흉부 X-ray 영상에서의 Vision Transformer 계열 모델을 활용한 다중 질환 분류 연구

강지은[1], 류다현[2], 장준혁[2], 조성은[3], 조경진[4]
[1]가천대학교 AI 소프트웨어학부(인공지능전공) 학부생
[2]서경대학교 컴퓨터공학과 학부생
[3]한신대학교 AI·SW 학과 학부생
[4]SK 주식회사

kje147459@gachon.ac.kr, {fbekgus413, joonsoon65}@skuniv.ac.kr, shap2819@hs.ac.kr, kjcho96@sk.com

Multiple disease classification using Vision Transformer series models in chest X-ray images

Jieun Kang[1], Dahyun Ryu[2], Junhyuk Jang[2], Sungeun Cho[3], Kyungjin Cho[4]
[1] School of Computing, Gachon University
[2] Dept. of Computer Engineering, Seo-Kyeong University
[3] School of Computing, Hanshin University
[4] SK Inc.

요 약

흉부 X-ray 영상은 다양한 질환 진단에 널리 활용되지만, 영상의학과 전문의가 아닌 임상의에게는 판독에 어려움이 따른다. 본 연구는 이러한 한계를 극복하기 위해 MIMIC-CXR-JPG v2.1.0 dataset을 활용하여 Vision Transformer와 Convolutional Neural Network(CNN) 계열 모델을 적용하고 흉부 질환 예측의 가능성을 탐구하였다. 성능 비교를 통해 Transformer 기반 접근법이 CNN 보다 우수함을 확인하였으며, 높은 성능의 모델과 Grad-CAM 시각화를 적용한 웹 기반 보조 시스템을 개발하여 예측 근거의 해석 가능성과 신뢰성을 강화하였다.

1. 서론

흉부 방사선 영상은 폐질환 진단에서 가장 표준적인 방법으로 사용되지만, 판독자 간 차이와 방대한 데이터 해석 부담이 여전히 문제로 남아 있다. 이를 보완하기 위해 딥러닝 기반 연구가 활발히 진행되었으며, 특히 Convolutional Neural Network(CNN)는 높은 성능을 보였다. 그러나 CNN은 주로 유도 편향에 의존하는 특성으로 인해 전체 폐 영상의 전역적 정보를 충분히 활용하기 어렵다는 한계를 가진다.

본 연구에서는 이러한 제약을 극복하기 위해 self-attention 메커니즘을 통한 Vision Transformer(ViT) [1] 기반 모델을 도입한다. 또한 Grad-CAM 기반 시각화를 활용해 모델의 판별 근거를 제시함으로써 의료 현장에서 요구되는 설명 가능성을 강화할 수 있다. 사전 학습 데이터는 MIMIC-CXR-JPG v2.1.0 dataset [2]을 활용하여 임상 환경에 가까운 학습이 가능하다.

2. 데이터 및 전처리

데이터셋은 MIMIC-CXR-JPG v2.1.0 dataset의 Postero-Anterior, Antero-Posterior, Erect 자세로 촬영된 흉부 X-ray 영상을 사용하였다. 동일 환자의 중복 촬영으로 인한 데이터 편향을 최소화하기 위해 각 study_id 당 하나의 이미지를 촬영 일시 기준으로 선별하였다. 이후 전체 데이터를 환자 단위로 무작위 분할하여 train/valid/test = 8:1:1 비율로 구성하였다.

라벨의 -1(불확실)과 NaN 은 0 으로 변환하였다. 또한 모든 질병 라벨이 0인 경우 'No Finding'을 1로 설정하였다. 임상적으로 불명확한 'Pleural Other' 항목은 제외하였다. 이를 통해 다중 라벨 구조를 유지하면서 학습에 불필요한 잡음을 줄였다.

영상은 모델 입력에 맞추어 224×224 크기로 리사이즈하고, ImageNet 사전 학습 모델과 일관성을 위해 평균과 표준편차로 정규화하였다. 데이터 증강은 랜덤 크롭, 이동, 밝기, 대비 조정을 적용하였으며, 좌우 반전은 심장 위치의 임상적 특성을 고려하여 제외하였다. 이때 random seed 값은 42로 고정하였다.

3. 모델 및 학습 방법

본 연구에서는 흉부 X-ray 영상의 폐질환 다중 분류를 위해 ViT 계열의 다양한 모델을 적용하여 비교 실험을 수행하였다. 구체적으로 ViT-Small/16, Swin Transformer-Tiny [3], Convolutional Neural Networks Meet Vision Transformer(CMT-S) [4] 모델을 각각 학습하였다. 추가로, 성능 차이를 확인하기 위해 CNN 의 대표적인 모델인 Residual Network(ResNet50) [5]을 학습하여 비교 실험에 활용하였다. 이를 통해 의료 영상 데이터에 적합한 모델 구조를 탐색하고자 하였다.

ViT 는 이미지를 고정 크기 패치로 분할한 뒤 Transformer 인코더에 입력하여 전역 정보를 학습하는 구조로, 전체 영상의 패턴을 효과적으로 반영한다. Swin Transformer 는 윈도우 기반 self-attention 을 도입하여 연산량을 줄이면서도 계층적 표현 학습을 가능하게 하여, 다양한 해상도의 정보를 통합적으로 학습할 수 있다. CMT 는 합성곱 기반 특징 추출과 Transformer 블록을 결합한 구조로, 지역적 세부 특징과 전역 문맥 정보를 동시에 학습할 수 있다.

학습 과정에서는 ImageNet 데이터셋으로 사전 학습된 가중치를 활용하여 전이 학습을 수행함으로써, 학습 안정성과 수렴 속도를 향상시켰다. 손실 함수는 다중 라벨 환경을 고려하여 BCEWithLogitsLoss 를 채택하였다. 배치 크기는 32, 옵티마이저는 AdamW [6], 초기 학습률은 $1e-4$, 스케줄러를 적용하여 30 에폭 동안 학습을 수행하였다.

4. 실험 결과

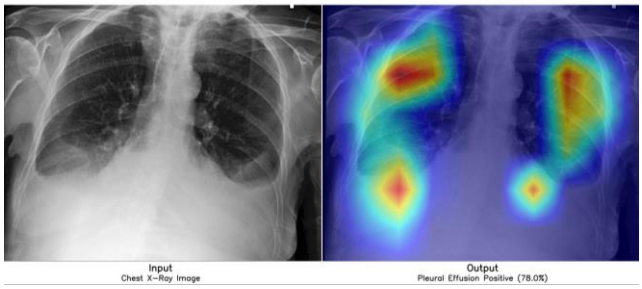
평가 지표는 흉부 X-ray 영상의 불균형 특성을 고려하여 Accuracy 와 mean Area Under ROCurve(mAUROC)를 사용하였다.

<표 1> 모델별 성능 비교

Model	Accuracy ↑	mAUROC ↑
ResNet-50	0.7765	0.7272
ViT-S	0.7741	0.7455
Swin-T	0.8127	0.7914
CMT-S	**0.8143**	**0.7934**

실험 결과, 세 모델 모두에서 임상적으로 의미 있는 수준의 성능을 보였으며, 특히 CMT-S 모델은 mAUROC 0.7934 로 가장 우수한 예측 성능을 기록하였다(표 1). ViT 모델은 깊은 구조와 높은 표현력을 바탕으로, 전반적으로 양호한 성능을 보였으나, 특정 질환에서 성능 편차가 관찰되었다. Swin 모델은 경량화 구조임에도 불구하고 전반적으로 안정적인 결과를 보여, 자원이 제한된 환경에서도 활용 가능성을 확인하였다. ResNet-50 은 Transformer 계열 모델보다 상대적으로 낮은 성능을 보였으며, 이를 통해 Transformer 기반 접근법의 성능적 우위를 확인하였다.

또한, 원본 X-ray 영상과 Grad-Cam 시각화를 통해 모델이 실제로 병변이 위치한 영역에 집중함을 확인하였다(그림 1). 이는 단순한 예측 성능을 넘어, 모델의 의사결정 과정을 시각적으로 해석할 수 있음을 보여주며 임상적 신뢰성을 높이는 근거로 작용한다.



(그림 1) CMT 모델의 Grad-CAM 시각화 결과

5. 결론

본 연구는 영상의학과 전문의가 아닌 임상의들의 흉부 X-ray 판독을 지원하기 위해 Transformer 계열 모델을 적용한 웹 기반 보조 시스템을 개발하였다. 연구 결과, 해당 모델들은 판독 정확도를 향상하고 임상적 의사결정을 지원할 수 있는 보조 도구로 발전할 가능성을 보여주었다. 향후에는 다양한 임상 환경과 대규모 데이터셋을 활용하여 모델의 일반화 성능과 효율성을 높이고, 실제 임상 workflow 에 통합될 수 있도록 발전시킬 필요가 있다.

※ 본 논문은 과학기술정보통신부 대학디지털교육역량강화 사업의 지원을 통해 수행한 한이음 드림업 프로젝트 결과물입니다.

참고문헌

1. Dosovitskiy, A., et al., *An image is worth 16x16 words: Transformers for image recognition at scale.* arXiv preprint arXiv:2010.11929, 2020.
2. Johnson, A.E., et al., *MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs.* arXiv preprint arXiv:1901.07042, 2019.
3. Liu, Z., et al. *Swin transformer: Hierarchical vision transformer using shifted windows*. in *Proceedings of the IEEE/CVF international conference on computer vision*. 2021.
4. Guo, J., et al. *Cmt: Convolutional neural networks meet vision transformers*. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.
5. He, K., et al. *Deep residual learning for image recognition*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
6. Loshchilov, I. and F. Hutter, *Decoupled weight decay regularization.* arXiv preprint arXiv:1711.05101, 2017.