의료 데이터의 멀티 모달 학습을 기반으로 한 임상 기록 생성 모델

유민서[1], 김현희[2]
[1]동덕여자대학교 문헌정보학과, [2]동덕여자대학교 정보통계학과
20220502@dongduk.ac.kr, heekim@dongduk.ac.kr

Clinical Note Generation Model Based on Multimodal Learning of Medical Data

Minseo Yoo[1], Hyon Hee Kim[2]
[1]Dept. of Library and Information Science, Dongduk Women's University
[2]Dept. of Statistics and Information Science, Dongduk Women's University

요 약

대한민국 의료공백에 의해 영상의학 진단이 지체됨에 따라 많은 환자들이 치료 시기를 놓치고 있다. 본 연구에서는 진단 가속을 위해 흉부 X-ray 이미지와 임상 노트 텍스트로 구성된 데이터를 멀티모달 학습시키고, 흉부 X-ray 이미지에 대한 임상 기록을 생성하는 모델을 제안하였다. 이미지 임베딩 생성에는 PubMed 텍스트/이미지 쌍을 학습한 BiomedCLIP을 사용하고, 이미지 임베딩을 텍스트화하고 최종 텍스트 생성하는 과정에는 PLM 모델 T5를 사용한다. T5는 경량화된 모델이므로 컴퓨팅 자원이 부족한 의료 실무 환경에서도 충분히 임상 노트를 생성을 수행할 수 있으며, 이를 통한 정밀의학의 실용화를 기대할 수 있다.

1. 서론

의료 위기에 대학병원에서의 영상의학 진단이 지체됨에 따라 질병 진단의 지연으로 적기를 놓치는 환자들이 늘고 있다[1]. 현재의 진단 의학은 영상의학 검사와 판독에 많은 시간과 인력이 필요하다는 점은 의료 인력 보조 도구의 필요성을 시사한다[2]. 인공지능이 보조 도구로 채택되어도 기존 연구는 대부분 LLM과 같은 거대 모델을 사용하며, 이를 사용하려면 GPU, TPU와 같은 처리장치가 필요하다. 현장에서 LLM을 학습시키기 위한 큰 초기 비용과 자원 투자 대신, 본 연구는 영상의학 진단의 보조도구로 활용될 수 있는 경량화된 임상 노트 생성 인공지능 모델을 제안하였다.

본 연구는 가장 많은 진단량을 갖는 흉부 X-ray 이미지와 진단 임상 노트를 기반으로 훈련되었으며, X-ray 이미지를 통해 흉부 질환 진단에 활용할 수 있는 AI가 선제적으로 검토한 진단 리포트를 생성하여 빠른 질병 진단 및 치료계획 수립에 보탬이 되고자 한다.

2. 관련 연구

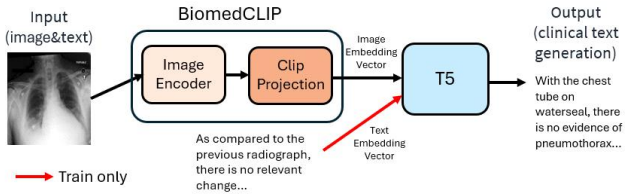
이미지 인코더로 사용한 BiomedCLIP모델[3]은 PubMed의 이미지-캡션 텍스트를 대조 학습하여 텍스트와 이미지를 같은 임베딩으로 훈련할 수 있는 모델이다. 제안한 모델은 텍스트로는 PubMedBERT[4], 이미지로는 ViT-B/16-224[5]를 사용하여 훈련했으며 분류, image-to-text 및 text-to-image, 질의응답의 downstream task에 적용할 수 있다.

텍스트 생성에 사용한 t5 모델은 Google에서 개발한 사전 학습된 자연어처리 모델로, 자연어처리 문제를 text-to-text로 처리하여 다양한 task에 대해 한 모델로 사용할 수 있는 모델이다. 웹에서 수집한 C4 거대 언어 데이터셋을 사전 학습하고 각 문제 상황에 대해 미세 조정하여 고안되었다.

3. 모델 학습 프로세스

본 연구는 멀티모달 학습을 통해 이미지와 진단에 대한 텍스트를 동시에 학습하여 feature 추출의 정확도를 높이고자 한다[6]. 기존 PubMed 텍스트 및 이미지뿐만 아니라, 흉부에 특화된 데이터의 중점적

학습에 유리한 데이터셋인 kaggle Curated CXR report generation dataset[7]을 사용하였다. 이 데이터셋은 MIMIC-CXR[8] 및 OpenI[9] 데이터를 후가공한 흉부 X-ray 및 진단 리포트이며, TFRecord로 저장된 93,347쌍 중 5,000쌍을 사용하였다.



[그림 1] 모델 학습 과정

그림 1은 멀티 모달 모델 학습 과정을 나타내고 있다. 먼저, 인코더에서 환자의 검사 진단 이미지 및 임상 텍스트의 멀티모달 데이터를 대조 학습할 수 있는 BiomedCLIP을 학습시킨다. 학습에 사용될 진단 이미지를 모델의 preprocess 함수를 이용해서 224*224 크기로 중앙 crop을 진행하고, 텐서로 변환한 뒤 정규화를 수행하는 방식으로 전처리하였다.

학습 코드는 가용 자원에서 구동하기 위해 DistributedSampler와 DistributedDataParallel로 분산학습을 수행하였다. 사용한 CLIP-T5 모델의 구조는 다음과 같다. 먼저 전처리된 이미지를 Biomedclip Encoder를 사용해서 토큰으로 변환한 뒤, clip_projection 단계에서 CLIP 모델의 출력 이미지 임베딩을 T5의 차원과 맞게 변환한다. CLIP 이미지 임베딩과 텍스트 토큰 임베딩을 결합하여 T5의 훈련하고 학습 손실을 계산한다.

최종 임상 텍스트 생성은 앞서 사용하였던 모델 중 가장 성능이 좋았던 모델을 불러와 새 데이터에서 진행하여 조건부 생성된 임상 노트 텍스트를 획득할 수 있다.

4. 결론 및 제언

본 연구의 학습 과정에서는 사용하고자 하는 의학 분과에 맞는 이미지와 텍스트를 의학 도메인에 특화된 BiomedCLIP 모델을 통해 도출한 멀티모달 벡터 임베딩을 통해 T5 모델이 텍스트를 생성하도록 설계하였고, 실제 평가 단계에서는 텍스트를 제외하고 사용할 수 있도록 하였다. 본 연구에서 제시한 모델은 용량이 적은 T5 모델을 사용하고, 가중치를 불러오는 방식으로 텍스트 인코더를 제거하여 현장 실무자의 작업환경에서 활용할 수 있도록 경량화되었다. 이에 기존의 LLM 기반 진단 생성보다 빠르게 현장에 적용될 수 있으며, AI 기반 정밀의학에 활용될 수 있을 것으로 기대된다.

참고문헌

[1] 박성제, 뇌혈전 의심 중3 응급실서 12시간 대기…부모 "현실 개탄스러워", 연합뉴스, 2024, https://www.yna.co.kr/view/AKR20240829100900051?input=1195m

[2] 권연아, 국내 병원들, AI 의료체계 속속 도입… 첨단의료 혁신은 진행 중, 바이오타임스, 2024, https://www.biotimes.co.kr/news/articleView.html?idxno=16055

[3] Zhang, Sheng, et al. "BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs." arXiv preprint arXiv:2303.00915 (2023).

[4] Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., ... & Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH), 3(1), 1-23.

[5] Dosovitskiy, Alexey. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).

[6] Mokady, Ron, Amir Hertz, and Amit H. Bermano. "Clipcap: Clip prefix for image captioning." arXiv preprint arXiv:2111.09734 (2021).

[7]https://www.kaggle.com/datasets/financekim/curated-cxr-report-generation-dataset

[8] https://physionet.org/content/mimic-cxr/2.1.0/

[9] https://openi.nlm.nih.gov/