Performance Evaluation of Vision Transformer-based Pneumonia Detection Model using Chest X-ray Images

Junyong Chang[1], Youngeun Choi[2], Seungwan Lee[1,2,*]

[1]Department of Radiological Science, Konyang University
[2]Department of Medical Science, Konyang University

Received: September 27, 2024. Revised: October 24, 2024. Accepted: October 31, 2024.

ABSTRACT

The various structures of artificial neural networks, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have been extensively studied and served as the backbone of numerous models. Among these, a transformer architecture has demonstrated its potential for natural language processing and become a subject of in-depth research. Currently, the techniques can be adapted for image processing through the modifications of its internal structure, leading to the development of Vision transformer (ViT) models. The ViTs have shown high accuracy and performance with large data-sets. This study aims to develop a ViT-based model for detecting pneumonia using chest X-ray images and quantitatively evaluate its performance. The various architectures of the ViT-based model were constructed by varying the number of encoder blocks, and different patch sizes were applied for network training. Also, the performance of the ViT-based model was compared to the CNN-based models, such as VGGNet, GoogLeNet, and ResNet. The results showed that the traninig efficiency and accuracy of the ViT-based model depended on the number of encoder blocks and the patch size, and the F1 scores of the ViT-based model ranged from 0.875 to 0.919. The training effeciency of the ViT-based model with a large patch size was superior to the CNN-based models, and the pneumonia detection accuracy of the ViT-based model was higher than that of the VGGNet. In conclusion, the ViT-based model can be potentially used for pneumonia detection using chest X-ray images, and the clinical availability of the ViT-based model would be improved by this study.

Keywords: Vision transformer, Deep learning, Pneumonia detection, Chest X-ray image

I. INTRODUCTION

인공지능 기술은 인공 신경회로를 통해 인간의 사고와 유사한 결과를 도출해낼 수 있는 기술로 그 응용분야가 점점 확장되고 있다. 인공지능 기술 중 딥러닝은 심층신경망과 다량의 데이터를 통해 특정 임무를 수행할 수 있는 모델을 스스로 학습하고, 정답에 가까운 결과를 도출해낼 수 있는 기술이다. 대표적인 딥러닝 모델로 convolutional neural network(CNN)[1], recurrent neural network(RNN)[2], transformer 기반 모델[3] 등이 있다. 본래 transformer 기반 모델은 언어와 문맥의 흐름을 학습하여 번역, 작문 등 자연어 처리를 위해 개발되었다[4,5]. 또한 transformer 내부 구조를 변형하여 영상을 입력하고, 입력 영상를 패치(patch) 단위로 나누어 학습할 수 있는 모델이 개발되었다. 이 모델을 Vision transformer(ViT) 라고 한다[6]. ViT 모델은 transformer 구조를 사용하여 확장성이 뛰어나며 빅데이터를 이용한 학습 시 성능 수렴 없이 학습 모델의 정확도가 향상되는 장점이 있다[6,7]. 또한 attention 기술을 이용하여 학습 영상을 광범위하게 학습함으로써 객체 탐지 및 영상 분류 분야에서 우수한 성능을 보여주고 있다[8,9].

* Corresponding Author: Seungwan Lee E-mail: slee1@konyang.ac.kr Tel: +82-42-600-8443
Address: Department of Radiological Science, Konyang University, 158 Gwanjeodong-ro, Seo-gu, Daejeon, Korea (35365)

폐렴 진단은 일반적으로 흉부 X-선 영상을 통해 이루어지는데, 폐렴이 있는 경우 X-선 영상에서 폐의 음영 및 내부 구조물의 분포 형태가 정상 폐와 비교하여 달라지는 것을 확인할 수 있다[10]. 그러나 음영 차이 및 구조물 분포 형태는 판독자의 주관적 해석에 따라 달라질 수 있고, 이와 같은 특징은 폐렴 진단 정확도 저하로 이어질 수 있다[11]. 하지만 폐렴 진단 시 딥러닝 모델을 이용하면 인간이 놓칠 수 있는 미세한 음영 및 구조물 분포 변화를 감지할 수 있으며 주관적 해석을 배제할 수 있기 때문에 재현성 및 정확도 높은 진단이 가능하다. 특히 ViT는 입력 영상을 패치 단위로 나누어 학습하기 때문에 학습 정보의 독립성을 확보할 수 있을뿐만 아니라 학습 시 주변 패치와의 상관관계를 파악함으로써 폐렴 진단의 정확도를 향상시킬 수 있을것으로 기대된다. 또한 ViT는 타 딥러닝 모델과 비교하여 학습에 필요한 비용이 작기 때문에 복잡한 네트워크를 효율적으로 학습시킬 수 있으며 과적합 문제를 방지할 수 있다. 이러한 장점 때문에 ViT 모델을 이용하여 흉부 X-선 영상 기반 폐 질환 검출을 위한 연구들이 수행된 바 있다[12-14]. 하지만 기존 연구에서는 네트워크 구조, 패치 크기 등 ViT 모델 학습조건 변화에 따른 성능 및 학습효율이 평가되지 않은 한계가 있다.

따라서 본 연구에서는 흉부 X-선 영상을 이용하여 폐렴을 진단할 수 있는 ViT 기반 모델을 개발하고, 학습 파라미터 변화에 따른 모델의 성능을 평가했다. 또한 기존 모델과 폐렴 진단 정확도 비교를 통해 본 연구에서 개발한 ViT 기반 모델의 성능을 검증했다.

II. MATERIAL AND METHODS

1. 데이터 준비

본 연구에서는 Kaggle 데이터베이스에서 제공하는 폐렴 및 정상 폐 X-선 영상을 사용했다. 모델훈련에는 256 × 256 크기의 정상 폐 영상 1,349장과 폐렴 폐 영상 3,883장, 총 5232장을 사용했다. 모델의 성능을 평가하기 위해 동일한 크기의 정상 폐 영상 234장과 폐렴 폐 영상 390장, 총 624장을 사용했다. 정확한 모델 성능 평가를 위해 훈련용 영상과 성능평가용 영상은 중복되지 않도록 하였다. 앞서 기술한 바와 같이 학습 데이터가 증가할수록 ViT 모델의 성능은 수렴 없이 향상되는 장점이 있다. 본 연구에서는 학습 데이터 양을 제외한 파라미터 변화에 따른 ViT 모델의 성능을 평가하고 타 모델과 비교하고자 했다. 따라서 기존 발표된 연구를 활용하여 훈련 영상 규모를 설정하였으며, 인위적인 데이터 증강 기술을 적용하지 않았다[15,16].

2. ViT 모델 구조 및 학습

일반적으로 ViT 모델은 학습 시 입력 영상을 특정 크기로 나누어 패치화 한다. 패치의 크기가 작을수록 국한된 영역의 특징을 추출하여 학습할 수 있지만, 주변 영역과의 상관관계 파악이 어렵고 학습 시간이 증가되는 단점이 있다. 본 연구에서는 패치 크기가 ViT 기반 폐렴 진단 모델에 미치는 영향을 평가하기 위하여 8 × 8, 16 × 16 및 32 × 32 크기의 패치를 설정하여 학습을 진행했다.

본 연구에서 개발한 ViT 네트워크는 Fig. 1과 같이 크게 입력 모듈, ViT encoder block 및 출력 모듈로 나뉜다. 입력 모듈의 첫 번째 단계는 linear projection으로 입력 영상을 사전에 설정한 패치 단위로 나눈 후 encoder block에 순차적으로 입력될 수 있도록 선형으로 정렬한다. 두 번째 단계는 patch and position embedding으로 정렬된 패치에 flatten 및 선형 변환 과정을 적용하여 벡터로 변환하고, 변환된 벡터에 class token을 추가하여 encoder block에 입력할 수 있는 데이터를 구성한다.

ViT encoder block은 layer normalization, multi-head self attention(MSA), multi layer perceptron (MLP) 단계로 구성되었다. 입력 모듈로부터 전달된 데이터는 layer normalization 과정을 거쳐 Eq. (1)과 같이 정규화 된다.

$$Layer\,Normalization(z_i) = \gamma \frac{z_i - \mu}{\sqrt{\sigma^2 + \epsilon}} + \beta \qquad (1)$$

z_i는 입력 모듈로부터 전달된 i번째 벡터, γ 및 β는 학습 가능 파라미터, μ는 입력 벡터의 평균, σ^2은 입력 벡터의 분산, ϵ는 최소 상수를 나타낸다.

정규화된 벡터 중 폐렴 진단 정확도를 향상 시킬 수 있는 특정 패치에 대해 높은 가중치를 부여하여 학습할 수 있도록 Eq. (2)와 같은 self-attention 과정을 거친다.

$$Selfattention(Q,K,V \mid \bar{z}_k) = Softmax(\frac{QK^{Transpose}}{\sqrt{z_k}})V \quad (2)$$

Q, K 및 V는 self attention 과정을 위해 정의한 query, key 및 value 행렬, $\bar{z}_k$는 k 차원을 갖는 정규화된 벡터이다. 본 연구에서는 encoder block의 성능을 향상시키기 위해 정규화된 벡터의 차원 수 만큼 self attention을 병렬로 반복하는 과정을 수행하였다. 다수의 self attention 과정을 통해 획득한 결과를 Eq. (3)과 같이 통합 및 선형 변환하여 MSA의 결과를 출력한다.

$$MSA(\bar{z}_k) = [SA_1(\bar{z}_k); \cdots ; SA_k(\bar{z}_k)]U_{msa} \quad (3)$$

$SA_j(\bar{z}_k)$는 j번째 self attention의 출력값, Umsa는 선형 변환 행렬을 나타낸다. MSA 출력값은 layer normalization 과정을 통해 재정규화된 후 MLP 과정을 거친다. MLP는 입력값을 비선형으로 변환하는 과정으로 모델이 입력영상으로부터 더 복잡한 패턴을 학습할 수 있게 도와주는 역할을 한다. Gaussian error linear unit(GELU) 활성화 함수는 자연어 및 영상 처리 분야에서 rectified linear unit(ReLU) 활성화 함수보다 우수한 성능을 보여준다[17,18]. 따라서 본 연구에서는 MLP 출력값의 비선형 변환를 위해 GELU 활성화 함수를 이용했다. ViT encoder block 내 skip connection을 추가하여 학습 과정에서 발생할 수 있는 정보 소실 문제를 방지하였다. ViT encoder block 개수 변화에 따른 학습 모델의 성능을 평가하기 위해 ViT encoder block 7-11 개를 선형 결합하여 각각의 네트워크를 구성하였다.

ViT encoder block의 출력값은 출력 모듈의 MLP head 층으로 입력된다. MLP head 층에서는 앞서 기술한 MLP 층과 동일한 과정이 수행되며 입력 벡터를 제외한 class token 만을 이용한다. 최종적으로 단순 linear 층을 통과하여 입력 영상의 폐렴 여부를 판별(classification)한다[19,20].

패치 크기 및 ViT encoder block 개수 변화에 따른 ViT 모델을 Table 1.에 정리하였다. ViT 모델 학습을 위해 학습률(learning rate) 4×10^{-3}, dropout 0.5, adaptive moment estimation(Adam) 최적화 함수, BCEWithLogitsLoss 손실 함수, 100회의 학습 횟수가 사용되었다.

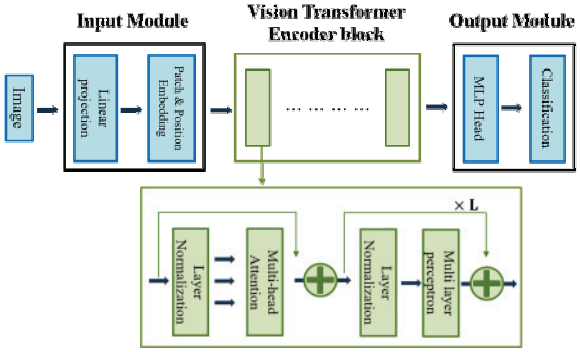

Fig. 1. Architecture of the ViT network used in this study.

Table 1. ViT models with various patch sizes and the number of the ViT encoder blocks.

Model #	Patch size	# of the ViT encoder block
1	8 × 8	7
2	8 × 8	8
3	8 × 8	9
4	8 × 8	10
5	8 × 8	11
6	16 × 16	7
7	16 × 16	8
8	16 × 16	9
9	16 × 16	10
10	16 × 16	11
11	32 × 32	7
12	32 × 32	8
13	32 × 32	9
14	32 × 32	10
15	32 × 32	11

본 연구에서 개발한 ViT 모델의 성능 검증 및 비교를 위해 VGGNet, ResNet 및 GoogLeNet을 이용하여 폐렴 진단 모델을 구성하였다. VGGNet은 16개의 3 × 3 convolution 층, 5개의 2 × 2 max-pooling 층, 3개의 fully-connected 층으로 구성하였다[21]. ResNet은 8개의 residual block으로 구성하였으며, 각 residual

block은 2개의 3 × 3 convolution 층이 skip connection 형태로 연결된 구조를 갖는다[22]. GoogLeNet은 19개의 3 × 3 convolution 층 및 4개의 3 × 3 max-pooling 층으로 구성하였으며, max-pooling을 동시에 수행한 후 병합되도록 하였다[23].

3. 성능 평가

본 연구에서 개발한 ViT 모델의 성능 평가를 위해 정상 및 폐렴 폐 X-선 영상을 학습 모델에 무작위로 입력했다. 입력 영상에 대하여 ViT 모델이 출력한 정상 또는 폐렴 진단 결과를 이용하여 Eq. (4) - (7)과 같은 정확도(accuracy), 정밀도(precision), 재현도(recall) 및 F1-score를 계산했다.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{4}$$

$$Precision = \frac{TP}{TP+FP} \tag{5}$$

$$Recall = \frac{TP}{TP+FN} \tag{6}$$

$$F1\,score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{7}$$

TP, TN, FP 및 FN은 각각 학습 모델이 폐렴으로 진단한 결과에 대한 진양성(True positive), 진음성(True negative), 위양성(False positive) 및 위음성(False negative)을 나타낸다. F1-score는 분류를 목적으로 하는 인공지능 기술의 성능을 평가하기 위해 범용적으로 사용되는 지표이다[24]. 본 연구에서 개발한 ViT 모델은 정상 및 폐렴 여부를 이진 분류하는 목적을 갖기 때문에 개발한 모델의 성능 평가에 적합한 지표라고 판단하여 사용했다.

III. RESULT

Fig. 2는 학습이 완료된 ViT 모델에 성능평가용 영상을 입력하여 획득한 결과를 시각화 한 것이다. 정상 또는 폐렴 진단 예측결과가 실제결과와 일치하지 않을 경우 붉은색 선으로 영상을 강조할 수 있도록 하였다. 이를 통해 ViT 모델이 정상 폐와 폐렴 폐를 어떻게 분류하는지 시각적으로 이해할 수 있다.

Fig. 3은 본 연구에서 개발한 ViT 및 비교 모델 학습 시 학습 횟수에 따른 정확도 및 손실값을 보여준다. 학습 횟수 20회 이상에서 ViT 모델의 학습 정확도 및 손실값이 특정값으로 수렴하는 경향을 확인하였다.

Table 2는 본 연구에서 개발한 ViT 및 비교 모델의 학습에 사용되는 파라미터 수(trainable parameters), 모델 크기(model size) 및 훈련 시간(training time)을 보여준다. 또한 Fig. 4는 encoder block 개수 및 패치가 크기가 다른 특정 ViT 및 비교 모델에 대한 학습 시간을 보여주고 있다. ViT 모델의 encoder block 개수가 증가함에 따라 학습에 사용되는 파라미터 수는 평균 1.31배, 모델 크기는 평균 1.31배, 훈련 시간은 평균 1.12배 증가되는 결과를 확인하였다. 패치 크기가 커짐에 따라 학습에 사용되는 파라미터 수는 평균 1.03배 증가하였지만 모델 크기 및 훈련 시간은 각각 평균 0.20 및 0.76배 감소하는 결과를 확인하였다. VGGNet, ResNet 및 GoogLeNet 대비 ViT #10 모델의 학습에 사용된 파라미터 수는 각각 0.06, 0.78 및 8.39배, 모델 크기는 각각 0.25, 1.51 및 4.90배, 훈련 시간은 0.71, 0.93 및 0.99배로 확인되었다. ViT #1~5 모델을 제외한 나머지 모델의 학습 시간은 비교 모델에 비해 짧은 결과를 확인하였다.

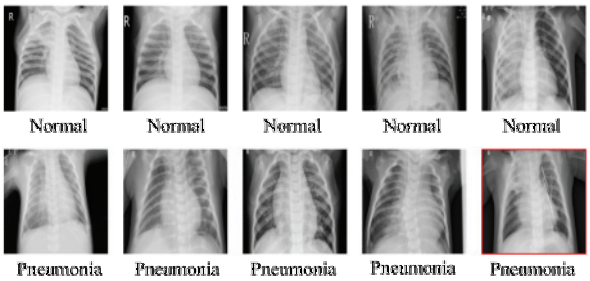

Fig. 2. Result images from the ViT-based pneumonia detection model with diagnostic decision.

Table 3은 본 연구에서 개발한 ViT 및 비교 모델를 통해 획득한 결과에 대한 정확도, 정밀도, 재현율 및 F1-score를 보여준다. 또한 Fig. 5는 encoder block 개수 및 패치가 크기가 다른 특정 ViT 및 비교 모델에 대한 F1-score 평가 결과를 보여주고 있다. Encoder block 개수가 증가함에 따라 ViT 모델

의 정확도, 정밀도, 재현율 및 F1-score는 평균 1.02배 향상되는 결과를 확인하였고, 패치 크기가 ViT 모델 성능에 미치는 영향은 미미한 결과를 확인하였다. VGGNet 대비 ViT #10 모델의 정확도, 정밀도, 재현율 및 F1-score는 각각 평균 1.05, 1.06, 1.04 및 1.05배 높은 결과를 확인하였다. ResNet 및 GoogLeNet 대비 ViT #10 모델의 정확도, 정밀도, 재현율 및 F1-score는 평균적으로 0.93배 낮은 결과를 확인하였다. Table 3 및 Fig. 5에서 확인할 수 있는 바와 같이 본 연구에서 개발한 모든 ViT 모델의 F1-score는 VGGNet과 동등하거나 높은 결과를 보여주었다. 또한 모든 ViT 모델의 F1-score가 0.9에 근접하거나 그 이상인 결과를 확인하였다.

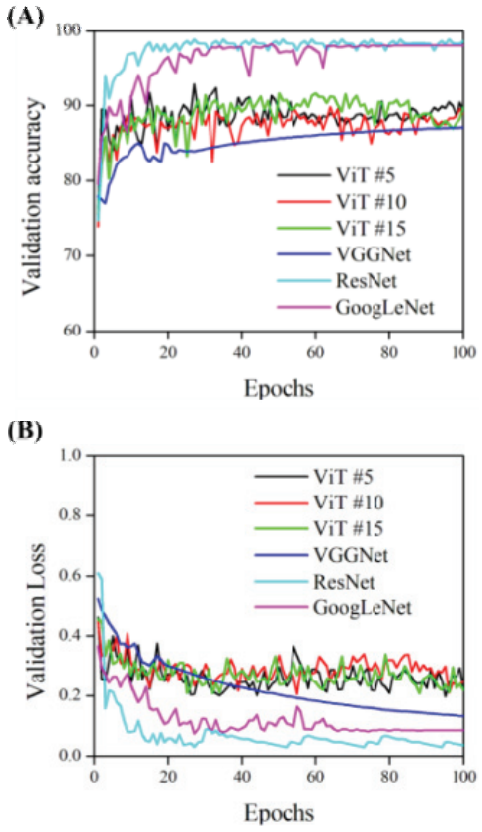

Fig. 3. (A) Validation accuracy and (B) validation loss for the ViT and comparative models.

IV. DISCUSSION

본 연구에서는 폐렴 진단이 가능한 ViT 기반 모델을 개발하고 흉부 X-선 영상을 이용하여 개발 모델의 성능을 평가하였다. 또한 CNN 기반 모델과의 비교를 통해 개발 모델의 성능을 검증하였다. 성능 평가 지표로 정확도, 정밀도, 재현율 및 F1 score를 분석하였으며 학습 가능 파라미터 수, 모델 크기 및 학습 속도 분석을 통해 개발 모델의 학습 효율을 평가하였다.

ViT 기반 모델의 encoder block의 개수가 증가함에 따라 해당 모델의 학습 효율은 감소하는 결과를 확인하였다. 이는 CNN 기반 모델과 비슷하게 네트워크의 깊이가 깊어짐에 따라 모델 크기가 증가하여 학습 효율이 감소하기 때문이다. 반면에 패치 크기의 증가는 ViT 기반 모델의 학습 효율을 향상시켰다. 이는 입력 영상 패치화 과정에서 패치 크기 증가에 따라 상대적으로 학습에 사용될 전체 패치 개수가 감소했기 때문이다. 따라서 ViT 기반 모델의 학습 효율 최적화를 위해 encoder block 개수 및 패치 크기의 적절한 설정이 필요하다고 판단된다. 또한 7~11개 ViT encoder block을 사용하는 모델 학습 시 16 × 16 크기 이상의 패치를 사용하면 VGGNet, ResNet 및 GoogLeNet에 비해 학습 시간을 줄일 수 있는 결과를 확인하였다.

Table 2. Trainable parameters, sizes and training time of the ViT and comparative models

	Trainable parameter	Model size (MB)	Training time (h)
ViT #1	5,545,729	809.41	5.39
ViT #2	6,335,489	924.42	5.58
ViT #3	7,125,249	1039.43	6.53
ViT #4	7,915,009	1154.45	6.81
ViT #5	8,704,769	1269.46	7.36
ViT #6	5,594,881	134.6	3.31
ViT #7	6,384,641	153.61	3.40
ViT #8	7,174,401	172.62	3.49
ViT #9	7,964,161	191.63	3.58
ViT #10	8,753,921	210.65	3.69
ViT #11	5,791,489	45.35	3.19
ViT #12	6,581,249	51.61	3.36
ViT #13	7,371,009	57.87	3.44
ViT #14	8,160,769	64.13	3.53
ViT #15	8,950,529	70.4	3.58
VGGNet	139,573,185	844.31	5.23
ResNet	11,170,753	139.87	3.97
GoogLeNet	1,043,105	42.95	3.70

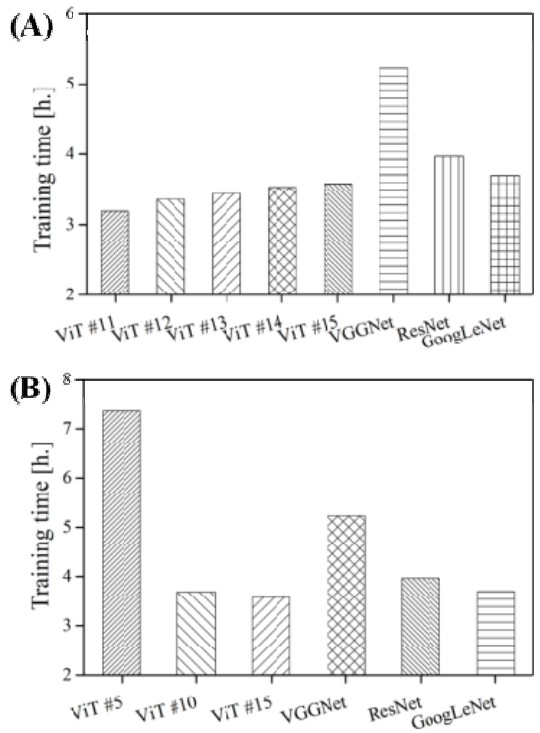

Fig. 4. Training time for the comparative models and the ViT models with (A) the different patch sizes and (B) the different number of encoder blocks.

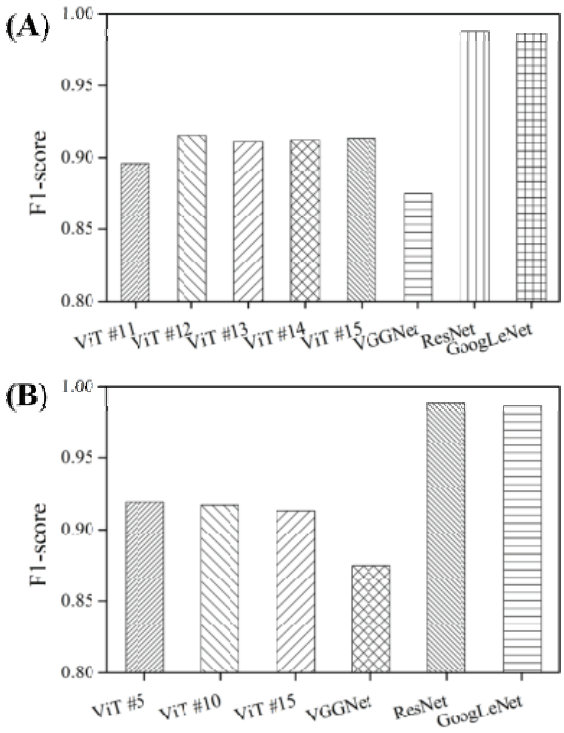

Fig. 5. F1-scores for the comparative models and the ViT models with (A) the different patch sizes and (B) the different number of encoder blocks.

Table 3. Performance evaluation results of the ViT and comparative models

	Accuracy	Precision	Recall	F1 score
ViT #1	0.873	0.849	0.902	0.875
ViT #2	0.872	0.852	0.899	0.875
ViT #3	0.895	0.889	0.906	0.898
ViT #4	0.913	0.908	0.925	0.916
ViT #5	0.916	0.905	0.933	0.919
ViT #6	0.889	0.892	0.895	0.894
ViT #7	0.884	0.886	0.892	0.889
ViT #8	0.899	0.905	0.902	0.903
ViT #9	0.896	0.902	0.899	0.900
ViT #10	0.914	0.920	0.914	0.917
ViT #11	0.892	0.892	0.901	0.896
ViT #12	0.912	0.914	0.917	0.915
ViT #13	0.907	0.911	0.911	0.911
ViT #14	0.909	0.908	0.916	0.912
ViT #15	0.910	0.908	0.919	0.913
VGGNet	0.870	0.871	0.879	0.875
ResNet	0.986	0.988	0.988	0.988
GoogLeNet	0.985	0.985	0.988	0.986

ViT encoder block 개수가 증가함에 따라 학습 모델의 폐렴 진단에 대한 성능은 향상되는 결과를 확인하였다. 이는 네트워크 깊이가 깊어짐에 따라 학습 모델의 정확도가 향상되는 기존 CNN 기반 모델과 유사한 결과이다. 또한 encoder block 개수 증가에 따른 모델의 성능 향상 정도는 큰 패치를 사용한 모델보다 작은 패치를 사용한 모델에서 강조되는 결과를 보였다. 이를 통해 크기가 작은 패치를 사용한 ViT 기반 모델은 encoder block 개수 증가를 통해 모델의 성능을 최대화 할 수 있는 반면, 크기가 큰 패치를 사용한 ViT 기반 모델은 encoder block 개수 증가를 통한 성능 향상 정도가 상대적으로 낮은 결과를 확인하였다. 본 연구에서 개발한 ViT 기반 모델의 성능은 VGGNet 보다 우수한 결과를 확인하였으며, 이는 흉부 X-선 영상을 이용한 폐렴 진단 정확도 향상을 위해 본 연구에서 개발한 ViT 기반 모델이 VGGNet을 대체할 수 있음을 보여준다. 본 연구에서 개발한 ViT 기반 모델의 F1-score는 0.9에 근접하거나 그 이상인 결과를 보

여주었다. 이는 ViT 기반 모델의 폐렴 진단 정확도가 높은 수준이며, 임상적 활용 가능성을 확인할 수 있는 결과이다. 하지만 ViT 기반 모델의 폐렴 진단 성능은 ResNet 및 GoogLeNet에 비해 다소 낮은 결과를 보였다. 이를 통해 폐렴 진단 정확도의 최대화를 위해 ViT 기반 모델의 추가적인 성능 향상이 필요함을 알 수 있다. 본 연구에서는 ViT 네트워크 설계 시 ViT encoder block 내 1개의 MSA 층을 사용하였다. 앞서 기술한 바와 같이 MSA 과정을 통해 특정 패치에 높은 가중치를 부여하여 학습의 정확도를 향상시킬 수 있다. 따라서 ViT encoder block 설계 시 다수의 MSA 층을 사용함으로써 학습 모델의 성능을 향상시킬 수 있을 것이다[5,25,26]. 또한 모델 학습 시 사용한 학습 데이터 양, 손실 함수, 학습 횟수 등 hyperparameter 변화를 통해 본 연구에서 개발한 ViT 기반 모델의 성능을 추가로 향상시킬 수 있을 것이다.

V. CONCLUSION

본 연구에서는 흉부 X-선 영상을 이용하여 폐렴을 진단할 수 있는 ViT 기반 모델을 개발하고, encoder block 개수 및 패치 크기 변화에 따른 모델의 학습 효율 및 성능을 평가하였다. 또한 기존 CNN 기반 모델과의 비교를 통해 본 연구에서 개발한 모델의 성능을 검증하였다. 연구 결과를 통해 ViT 기반 모델의 학습 효율 및 성능 향상을 위해 encoder block 개수 및 패치 크기 최적화가 필요함을 보여주었다. 또한 ViT 기반 모델의 성능 최대화를 위한 추가 연구의 필요성을 확인할 수 있었다. 결론적으로 본 연구의 결과는 흉부 X-선 영상을 이용한 폐렴 진단 딥러닝 모델 개발을 위해 이용될 수 있을 것으로 생각되며, ViT 기반 모델의 임상적 활용 가능성을 향상시킬 수 있을 것이다.

Acknowledgement

본 연구는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업입니다. (과제번호: RS-2023-00211810)

Reference

[1] S. Yu, K. Ma, Q. Bi, C. Bian, M. Ning, N. He, Y. Li, H. Liu, Y. Zheng, "MIL-VT:Multiple Instance Learning Enhanced Vision Transformer for Fundus Image Classification", Medical Image Computing and Computer Assisted Intervention, Vol. 12908, pp. 45-54, 2021. https://doi.org/10.1007/978-3-030-87237-3_5

[2] S. Bhojanapalli, A. Chakrabarti, D. Glasner, D. Li, T. Unterthiner, A. Veit, "Understanding Robustness of Transformers for Image Classification", IEEE/CVF International Conference on Computer Vision, pp. 10231-10241, 2021. https://doi.org/10.48550/arXiv.2103.14586

[3] K. Simonyan, A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition", 2014. https://doi.org/10.48550/arXiv.1409.1556

[4] T. Mikolov, M. Karafiat, L. Burget, J. Cernocky, S. Khudanpur, "Recurrent neural network based language model", Annual Conference of the International Speech Communication Association, pp. 1045-1048, 2010.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, I. Polosukhin, "Attention Is All You Need", Advances in Neural Information Processing Systems, 2017. https://doi.org/10.48550/arXiv.1706.03762

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale", International Conference on Learning Representation, 2021. https://doi.org/10.48550/arXiv.2010.11929

[7] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lucic, C. Schmid, "ViViT: A Video Vision Transformer", IEEE/CVF International Conference on Computer Vision, pp. 6836-6846, 2021. https://doi.org/10.48550/arXiv.2103.15691

[8] Q. Wang, B. Li, T. Xiao, J. Zhu, C. Li, D. Wong, L. Chao, "Learning Deep Transformer Models for Machine Translation", Advances in Neural Information Processing Systems, pp. 2797-2806, 2017. https://doi.org/10.48550/arXiv.1706.03762

[9] K. Irie, A. Zeyer, R. Schluter, H. Ney, "Language Modeling with Deep Transformers", Computation and Language, Annual Conference of the International Speech Communication Association, 2019. https://doi.org/10.48550/arXiv.1905.04226

[10] Hashmi MF, Katiyar S, Keskar AG, Bokde ND, Geem ZW, "Efficient Pneumonia Detection in Chest Xray Images Using Deep Transfer Learning", Diagnostics, Vol. 10, No. 6, pp. 417, 2020. https://doi.org/10.3390/diagnostics10060417

[11] M. Elemraid, M. Muller, D. Spencer, S. Rushton, R. Gorton, M. Thomas, K. Eastham, F. Hampton, A. Gennery, J. Clark, "Accuracy of the Interpretation of Chest Radiographs for the Diagnosis of Paediatric Pneumonia", PLoS ONE, Vol. 9, No. 8, 2014. https://doi.org/10.1371/journal.pone.0106051

[12] S. Park, G. Kim, Y. Oh, J. B. Seo, S. M. Lee, J. H. Kim, S. Moon, J. K. Lim, C. M. Park, J. C. Ye, "Self-evolving vision transformer for chest X-ray diagnosis through knowledge distillation", Nature Communications, Vol. 13, pp. 3848, 2022. https://doi.org/10.1038/s41467-022-31514-x

[13] T. Wang, Z. Nie, R. Wang, Q. Xu, H. Huang, H. Xu, F. Xie, X.-J. Liu, "PneuNet: deep learning for COVID-19 pneumonia diagnosis on chest X-ray image analysis using Vision Transformer", Medical & Biological Engineering & Computing, Vol. 61, pp. 1395, 2023. https://doi.org/10.1007/s11517-022-02746-2

[14] S. Singh, M. Kumar, A. Kumar, B. K. Verma, K. Abhishek, S. Selvarajan, "Efficient pneumonia detection using Vision Transformers on chest X-rays", Scientific Reports, Vol. 14, pp. 2487, 2024. https://doi.org/10.1038/s41598-024-52703-2

[15] C. Chen, Q. Fan, "CrossViT:Cross-attention Multi-Scale Vision Transformer for Image Classification", IEEE/CVF International Conference on Computer Vision, pp. 357-366, 2021. https://doi.org/10.48550/arXiv.2103.14899

[16] G. Okolo, S. Katsigiannis, N. Ramzan, "IEVIT: An enhanced vision transformer architecture for chest X-ray image classification", Computer Methods and Programs in Biomedicine, Vol. 226, 2022. https://doi.org/10.1016/j.cmpb.2022.107141

[17] H. Fang, J. Lee, N. Moosavi, I. Gurevych, "Transformers with Learnable Activation Functions", 2022. https://doi.org/10.48550/arXiv.2208.14111

[18] T. Li, F. Zhang, G. Xie, X. Fan, Y. Gao, M. Sun, "A high speed reconfigurable architecture for softmax and GELU in vision transformer", Electronics Letters, Vol. 59, No. 5, 2023. https://doi.org/10.1049/ell2.12751

[19] X. Mao, G. Qi, Y. Chen, X. Li, R. Duan, S. Ye, Y. He, H. Xue, "Towards Robust Vision Transformer", IEEE/CVF International Conference on Computer Vision and Pattern Recognition, pp. 12042-12051, 2022. https://doi.org/10.48550/arXiv.2105.07926

[20] Y. Bazi, L. Bashmal, M. Rahhal, R. Dayil, N. Ajlan, "Vision Transformers for Remote Sensing Image Classification", Remote Sensing, Vol. 13, No. 3, pp. 516, 2021. https://doi.org/10.3390/rs13030516

[21] K. Simonyan, A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition", The International Conference on Learning Representations, 2015. https://doi.org/10.48550/arXiv.1409.1556

[22] K. He, X. Zhang, S. Ren, J. Sun, "Deep Residual Learning for Image Recognition", IEEE/CVF International Conference on Computer Vision and Pattern Recognition, 2016. https://doi.org/10.48550/arXiv.1512.03385

[23] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, "Going Deeper with Convolutions", IEEE/CVF International Conference on Computer Vision and Pattern Recognition, 2015. https://doi.org/10.48550/arXiv.1409.4842

[24] Z. Vujovic, "Classification Model Evaluation Metrics", International Journal of Advanced Computer Science and Applications, Vol. 12, No. 6, 2021. https://doi.org/10.14569/IJACSA.2021.0120670

[25] S. Hong, G. Lee, W. Jang, S. Kim, "Improving Sample Quality of Diffusion Models Using Self-Attention Guidance", IEEE/CVF International Conference on Computer Vision, 2023. https://doi.org/10.48550/arXiv.2210.00939

[26] Z. Chen, Y. Duan, W. Wang, J. He, T. Lu, J. Dai, Y. Qiao, "Vision Transformer Adapter for Dense Predictions", The International Conference on Learning Representations, 2023. https://doi.org/10.48550/arXiv.2205.08534

흉부 X-선 영상을 이용한 Vision transformer 기반 폐렴 진단 모델의 성능 평가

장준용[1], 최용은[2], 이승완[1,2,*]

[1]건양대학교 방사선학과
[2]건양대학교 의과학과

요 약

Convolutional neural network(CNN), recurrent neural network(RNN)와 같은 다양한 인공 신경망이 연구되고 있으며, 타 인공지능 기반 모델의 기초 구조로 활용되고 있다. 그 중, 트랜스포머를 기반으로 하는 인공 신경망은 자연어 처리 분야에서 그 성능이 입증되었고, 활발하게 연구되고 있는 구조이다. 최근 트랜스포머 기반 인공 신경망의 내부구조 변경을 통해 영상처리가 가능한 Vision transformer(ViT) 모델이 개발되었다. 비젼 영상처리에 있어 ViT 모델의 정확도와 성능은 다양한 연구를 통해 입증되었다. 본 연구에서는 흉부 X-선 영상을 이용하여 폐렴을 진단할 수 있는 ViT 기반 모델을 개발하고, 개발 모델의 학습효율 및 성능을 정량적으로 평가하였다. ViT 기반 모델의 구조는 encoder block의 개수를 다르게 하여 설계하였고, 신경망 학습 시 패치의 크기를 다르게 설정하였다. 또한 개발한 ViT 기반 모델을 검증하기 위하여 기존 CNN 기반 모델인 VGGNet, GoogLeNet 및 ResNet 모델과 성능 비교를 수행하였다. 연구결과 ViT 기반 모델의 학습효율 및 성능은 encoder block의 개수 및 학습 패치 크기에 따라 변화함을 확인하였고 F1 score가 최소 0.875, 최대 0.919로 측정되었다. 32 × 32 크기의 패치를 이용하여 학습한 ViT 기반 모델의 학습효율은 기존 CNN 기반 모델에 비해 우수한 것으로 확인되었으며, 본 연구에서 설계한 모든 ViT 기반 모델이 VGGNet 보다 폐렴 진단의 정확도가 높은 결과를 확인하였다. 결론적으로 본 연구에서 개발한 ViT 기반 모델은 흉부 X-선 영상을 이용한 폐렴 진단에 잠재적으로 사용될 수 있으며, 본 연구를 통해 ViT 기반 모델의 임상적 활용가능성을 향상시킬 수 있을 것이다.

중심단어: Vision transformer, 딥러닝, 폐렴 진단, 흉부 X-선 영상

연구자 정보 이력

	성명	소속	직위
(제1저자)	장준용	건양대학교 방사선학과	학부생
(공동저자)	최용은	건양대학교 의과학과	대학원생
(교신저자)	이승완	건양대학교 방사선학과, 의과학과	교수