ISSN : 1226-7244 (Print)
ISSN : 2288-243X (Online)
j.inst.Korean.electr.electron.eng.Vol.28,No.3,390~396,September 2024
https://doi.org/10.7471/ikeee.2024.28.3.390
논문번호 24-03-21
148

특징 매칭을 이용한 페어와이즈 어텐션 강화 모델에 대한 연구

Research on Pairwise Attention Reinforcement Model Using Feature Matching

임 준 식*, 주 영 석*★

Joon-Shik Lim*, Yeong-Seok Ju*★

Abstract

Vision Transformer (ViT) learns relationships between patches, but it may overlook important features such as color, texture, and boundaries, which can result in performance limitations in fields like medical imaging or facial recognition. To address this issue, this study proposes the Pairwise Attention Reinforcement (PAR) model. The PAR model takes both the training image and a reference image as input into the encoder, calculates the similarity between the two images, and matches the attention score maps of images with high similarity, reinforcing the matching areas of the training image. This process emphasizes important features between images and allows even subtle differences to be distinguished. In experiments using clock-drawing test data, the PAR model achieved a Precision of 0.9516, Recall of 0.8883, F1-Score of 0.9166, and an Accuracy of 92.93%. The proposed model showed a 12% performance improvement compared to API-Net, which uses the pairwise attention approach, and demonstrated a 2% performance improvement over the ViT model.

요 약

Vision Transformer(ViT)는 패치 간의 관계를 학습하지만, 색상, 질감, 경계와 같은 중요한 특징을 간과할 경우 의료 분야나 얼굴 인식 등에서 성능 한계가 발생할 수 있다. 이를 해결하기 위해 본 연구에서는 Pairwise Attention Reinforcement(PAR) 모델을 제안한다. PAR 모델은 학습 이미지와 참조 이미지를 인코더에 입력하여 두 이미지 간의 유사성을 계산한 후, 높은 유사성을 보이는 이미지 어텐션 스코어 맵을 매칭하여 학습 이미지의 매칭 영역을 강화한다. 이를 통해 이미지 간의 중요한 특징이 강조되며, 미세한 차이도 구별할 수 있다. 시계 그리기 검사 데이터를 사용한 실험에서 PAR 모델은 Precision 0.9516, Recall 0.8883, F1-Score 0.9166, Accuracy 92.93%를 기록하였다. 본 모델은 Pairwise Attention 방식을 이용한 API-Net 대비 12% 성능이 향상되었으며, ViT 모델 대비 2%의 성능 향상을 보였다.

Key words : Vision Transformer, Clock Drawing Test, Attention Mechanism, Dementia Classification, Image Processing

* Dept, of Computer Engineering, Gachon University
★ Corresponding author
E-mail : jys4542@gachon.ac.kr
※ Acknowledgements
Manuscript received Sep. 5, 2024; revised Sep. 18, 2024; accepted Sep. 24, 2024.

Ⅰ. 서론

Transformer는 셀프 어텐션 메커니즘을 통해 장거리 의존성을 효과적으로 모델링하며 자연어 처리뿐만 아니라 이미지 분류에도 적용된다[1]. Vision Transformer (ViT)는 이미지를 패치로 나눠 Transformer 인코더에 입력하여 이미지 내 장거리 관계를 포착하는 데 효과적이다[2], [3]. 그러나 ViT는 이미지를 독립적으로 처리하므로, 시각적으로 유사한 이미지를 분류하는 데 어려움

을 겪을 수 있다. 특히 다양한 각도, 조명, 배경을 가진 데이터 셋에서 이러한 한계가 두드러지며, 의료 영상 분석이나 얼굴 인식처럼 높은 정밀도가 요구되는 작업에서는 더욱 심화된다[5]-[9].

이 문제를 해결하기 위해 기존 연구들은 다양한 기법을 제안해왔다. 예를 들어, 데이터 증강 기법을 활용하여 모델의 일반화 성능을 향상시키거나[10], [11], 복합적인 손실 함수를 도입하여 모델이 더 정밀한 특징을 학습할 수 있도록 하는 연구들이 있었다[12], [13]. 또한, 두 이미지 간의 상호작용을 활용하여 이미지 간의 관계를 모델링하는 방법도 제안되었다[5]. 그러나 이러한 방법들은 주로 이미지 간의 전반적인 구조와 패턴 학습에 중점을 두기 때문에, 위치나 모양과 같은 세부적인 시각적 차이를 감지하는 데 한계가 있다[12], [13], [15]. 그 결과, 유사한 이미지 간의 미세한 차이를 감지하는데 어려움을 겪는 경우가 발생한다[14]-[16].

본 연구에서는 이 문제를 해결하기 위해 두 이미지의 유사성을 활용해 어텐션 스코어를 강화하는 Pairwise Attention Reinforcement(PAR) 모델을 제안한다. PAR 모델은 학습 이미지와 참조 이미지를 함께 사용하여, 두 이미지 간의 유사도를 기반으로 학습 이미지의 어텐션을 강화한다. 유사도가 0.7 이상일 때 두 이미지의 어텐션 스코어 맵을 매칭하여 학습 이미지의 매칭 특징을 강화함으로써, 유사한 이미지 간 미세한 차이를 정밀하게 구분할 수 있다. PAR 모델은 Precision 0.9516, Recall 0.8883, F1-Score 0.9166, Accuracy 92.93%의 성능을 기록하였으며, Pairwise Attention 방식을 이용한 API-Net 대비 12% 성능이 향상되었으며, ViT 모델 대비 1~2%의 성능 향상을 보였다.

Ⅱ. 관련 연구

2.1. ViT의 어텐션 메커니즘 개선 연구

ViT의 성능 향상을 위해 어텐션 메커니즘을 개선하는 연구가 활발히 이루어지고 있다.[12]는 Dual Cross-Attention 학습을 통해 미세한 시각적 범주화를 개선했으며, [17]은 비균일 어텐션을 통해 병리학적 이미지 분류 성능을 향상시켰다. 이들은 어텐션 메커니즘을 강화하여 ViT의 성능을 높이고자 했다. 그러나 이들 연구는 여전히 개별 이미지의 어텐션에 중점을 두고 있어, 이미지 간의 상호작용을 충분히 반영하지 못하는 한계를 가지고 있다. 본 연구는 학습 이미지와 참조 이미지를 활용하여 어텐션 메커니즘을 강화함으로써 이러한 한계를 극복하려고 한다.

2.2. 유사 이미지 비교를 통한 성능 향상 연구

시각적으로 유사한 이미지를 비교하여 모델의 성능을 개선하려는 다양한 연구가 수행되었다. [5]는 AI 기반 시계 그리기 검사(CDT) 평가에서 위치와 형태의 오류를 감지하기 위해 Pairwise Interaction Network를 적용하였다. 이 접근법은 유사 이미지를 비교함으로써 중요한 특징을 강조하는 데 효과적임을 입증하였다. 그러나 기존 연구에서는 이미지 간 유사성 비교 과정에도 불구하고 이미지 핵심 특징이 충분히 반영되지 않아 이미지 분류 정확도에 한계를 가지고 있다. 이러한 문제를 해결하기 위해 본 연구에서는 유사 이미지 비교 방법을 확장하고, 이미지 핵심 특징 강화를 고려할 수 있는 새로운 접근법을 제안한다.

2.3. Transformer 기반 모델의 확장 및 변형 연구

ViT의 기본 구조를 변형하여 성능을 개선하려는 다양한 연구들이 진행되고 있다. [3]의 Swin Transformer는 윈도우 기반의 계층적 어텐션을 도입하여 성능을 높였고, [18]의 Pyramid Vision Transformer는 피라미드 구조를 통해 밀집 예측 작업에서 우수한 성능을 보였다. 이러한 변형 모델들은 ViT의 구조적 한계를 극복하려는 시도로, 다양한 이미지 처리 작업에서 우수한 성능을 보이고 있다. 그러나 이러한 연구들은 주로 모델의 구조적 변형에 중점을 두고 있어, 이미지 간의 상호작용을 충분히 반영하지 못하는 한계를 가진다. 본 연구는 ViT의 구조적 변형과 함께 유사 이미지 비교를 통한 어텐션 강화 기법을 결합하여 보다 향상된 성능을 달성하고자 한다.

Ⅲ. 방법론

3.1. 데이터 및 전처리

본 연구에서는 [5]의 데이터를 사용하였으며, 해당 데이터는 2019년부터 2021년까지 태국 방콕의 King Chulalongkorn Memorial Hospital에서 수집된 3,108개의 시계 그림 이미지로 구성되어 있다. 참가자 연령은 29세에서 90세 사이였으며, 성비는 3:1(여성 : 남성)이다. MoCA 평가에서 참가자들은 11시 10분을 가리키는 시계를 그렸다.

수집된 이미지는 256×256 픽셀로 크기가 조정되었으며, 데이터 다양성을 확보하기 위해 밝기, 대비, 크기 조

정 등의 이미지 증강 기법을 적용하였다. 이때, 시계의 정확한 스코어링에 영향을 줄 수 있는 회전 및 반전 변형은 제외되었다. 최종적으로 3,108개의 원본 이미지를 23배 증강하여 총 71,484개의 이미지를 생성하였다. 실험에서는 71,484개 이하의 이미지를 사용했을 때 성능이 충분하지 않았으나, 71,484개의 이미지를 사용했을 때 성능이 크게 향상되었으며, 이후 더 많은 이미지를 생성해도 성능 차이는 거의 없었다.

3.2. Pairwise Attention Reinforcement(PAR)

본 연구에서는 두 개의 이미지를 무작위로 선택하여 하나를 학습 이미지로, 다른 하나를 참조 이미지로 설정하고 두 이미지의 유사성을 활용하여 어텐션을 강화하고 강화된 어텐션을 이용하여 학습을 수행한다. 학습 이미지는 분류를 위한 학습에 사용되며 참조 이미지는 학습 이미지의 어텐션을 강화하기 위해 참조되는 이미지이다. 이를 통해 모델은 유사한 이미지 쌍의 미세한 차이를 학습할 수 있다. PAR 모델은 ViT 기반으로 구성되었다.

그림 1은 PAR 모델의 구조를 나타낸다. 이 모델은 사전 학습된 ViT를 사용하며, 어텐션 강화는 마지막 블록에서 이루어진다. 마지막 블록은 고차원적인 특징을 학습하여 이미지의 핵심 패턴을 더 잘 이해하게 된다. 그림 1의 (a)와 (c)는 참조 이미지와 학습 이미지의 입력 과정을 각각 보여주며, (b)는 PAR 모델의 핵심 알고리즘을 설명한다. ①번은 두 이미지의 클래스 토큰 유사도와 어텐션 스코어 맵 유사도를 확인하는 과정, ②번은 매칭되는 부분을 식별하는 과정, ③번은 학습 이미지의 어텐션 스코어 맵을 강화하는 과정을 나타낸다. 각 과정에 대한 자세한 설명은 그림 2,3 에서 확인할 수 있다.클래스 토큰은 전체 이미지 분류를 위해 학습되는 파라미터이다. 그림 2는 클래스 토큰의 코사인 유사도(CCS)를 나타내며, 오른쪽은 벡터 평면에 나타낸 클래스 토큰 벡터를 보여준다. 두 이미지의 CCS가 높을수록 구조나 주요 요소가 유사하다고 판단된다. CCS를 구하는 공식은 (1)과 같다:

$$CCS = \frac{V_R \cdot V_L}{\| V_R \| \| V_L \|} \quad (1)$$

여기서 V_R과 V_L은 각각 참조 이미지와 학습 이미지의 클래스 토큰 벡터이다. 분자는 두 벡터의 내적을 나타내고, 분모는 각 벡터의 노름(norm)을 곱한 값으로, 이는 두 벡터 간의 코사인 유사도를 나타낸다.

어텐션 스코어 맵의 코사인 유사도(ASCS)가 높을수록 두 이미지의 어텐션 스코어 맵이 유사한 것으로 판단되며, ASCS 구하는 공식은 (2)와 같다:

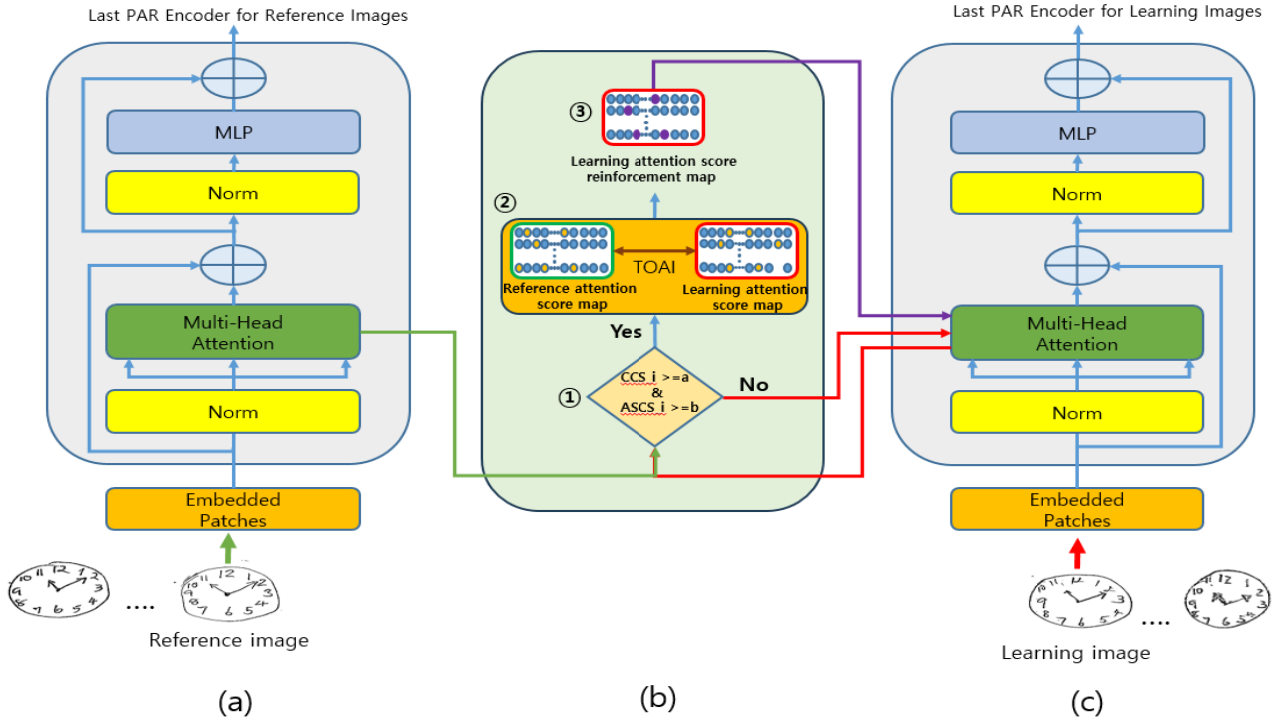

Fig. 1. Pairwise Attention Reinforcement (PAR) structure. CCS: Class token Cosine Similarity, ASCS: Attention Score map Cosine Similarity

그림 1. 쌍별 어텐션 강화(PAR) 구조. CCS: 클래스 토큰 코사인 유사도, ASCS: 어텐션 스코어 맵 코사인 유사도

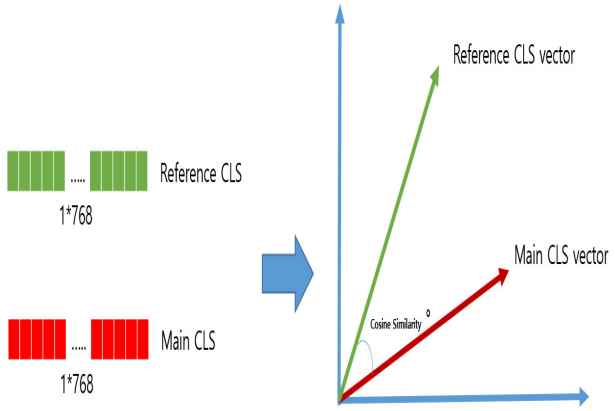

Fig. 2. CCS (Class token Cosine Similarity).
그림 2. CCS(클래스 토큰 코사인 유사도)

$$ASCS = \frac{1}{N}\sum_{i=1}^{N}\frac{a_R^i \cdot a_L^i}{\| a_R^i \| \ \| a_L^i \|} \quad (2)$$

여기서 a_R^i와 a_L^i은 각각 참조 이미지와 학습 이미지의 i번째 패치의 어텐션 스코어 맵이며, N은 패치의 수를 나타낸다. ASCS는 N개 패치의 코사인 유사도의 평균이다. 수식 1과 수식 2에서 나온 값이 그림 1의 ①번에 나타난 임계값 a와 b 보다 크면 ②로 넘어가고, 그렇지 안으면 학습 이미지의 어텐션 스코어가 강화되지 않는다. 임계값 a,b는 다양한 값으로 실험하였으며 그 결과는 표 2에 있다.

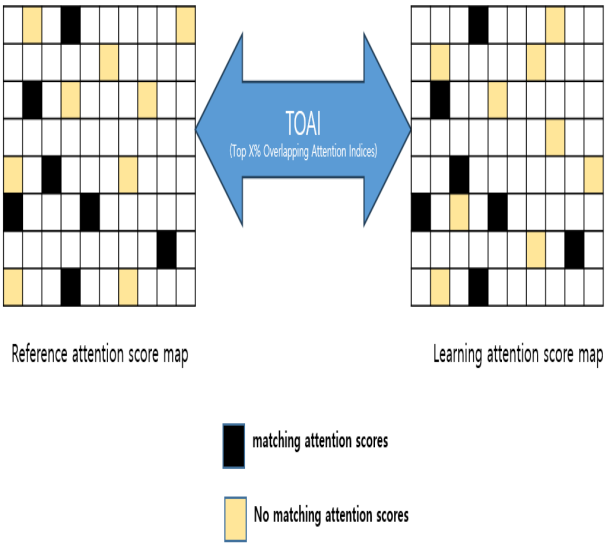

Fig. 3. TOAI (Top X% Overlapping Attention Indices).
그림 3. TOAI(상위 X% 중첩 어텐션 인덱스)

그림 3는 그림 1의 ②부분을 나타낸다. 그림 3는 두 이미지 간의 어텐션 스코어 맵을 비교하는 TOAI(Top X% Overlapping Attention Indices)를 보여준다. TOAI는 학습 이미지와 참조 이미지의 각 어텐션 맵에서 상위 X%의 값의 위치를 저장하고 두 어텐션 맵에서 저장된 위치가 매칭되는 영역을 식별하는 방법이다.

TOAI를 통해 매칭되는 영역이 식별된 후, 학습지 이미지의 맵에서 매칭 영역에 대한 스코어 가 강화된다. 이 과정은 그림 1의 ③번 과정에서 수행된다.

PAR 모델 아키텍처는 CDT(Clock Drawing Test) 데이터를 이용해 인지 기능의 저하를 분류 함으로서 성능을 입증하였다.

Ⅳ. 실험 결과

Table 1. Performance analysis according to CCS and ASCS thresholds.
표 1. CCS 및 ASCS 임계값에 따른 성능 분석

CCS, ASCS	Precision	Recall	F1-Score	Accuracy
CCS=0.9 ASCS=0.9	0.9366	0.8883	0.9116	92.22%
CCS=0.8 ASCS=0.8	0.9466	0.8683	0.905	92.45%
CCS=0.7 ASCS=0.7	0.9516	0.8883	0.9166	92.93%
CCS=0.6 ASCS=0.6	0.93	0.9033	0.915	92.86%
CCS=0.5 ASCS=0.5	0.9433	0.8933	0.9166	92.72%

표 1은 클래스 토큰 코사인 유사도(CCS)와 어텐션 스코어 맵 코사인 유사도(ASCS) 임계값에 따른 성능 변화를 보여준다. 실험 결과, CCS와 ASCS 임계값을 0.7로 설정했을 때 Precision 0.9516, Recall 0.8883, F1-Score 0.9166, Accuracy 92.93%로 가장 높은 성능을 기록했다. 이러한 결과는 Precision과 Recall의 균형을 최적화하여 우수한 성능을 발휘한 것으로 해석된다.

특히, 0.7 임계값에서는 Precision과 Recall이 적절히 조화되면서 F1-Score와 Accuracy도 최대치를 기록했다. 반면, 0.9 임계값에서는 Precision이 약간 상승했으나, Recall과 F1-Score, Accuracy는 다소 감소하여 전반적인 성능이 저하되었다. 따라서 0.9 이상의 높은 임계값은 성능을 떨어뜨릴 수 있음을 확인할 수 있다.

이러한 결과를 종합해 볼 때, CCS와 ASCS 임계값을 0.7로 설정하는 것이 Precision과 Recall의 균형을 최적화하고, 모델 성능을 극대화하는 데 가장 적합한 설정임을 확인할 수 있다. 따라서, 본 연구에서는 CCS와

ASCS 임계값 0.7을 최종적으로 채택하여 실험을 진행하였다.

Table 2. Performance analysis by TOAI threshold.
표 2. TOAI 임계값에 따른 성능 분석

TOAI	Precision	Recall	F1-Score	Accuracy
10%	0.935	0.8966	0.9133	92.61%
20%	0.9516	0.8883	0.9166	92.93%
30%	0.8783	0.895	0.885	91.92%
40%	0.9333	0.8816	0.905	92.30%
50%	0.92	0.895	0.9066	92.33%

표 2는 TOAI(Top X% Overlapping Attention Indices) 임계값에 따른 성능을 보여준다. TOAI를 20%로 설정했을 때, Precision 0.9516, Recall 0.8883, F1-Score 0.9166, Accuracy 92.93%로 가장 높은 성능을 기록했다. 이는 20% 설정이 Precision과 Recall의 최적 균형을 제공했음을 의미한다. 반면, 10% 또는 30% 이상의 설정에서는 성능이 상대적으로 낮아졌다. 이 결과는 TOAI 20% 설정이 성능 최적화를 위한 가장 효과적인 선택임을 시사한다.

본 연구에서는 PAR(Pairwise Attention Reinforcement) 모델의 성능을 평가하기 위해 ResNet-152, VGG 16, DenseNet-121, API-Net, 그리고 ViT 모델들과 비교 실험을 수행하였다. 표 3에서 확인할 수 있듯이, PAR 모델이 기존 모델들에 비해 전반적으로 성능이 향상되었음을 알 수 있다.

PAR 모델은 다양한 설정을 통해 평가되었으며, 인코더 레이어 크기(B, L), 패치 크기(16, 32), 입력 이미지 해상도(224, 384) 등의 변수를 조정하여 성능을 측정하였다. 예를 들어, PAR_B_16_224 모델은 16×16 크기의 패치와 224×224 해상도의 이미지를 사용하는 기본 크기의 모델이다.

특히, API-Net은 Pairwise Attention 기법을 사용하는 모델로 성능 비교의 기준이 되었으나, PAR 모델은 모든 성능 지표에서 API-Net을 능가하는 성능을 보였다. 예를 들어, PAR_B_16_224 모델은 Precision(정밀도)에서 0.9516을 기록하며, API-Net에 비해 약 16% 성능이 향상되었다. 또한, Accuracy(정확도)에서는 API-Net의 80.33%와 비교해 92.93%로 약 12% 더 높은 성능을 기록했다. ResNet-152(76.68%), VGG16(76.68%), DenseNet-121(77.40%)과 비교했을 때도 약 16% 이상의 성능 차이가 있었다.

ViT(Vision Transformer) 모델들과 비교하면, PAR_B_16_224 모델은 ViT_B_16_224 모델의 91.76%보다 약 1% 높은 92.93%의 정확도를 기록했다. 또한, ViT_L_32_384 모델의 91.89%와 비교했을 때도 약 1% 더 우수한 성능을 보였다. PAR 모델은 ViT 모델 전반에 걸쳐 약 1~2% 성능 향상을 보여주었다.

V. 결론

본 연구에서는 ViT 모델의 한계를 극복하기 위해 PAR 모델을 제안하였다. PAR 모델은 유사 이미지 간의 미세한 차이를 더 효과적으로 감지할 수 있도록 클래스 토큰 코사인 유사도와 어텐션 스코어 맵 코사인 유사도를 활용

Table 3. PAR model performance comparison with other models.
표 3. PAR 모델과 다른 모델 성능 비교

models	Precision	Recall	F1-Score	Accuracy
ResNet-152	0.7654	0.7668	0.7581	76.68%
VGG16	0.7628	0.7668	0.7608	76.68%
DenseNet-121	0.7764	0.7740	0.7708	77.40%
API-Net	0.8028	0.8033	0.8013	80.33%
Vit_B_16_224	0.9266	0.87	0.8966	91.76%
Vit_L_16_224	0.9216	0.8566	0.8883	91.72%
Vit_B_32_384	0.9183	0.86	0.885	91.31%
Vit_L_32_384	0.9266	0.8733	0.8983	91.89%
PAR_B_16_224	0.9516	0.8883	0.9166	92.93%
PAR_L_16_224	0.9466	0.8866	0.915	92.57%
PAR_B_32_384	0.94	0.89	0.9133	92.56%
PAR_L_32_384	0.9333	0.8933	0.9116	92.81%

하여 어텐션 스코어를 강화하였다. 다양한 실험 결과, PAR 모델은 기존 ViT 모델에 비해 전반적으로 우수한 성능을 나타냈으며, 특히 Precision, Recall, F1-Score, Accuracy에서 중요한 성능 향상을 달성하였다.

PAR 모델의 우수한 성능은 특징 매칭 어텐션 강화 기법이 이미지 간의 미세한 차이를 더 정밀하게 구분할 수 있도록 도움을 주었음을 시사한다. 이러한 성능 향상은 정밀한 분류 작업에서 특히 유용할 것으로 기대된다. 본 연구의 결과는 ViT 기반 모델의 성능을 향상시키는 새로운 접근법으로서, 다양한 컴퓨터 비전 응용 분야에서의 활용 가능성을 보여준다.

References

[1] A. Vaswani et al., "Attention is all you need," *in Adv. Neural Inf. Process.* Syst., 2017, vol.30, pp.5998-6008. DOI: 10.48550/arXiv.1706.03762

[2] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:*2010.11929, 2020. DOI: 10.48550/arXiv.2010.11929

[3] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," *in Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp.10012-10022. DOI: 10.1109/ICCV48922.2021.00986

[4] C. F. Chen, Q. Fan, and R. Panda, "CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification," *in Proc. IEEE/CVF Int. Conf. Comput. Vis.,* 2021, pp.10012-10022. DOI: 10.48550/arXiv.2103.14899

[5] R. Raksasat, S. Teerapittayanon, S. Itthipuripat, K. Praditpornsilpa, A. Petchlorlian, T. Chotibut, and I. Chatnuntawech, "Attentive pairwise interaction network for AI-assisted clock drawing test assessment of early visuospatial deficits," *Sci. Rep.*, vol.13, no.1, p.18113, 2023. DOI: 10.1038/s41598-023-44723-1

[6] S. Chen et al., "Automatic dementia screening and scoring by applying deep learning on clock-drawing tests," *Sci. Rep.*, vol.10, no.1, p.20854, 2020. DOI: 10.1038/s41598-020-74710-9

[7] J. Yao et al., "Extended Vision Transformer (ExViT) for Land Use and Land Cover Classification: A Multimodal Deep Learning Framework," *IEEE Transactions on Geoscience and Remote Sensing,* 2023. DOI: 10.1109/TGRS.2023.3284671

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:*1409.1556, 2014. DOI: 10.48550/arXiv.1409.1556

[9] K. He et al., "Deep residual learning for image recognition," *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp.770-778. DOI 10.1109/CVPR.2016.90

[10] H. Inoue, "Data augmentation by pairing samples for images classification," *arXiv preprint arXiv:*1801.02929, 2018. DOI: 10.48550/arXiv.1801.02929

[11] D. Yarats, I. Kostrikov, and R. Fergus, "Image augmentation is all you need: Regularizing deep reinforcement learning from pixels," *in Int. Conf. Learn. Represent.*, 2021. DOI: 10.48550/arXiv.2004.13649

[12] H. Zhu, W. Ke, D. Li, J. Liu, L. Tian, and Y. Shan, "Dual cross-attention learning for fine-grained visual categorization and object re-identification," *in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.,* 2022, pp.4692-4702. DOI: 10.48550/arXiv.2205.02151

[13] X. Peng et al., "Optical Remote Sensing Image Change Detection Based on Attention Mechanism and Image Difference," *IEEE Transactions on Geoscience and Remote Sensing,* vol.59, no.9, pp.7426-7440, Sep. 2021. DOI: 10.1109/TGRS.2020.3033009

[14] S. Mehta and M. Rastegari, "MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer," *arXiv preprint arXiv:*2110.02178, 2021. DOI: 10.48550/arXiv.2110.02178

[15] M. Dehghani et al., "Patch n' Pack: NaViT, a Vision Transformer for any Aspect Ratio and Resolution," *arXiv preprint arXiv:*2307.06304, 2023. DOI: 10.48550/arXiv.2307.06304

[16] K. Xu, P. Deng, and H. Huang, "Vision Transformer: An Excellent Teacher for Guiding Small Networks in Remote Sensing Image Scene

Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol.60, pp.1-15, 2022. DOI: 10.1109/TGRS.2022.3152566
[17] T. Stegmüller, B. Bozorgtabar, A. Spahr, and J. P. Thiran, "Scorenet: Learning non-uniform attention and augmentation for transformer-based histopathological image classification," *in Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2023, pp.6170-6179.
[18] W. Wang et al., "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," *in Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 568-578. DOI: 10.1109/ICCV48922.2021.00061
[19] S. Amini et al., "An AI-assisted online tool for cognitive impairment detection using images from the clock drawing test," *MedRxiv,* 2021. DOI: 10.1101/2021.03.06.21253047
[20] Q. Chen, J. Fan, and W. Chen, "An improved image enhancement framework based on multiple attention mechanism," *Displays*, vol.70, pp.102091, 2021. DOI: 10.1016/j.displa.2021.102091

BIOGRAPHY

Yeong-Seok Ju (Member)

2013 : BS degree in Multimedia Engineering, Nazareth University.
2018 : MS degree in Multimedia Engineering, Kongju University.
2021~present : PhD course in Computer Engineering, Gachon University

Joon-Shik Lim (Member)

1986 : BS degree in computer engineering, Inha University.
1989 : MS degree in computer engineering, ALABAMA University.
1994 : PhD degree in computer engineering, LOUISIANA STATE University.
1995~Present: Professor, Department of Computer Engineering, Gachon University
Areas of Interest: neuro-fuzzy systems, biomedical systems, Artificial Intelligence