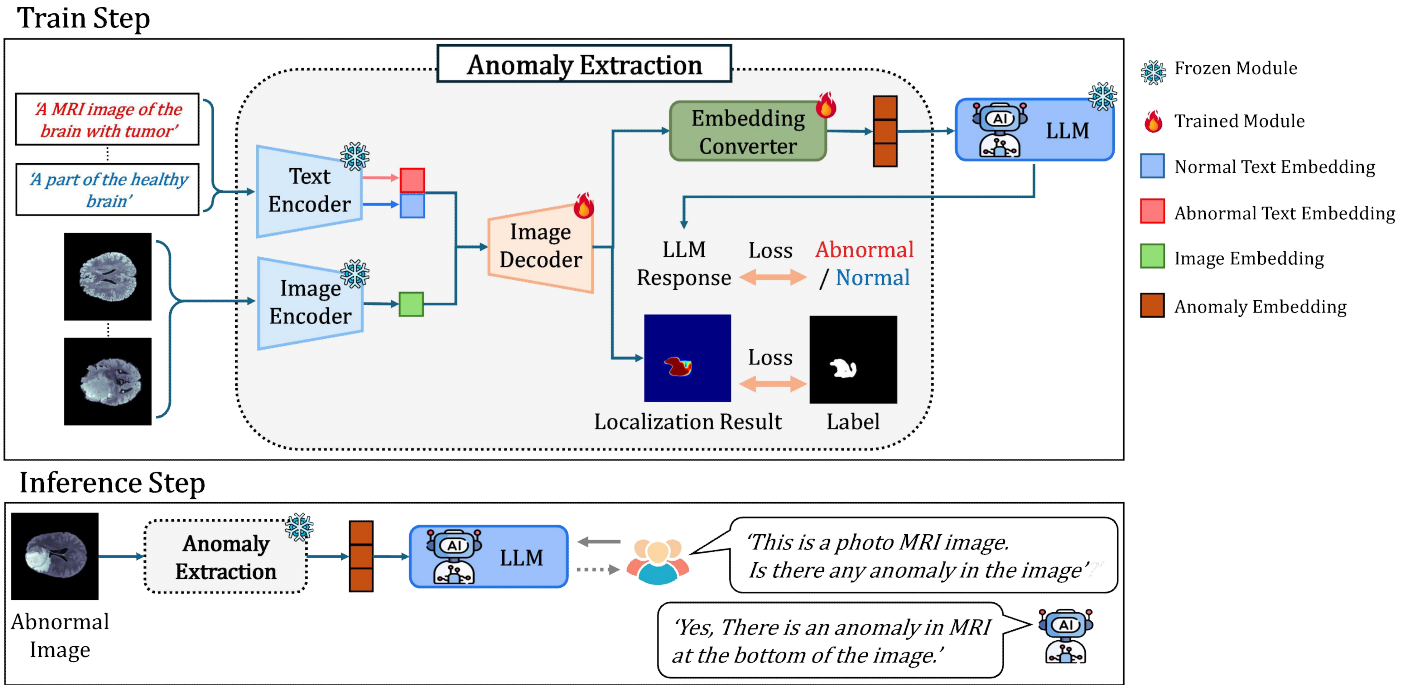
1. 서 론

의료 영상은 의학적 목적을 위해 인체의 내부 구조와 상태를 촬영한 이미지 데이터로, 검진 목표에 따라 뇌, 망막 등 다양한 신체 부위를 나타낸다. 이러한 의료 영상은 의사가 직접 질병을 진단하고 치료 계획을 수립하는 데 중요하기 때문에 MRI, CT, X-ray와 같은 고정밀 촬영 기술이 활용된다. 그러나, 의료 영상은 스캔 지점 및 환자의 신체적 특징에 따라 광범위한 형태와 패턴을 보여 의료 영상을 분석하는 데 많은 시간이 들고 오진할 가능성이 생긴다. 그러므로, 전문 인력의 진료 부담을 줄이고 정확한 진단을 돕기 위해선 효과적인 조기 진단 기술이 필요하다.

이에 대한 대안으로 주목받은 의료 이상 탐지는 주어진 의료 영상에서 뇌종양이나 망막 부종과 같은 이상을 식별하는 기술이다. 기존의 의료 이상 탐지 연구들은 입력 영상을 압축하고 압축된 특징 벡터를 다시 영상으로 복원하는 재구축 모델을 기반으로 이상을 검출한다. 예를 들어, Shvetsova는 정상 클래스의 의료 영상만을 학습한 Autoencoder가 이상을 가진 영상을 잘 복원하지 못하는 특성을 이용해, 복원 오차가 일정 수치를 초과한 영상을 이상 클래스로 판단하였다[1]. 이러한 방법은 학습을 위해 이상 클래스 없이 정상 클래스의 데이터만 활용하는 장점이 있지만, 판별에 필요한 임곗값을 설정하기 위해 사람의 주관이 필요하고, 이상이 식별된 부위를 명시적으로 표현하지 못하는 단점이 있다.

최근, 자연어 질의를 바탕으로 영상에 대한 추론과 질의응답을 할 수 있는 대형 비전-언어 모델(Large Vision-Language Model, LVLM)이 등장하였다. LVLM은 대형 언어 모델(Large Language Model, LLM)에 사전 훈련된 이미지 인코더를 추가하여 자연어와 시각 정보를 함께 이해한다. 예를 들어, PandaGPT는 ImageBind의 이미지-텍스트 인코더들을 사용하여 이미지와 텍스트의 임베딩들을 추출하고, 이를 LLM인 Vicuna에 전달하여 이미지에 대한 질의응답을 수행한다[2]. 이러한 방법은 LLM이 이미지를 직접 해석하고 판단하게 할 수 있어, 의료 이상 탐지를 위한 잠재적인 대안이 된다. 그러나, 대부분의 LVLM은 범용적인 이미지-텍스트 데이터를 중심으로 학습되었기 때문에, 의료 영상의 미세 병변이나 구조적 이상을 이해하는 데 한계를 가진다.

이러한 한계를 극복하기 위해, 본 연구에서는 자연어 질의를 통해 임곗값 없이 의료 영상의 이상 여부를 판단하는 의료 이상 탐지 기법을 제안한다. 제안하는 기법은



(그림 1)제안된 기법의 훈련 및 추론 과정

ImageBind와 LLaMA로 구성된 LVLM을 기반으로 의료 영상에 대한 이상 여부 판별을 수행한다. 또한, 의료 영상에 나타난 이상 영역을 명시적으로 표현하기 위해, 이미지를 LLM에 입력할 수 있도록 변환하는 임베딩 변환기와 이상 영역 지도를 생성하는 디코더를 LVLM에 도입한다. 제안하는 기법의 효과를 검증하기 위해, 본 연구에서는 의료 영상 데이터셋을 기반으로 비교 실험을 수행한다. 실험 결과는 제안 기법이 기존 재구축 기반 의료 영상 이상 탐지 방법에 비해 우수한 성능을 가짐을 입증한다.

본 논문의 구성은 다음과 같다. 2장에서는 본 논문에서 제안하는 기법에 관해 설명한다. 3장에서는 정성적 및 정량적 평가를 통해 제안 기법의 의료 영상 이상 탐지 성능을 제시한다. 마지막으로 4장에서 결론을 제시한다.

2. 본 론

2.1 ImageBind와 LLM

본 연구에서는 ImageBind와 LLaMA를 기반으로 대형 비전-언어 모델을 구축한다. ImageBind의 이미지 인코더는 MRI 의료 영상을 입력받아 시각적 특징을 추출한다. 동시에, 텍스트 인코더는 정상과 이상 상태를 표현하는 "A brain MRI of the brain with structural defect", "a brain MRI of the brain without tumor"와 같은 텍스트를 이용하여 이상 상태를 나타내는 텍스트 임베딩 값과 정상 상태를 나타내는 텍스트 임베딩 값을 둘 다 나타낸다.

의료 영상 이상 탐지 결과를 사람이 이해할 수 있도록 제공하기 위해, LLaMA는 이상 유무를 판별하는 문장과 이상 위치를 설명하는 내용을 자연어로 출력한다. 이는 사전 학습된 언어 지식과 의료 영상에 대한 입력 임베딩을 결합하여 추론을 수행하기 때문에 가능하며, 따라서 별도의 임곗값 설정 없이도 정확한 이상 탐지 결과를 제시한다.

2.2 이미지 디코더와 임베딩 변환기

본 연구에서는 의료 영상에서 식별된 이상 부위를 시각적으로 나타내기 위해, 이상 확률 지도를 생성하는 이미지 디코더를 도입한다. 인코더에서 추출된 이미지 임베딩과 정상, 이상을 나타내는 텍스트 임베딩은 디코더에서 임베딩 간의 유사도를 구하는 과정을 통해 픽셀 단위 분류를 수행한다. 구체적으로, 각 픽셀 위치의 이미지 임베딩과 텍스트 임베딩 간 내적을 계산하여, 해당 픽셀이 정상 혹은 이상일 확률을 추정하는 이진 분류 맵을 생성한다. 이 과정을 통해 텍스트의 의미에 따라 이미지의 시각 정보와의 유사도를 계산하고, 각 위치가 이상 상태에 속할 확률을 나타내는 이상 확률 지도를 생성한다.

LLM이 생성된 이상 확률지도를 이해하기 위해선, 이를 LLM이 이해할 수 있는 임베딩으로 변환하는 과정이 필요하다. 본 연구는 LLM에 입력할 수 있는 임베딩으로 이미지를 변환하는 '임베딩 변환기'를 제안한다. 임베딩 변환기는 컨볼루션 신경망을 통해 이상 확률지도 정보를 LLM 입력 형식에 맞는 확률지도 임베딩으로 변환한다. 또한 LLM의 의료 이상 탐지를 보조하기 위해, 의료 관련 도메인 지식을 인코딩한 의료 지식 임베딩을 포함한다. 이후, 확률지도 임베딩과 의료 지식 임베딩을 동일 차원으로 정렬한다. 마지막으로 두 표현을 하나의 입력 시퀀스로 연결해 LLM에 주입함으로써, 이상 확률지도에 담긴 정보를 자연어 설명으로 매핑 한다. LLM은 고정된 상태로 임베딩 변환기가 출력한 임베딩을 입력받아 의료 영상의 이상 유무 및 위치에 관한 자연어 응답을 생성한다. 정답 텍스트

와 LLM의 응답 간의 차이를 손실함수로 계산하여 임베딩 변환기에만 역전파 한다. 이 과정을 반복함으로써 임베딩 변환기는 이상 확률 지도를 LLM이 적절한 응답을 생성할 수 있도록 정합성 있는 임베딩을 생성하도록 한다.

2.3 의료 영상 이상 탐지

본 문단에서는 제안하는 의료 영상 이상 탐지 기법의 추론 절차를 서술한다. 모델은 입력된 의료 영상을 기반으로 이상 여부 및 위치를 식별하고, 그 결과를 자연어 형태로 출력한다. 입력된 영상은 ImageBind의 이미지 인코더를 통해 시각적 특징을 추출하며, 동시에 정상·이상 상태를 표현한 텍스트는 텍스트 인코더를 통해 임베딩 된다. 디코더는 두 임베딩 간 유사도를 계산하여 픽셀 단위 이상 확률 지도를 생성하고, 이를 임베딩 변환기를 통해 LLM에 입력할 수 있는 형태로 변환한다. LLM은 이를 바탕으로 이상 유무 및 위치 정보를 포함한 자연어 응답을 생성하며, 별도의 임곗값 없이 직관적인 이상 탐지를 가능하게 한다.

3. 실 험

3.1 실험 방법

제안된 기법의 성능을 검증하기 위하여, 우리는 뇌 MRI 영상으로 구성된 BraTs2021[3] 데이터셋에 대해 다양한 실험을 수행하였다. 특히, 우리는 데이터셋의 각 3D MRI 볼륨을 Axial 방향으로 슬라이스한 뒤, 60에서 100번까지의 단면만을 추출하여 2D 이미지로 저장해 사용한다. 훈련-검증-평가 데이터셋은 7:2:1의 비율로 설정했으며, 제안 기법과 비교 모델은 학습을 위해 배치크기는 1, 학습 횟수는 10, 학습률은 0.0001로 설정하였다.

3.2 평가 지표

제안 기법의 이상 탐지 성능을 정량적으로 평가하기 위해 AUC와 Accuracy를 사용한다. AUC는 모델이 각 영상에 대해 이상과 정상 간의 구분을 얼마나 안정적으로 수행하는지를 나타내며, 민감도와 특이도를 종합적으로 반영하는 지표이다. 0과 1 사이의 값을 가지며, 이 값이 1에 가까울수록, 이상과 정상 구분을 안정적으로 수행함 의미한다. Accuracy는 전체 예측 중 실제 정답과 일치하게 분류한 비율을 의미하는 지표로, 분류 모델의 정확도를 평가하는 데 사용된다. 마찬가지로 0과 1 사이의 값을 가지며, 값이 높을 수록 모델의 전반적인 분류 성능이 우수함을 뜻한다.

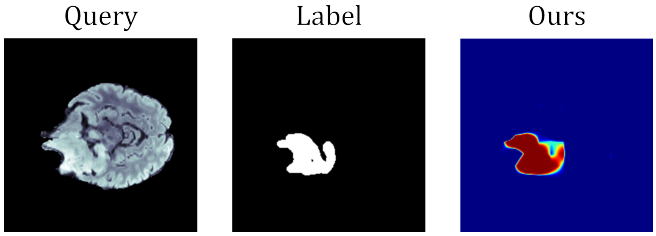
3.3 실험 결과

제안 기법의 성능 비교 대상으로, 대표적인 재구축 기반 이상 탐지 방법인 Autoencoder 계열 모델들을 선정하였다[1, 4, 5]. 해당 모델들은 이상 여부를 판단하기 위해 임곗값을 요구하므로, 검증 셋에서 최대 recall 값을 보이는 임곗값을 선택했다.

<표 1> 제안 기법의 성능 비교

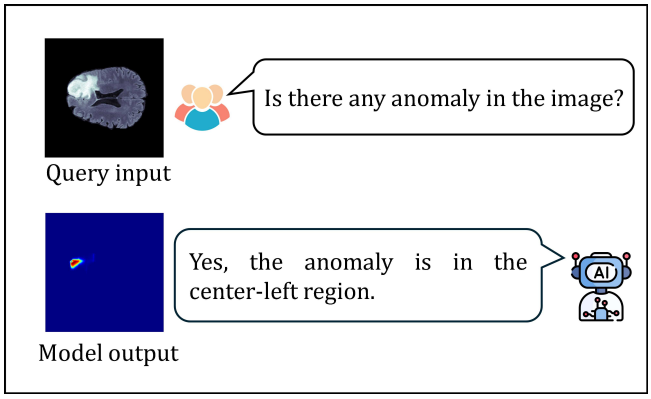
	AUC	Accuracy
Autoencoder	49.98	55.00
Masked Autoencoder	55.59	41.01
Memory Autoencoder	60.57	82.93
Proposed Scheme	**89.07**	**85.90**

표 1은 AUC와 Accuracy 측면에서 Autoencoder 계열 모델들과 제안 기법의 이상 탐지 성능을 비교한 결과를 나타낸다. 동일한 훈련 조건에서 제안 기법은 Autoencoder 대비 AUC는 약 39.1%, Accuracy는 약 30.9% 높은 성능을 기록하였다. Masked Autoencoder[4]는 AUC 55.59, Accuracy 41.01로 Autoencoder 대비 AUC 는 소폭 개선되었으나 Accuracy는 낮게 나타났다. Memory Autoencoder [5]는 AUC 60.57, Accuracy 82.93으로 상대적으로 높은 정확도를 기록하였다. 이는 복원 오차 기반의 Autoencoder 계열 모델들과 달리, 제안 기법이 이미지와 텍스트 임베딩의 의미적 유사도와 LLM 기반 자연어 추론을 활용해 이상 유무 및 위치를 직접 판별함으로써 다양한 이상 패턴에 효과적으로 대응한 결과이다.



(그림 2) 제안 기법의 이상탐지 응답 예시

그림 2는 왼쪽부터 입력된 MRI 영상, 정답 이상 마스크, 제안 기법이 생성한 이상 확률 지도를 순서대로 보여준다. 이상 확률 지도에서 붉은색은 이상 가능성이 큰 영역을, 파란색은 정상으로 판단되는 영역을 의미한다. 제안 기법은 입력 영상에서 이상이 있는 부위를 효과적으로 구분하며, 정답 마스크에 가까운 이상 확률 지도를 생성함을 알 수 있다.



(그림 3) 제안 기법의 이상 탐지 응답 예시

그림 3은 제안한 기법이 사용자의 질의에 응답하는 과정의 예시를 나타낸다. 상단의 Query input은 실제 뇌 MRI 영상이며, 사용자는 일반적인 질의를 입력한다. 이에 대한 하단의 Model output은 이상 유무와 해당 위치 정보

를 포함한 자연어 응답을 포함한다. 그림에서 하단에 시각화된 이상 확률 지도와 모델이 응답한 위치가 일치함을 볼 수 있다.

4. 결 론

본 연구는 비전-언어 모델과 임베딩 변환기를 결합하여, 의료 영상 내 이상 여부 및 위치를 자연어로 직관적으로 표현할 수 있는 이상 탐지 기법을 제안하였다. 제안된 모델은 이상 확률 지도를 기반으로 LLM이 이해할 수 있는 임베딩을 생성하고, 이를 통해 픽셀 수준의 정보를 반영한 자연어 응답을 생성할 수 있다. BraTS2021 데이터셋을 기반으로 수행한 실험에서는 AUC와 Accuracy 지표 모두에서 기존 재구축 모델 대비 우수한 성능을 보였다. 또한, 수동 임곗값 설정 없이 이상 여부를 직접 식별할 수 있는 구조를 통해 모델의 실용성과 확장 가능성을 입증한다. 향후 연구에서는 다양한 의료 영상 모달리티와 병변 유형에 대해 제안 기법의 적용 가능성을 확장할 예정이다.

사 사 문 구

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. RS-2023-00252257, No. RS-2024-00397293).

참 고 문 헌

[1] Shvetsova, Nina, et al. “Anomaly detection in medical imaging with deep perceptual autoencoders.“ IEEE Access 9 (2021): 118571-118583.

[2] Su, Yixuan, et al. “Pandagpt: One model to instruction-follow them all.“ arXiv preprint arXiv:2305.16355 (2023).

[3] Baid, Ujjwal, et al. “The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification.“ arXiv preprint arXiv :2107.02314 (2021).

[4] He, Kaiming, et al. “Masked autoencoders are scalable vision learners.“ Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2022).

[5] Gong, Dong, et al. “Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection.“ Proceedings of the IEEE/CVF international conference on computer vision. (2019).