녹내장 진단을 위한 ViT 기반 모델의 해석 가능성 비교: CNN과의 시각적 해석력 차이 분석 및 Occlusion, Grad-CAM 기법 활용

유호명, 하정욱*
서강대학교

hmyou6895@sogang.ac.kr, *jwha6169@sogang.ac.kr

Interpretability Comparison of ViT-Based Models for Glaucoma Diagnosis: Visual Explainability Analysis with CNNs Using Occlusion and Grad-CAM

You Ho Myung, Jeung Uk Ha*

Sogang University

요약

녹내장은 조기 진단이 필수적인 주요 실명 원인이다. 그러나 기존 딥러닝 연구는 Grad-CAM이나 Attention Map에 의존하여 해석 가능성이 정성적 수준에 머물렀다. 본 연구는 CNN 기반 MobileNetV3-Large와 ViT 기반 MobileViT-S를 동일 조건에서 학습시키고, Grad-CAM과 Occlusion을 결합한 정량적 지표(Consistency, ROI Energy Ratio, Quantitative Drop)를 통해 비교하였다. 단순 시각적 해석에 그치지 않고 지표별 유의미성을 통계적으로 검증한 결과, ViT는 시신경 유두에 국한되지 않고 RNFL을 포함한 전역적 패턴을 활용하는 경향을 보였다.

Ⅰ. 서 론

녹내장은 전 세계적으로 발생하는 실명의 주요 원인 중 하나로, 조기 진단과 적절한 치료 여부에 따라 시력 보존 가능성이 크게 달라진다. 최근 안저(fundus) 사진을 활용한 딥러닝 기반 진단 모델이 임상 현장에 도입되면서 진단의 정확도와 효율성이 크게 향상되었다. 그러나 여러 임상 연구에서는 시신경 유두(optic disc) 영역만을 기반으로 한 진단에 한계가 있을 수 있음을 보고하고 있으며, 보다 안정적이고 신뢰할 수 있는 진단을 위해 시신경 유두뿐 아니라 섬유층(retinal nerve fiber layer; RNFL)이나 주변 혈관 구조 등 전역적인 패턴의 동반 분석이 권장된다[1][2].

인공지능 학회에서도 CNN(Convolutional Neural Network)과 Vision Transformer(ViT) 계열 모델이 최근 활발히 연구되고 있으나, 대부분의 연구는 Grad-CAM이나 Attention Map을 이용해 모델이 주목하는 영역을 정성적(qualitative)으로만 해석하는 수준에 머물렀다[3]. 이러한 한계로 인해 녹내장을 진단하는 모델이 집중하는 패턴을 정량적으로 비교·분석하고, 임상적으로 해석 가능한 결과를 도출하는 데는 한계가 존재했다. 본 연구는 이러한 문제를 해결하기 위해 동일한 학습 조건하에서 VNet(MobileNetV3-Large)과 ViT(MobileViT-S) 모델을 학습한 뒤, Grad-CAM과 Occlusion 기법을 결합하여 정량적인 분석 프레임워크를 설계하였다. Consistency(Cosine, Pearson), ROI Energy Ratio, Quantitative Drop(절대·상대) 등의 지표를 도입해 내부(AIROGS-Light)와 외부(ORIGA) 데이터셋에서 모델의 주목 패턴을 비교하고, ViT 기반 모델의 전역적 패턴 학습 특성과 우수한 일반화 성능을 실험적으로 검증하였다.

Ⅱ. 본론

2.1 데이터셋

본 연구에서는 세 가지 데이터셋을 활용하였다.

- 내부 데이터셋 (AIROGS-Light)

AIROGS 공개 데이터셋을 기반으로 Kaggle 플랫폼에서 약 8,000장의 훈련 이미지, 770장의 검증 이미지, 770장의 테스트 이미지를 다운로드 받은 후 전처리하여 실험에 활용하였다[4][9].

- 외부 데이터셋Ⅰ (ORIGA)

Kaggle 플랫폼에서 제공되는 ORIGA 공개 데이터셋을 동일한 전처리 조건으로 변환하여 사용하였다[5][10].

- 외부 데이터셋 Ⅱ(REFUGE2)

Kaggle 플랫폼에서 제공되는 REFUGE2 공개 데이터셋을 동일한 전처리 조건으로 변환하여 사용하였다[11][12].

2.2 모델 구조 및 학습 설정

- VNet (MobileNetV3-Large)

MobileNetV3-Large는 CNN 기반의 경량 모델로, 모바일 환경에서 효율성과 속도를 고려한 구조를 특징으로 한다[6]. 학습 코드 구현은 Kaggle Notebook PyTorch Easy Setup for Glaucoma Detection; 92.6%을 참고하여 수정하였다 [7].

- ViT (MobileViT-S)

Transformer 기반 경량 모델로, 전역적 관계 학습과 국소적-지역적 융합 구조를 통해 효율성과 표현력을 동시에 확보한다[8].

두 모델 모두 ImageNet 사전 학습 가중치를 사용하며, 입력 크기 512*512, 배치 크기 16, 학습률 1e-3, StepLR 스케줄러, 11 epochs, Adam 옵티마이저, 동일한 데이터 증강 기법으로 학습하여 비교의 공정성을 확보하였다. 단, 시드는 고정하지 않았다.

2.3 설명 가능성 분석

- Consistency (Cosine, Pearson): Grad-CAM과 Occlusion 맵 간의 유사도를 측정하는 지표

$$Consistency_{\cos}(G,O) = \frac{\langle G, O \rangle}{\| G \| \, O \|}$$

$$Consistency_{pearson}(G,O) = \frac{Cov(G,O)}{\sigma_G \sigma_O}$$

- ROI Energy Ratio: 상위 15% 활성화 영역이 Grad-CAM 맵에서 차지하는 에너지 비율

$$EnergyRatio(G) = \frac{(\Sigma_{i \in ROI} G_i)}{(\Sigma_i G_i)}$$

- Quantitative Drop (Q-Drop, 절대·상대): Grad-CAM 상위 영역을 마스킹했을 때 모델 출력 확률의 절대 및 상대 감소량

$$QDrop_{abs} = \max(0, P(y \mid x) - P(y \mid x^{ROI-masked}))$$

$$QDrop_{rel} = \max(0, \frac{P(y|x) - P(y|x^{ROI-masked})}{P(y|x)})$$

Grad-CAM은 각 모델의 마지막 컨볼루션 층(VNet: features 블록의 마지막 Conv, MobileViT-S: conv_head)을 타깃으로 하여 heatmap을 생성하였다. 이는 class-specific activation이 가장 잘 반영되는 층으로, 두 모델 간 주목 패턴을 공정하게 비교하기 위함이다. Occlusion의 패치 크기(patch size)는 32, 64, 96으로 변화시키며 패치 민감도 실험을 수행했고, 각 조건에서의 Consistency 차이를 Wilcoxon signed-rank 검정과 Holm-Bonferroni 보정으로 분석하였다.

2.4 실험 결과

2.4.1 분류 성능

내부 및 외부(ORIGA+REFUGE2 combined) 데이터셋에서의 성능은 [Table 1] 과 같다.

[Table 1] 내부·외부 데이터셋에서의 모델 분류 성능 비교

Model	Dataset	ROC-AUC	Accuracy	F1-score	AUPRC
VNet	Internal	0.9834	94.03%	0.9404	0.9825
ViT		0.9828	93.77%	0.9383	0.9809
VNet	External	0.8850	86.00%	0.6441	0.7073
ViT		0.9032	86.10%	0.6667	0.7355

내부 데이터셋에서는 두 모델 모두 높은 정확도와 AUC를 기록했으나, 외부 데이터셋에서는 ViT가 소폭 우수한 성능을 나타냈다.

2.4.2 정량적 해석 가능성 비교

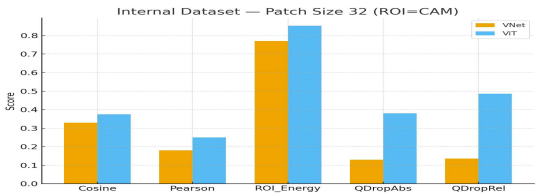

그림 1 내부 데이터셋 정량 지표 (Grad-CAM+Occlusion, n=200)

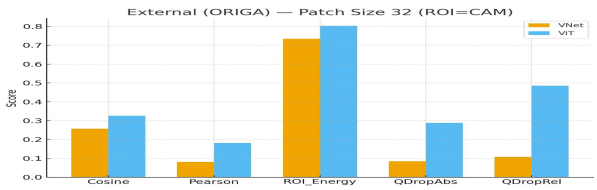

그림 2 외부(ORIGA) 데이터셋 정량 지표 (Grad-CAM+Occlusion, n=200)

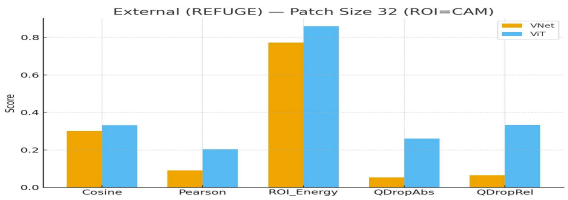

그림 3 외부(REFUGE2) 데이터셋 정량 지표 (Grad-CAM+Occlusion, n=200)

내부 및 외부 데이터셋에 대해 Grad-CAM과 Occlusion 기반 지표를 산출한 결과, ViT는 전반적으로 VNet보다 높은 값을 기록하였다. Consistency(Cosine, Pearson)는 ViT가 대체로 더 높은 경향을 보였으나, 특정 영상(예: 강한 빛반사)에서는 Grad-CAM과 Occlusion 맵이 동시에 비정상적으로 반응하여 지표 해석이 모호해지는 사례가 관찰되었다. ROI Energy Ratio 또한 ViT에서 일관되게 높게 나타나 전역적 패턴 활용을 시사했으나, 단독 지표로는 제한적인 해석에 머물 수 있다. 반면 Quantitative Drop(절대·상대)은 내부(AIROGS-Light)와 외부(ORIGA, REFUGE2) 모두에서 큰 차이를 보였으며, ViT가 ROI 영역 차단 시 출력 확률이 뚜렷하게 감소하였다. 이는 ViT가 단순히 국소적 OD 영역이 아닌 RNFL을 포함한 전역적 특징에 의존함을 강하게 시사한다.

2.4.3 패치 민감도 및 통계 검정

Occlusion 패치 크기를 32, 64, 96으로 변화시켜 분석한 결과, 전반적으로 ViT의 우위가 유지되었다. 그러나 지표별 신뢰도에는 차이가 있었다. Consistency(Cosine, Pearson)는 일부 조건에서 ViT가 더 높은 값을 보였으나, 내부 데이터셋에서는 통계적 유의성이 제한적이었고(예: Cosine, $p \approx 0.068$), 외부 REFUGE2에서는 차이가 뚜렷하지 않았다($p \approx 0.34$). ROI Energy Ratio는 내부·외부 모두에서 ViT가 높았으나, 효과 크기(Cohen's $d \approx 0.26 - 0.37$)는 상대적으로 작아 전역적 패턴 활용을 보조적으로 시사하는 수준이었다. 반면 QDrop(절대·상대)은 내부와 외부 전부에서 $p < 10^{-6}$ 수준의 강한 유의성과 큰 효과 크기(Cohen's $d \approx 0.60 - 0.92$)를 보여, ViT의 전역적 주목 패턴을 가장 신뢰성 있게 반영하는 지표로 확인되었다.

2.4.4 시각적 분석

[그림 4]은 내부 데이터셋에서 VNet과 ViT 모델의 Grad-CAM 시각화 예시를 보여준다. VNet은 대부분의 케이스에서 시신경 유두(OD) 주변에 국소적으로 집중된 활성화를 보이는 반면, ViT는 OD를 중심으로 보다 넓게 감싸는 전역적 활성화를 나타냈다. [그림 5]와 같이 일부 사례에서는 ViT의 활성화가 망막신경 섬유층(RNFL) 방향으로 길게 확장되는 패턴이 관찰되었다.

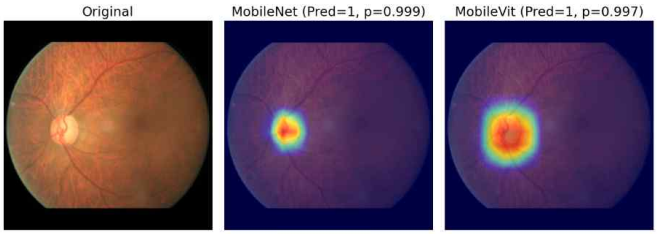

그림 4 ViT·VNet Grad-CAM (내부 데이터셋, 녹내장)

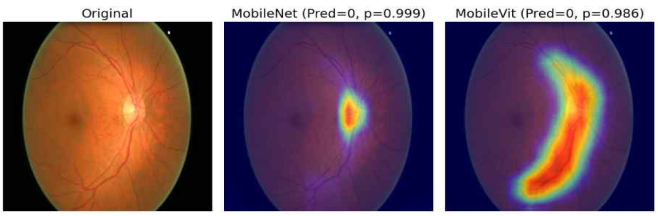

그림 5 ViT·VNet Grad-CAM (외부 데이터셋, 정상 안구)

Ⅲ. 결론

본 연구에서는 동일 조건하에서 학습된 CNN(MobileNetV3-Large, VNet)과 ViT(MobileViT-S, ViT) 모델의 시각적 해석 가능성을 정량적으로 비교하였다. Grad-CAM과 Occlusion을 결합해 여러 지표를 분석한 결과, ViT가 VNet보다 전역적 패턴을 더 폭넓게 활용하는 경향이 확인되었다. 다만 Consistency(Cosine, Pearson)는 특정 영상 조건(예: 빛반사 등)에서 해석이 왜곡될 수 있으며, ROI Energy Ratio 역시 보조적 지표로서 한계가 있었다. 반면 QDrop(절대·상대)은 내부·외부 데이터셋과 다양한 패치 크기에서 일관되게 유의한 차이를 보였으며, ViT의 전역적 주목 패턴을 가장 신뢰성 있게 반영하는 지표로 나타났다. 이러한 분석은 임상적으로 의미가 있으나, 실제 환자 데이터를 통한 후속 검증이 필요하다.

참 고 문 헌

[1] European Glaucoma Society, "Terminology and Guidelines for Glaucoma, 5th Edition," Br. J. Ophthalmol., vol. 105, suppl. 1, pp. 1-169, 2021, (https://doi.org/10.1136/bjophthalmol-2020-EGSguidelines).

[2] European Glaucoma Society, "Terminology and Guidelines for Glaucoma, 4th Edition - Summary," Br. J. Ophthalmol., vol. 101, no. 6, pp. 130-195, 2017, (https://doi.org/10.1136/bjophthalmol-2016-EGSguideline).

[3] Gu, B., Sidhu, S., Weinreb, R. N., Christopher, M., Zangwill, L. M., and Baxter, S. L., "Review of visualization approaches in deep learning models of glaucoma," Asia-Pacific Journal of Ophthalmology (Phila), vol. 12, no. 4, pp. 392-401, 2023.

[4] de Vente, C., Orlando, G., van Rijsbergen, J. J., et al., "AIROGS: Artificial Intelligence for Robust Glaucoma Screening Challenge," arXiv preprint arXiv:2302.01738, 2023, (https://arxiv.org/abs/2302.01738).

[5] Zhang, Z., Yin, F., Liu, J., Wong, W. K., Tan, N. M., Lee, B. H., Cheng, J., and Wong, T. Y., "ORIGA(-light): An online retinal fundus image database for glaucoma analysis and research," Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC), pp. 3065-3069, Aug. 2010, (https://doi.org/10.1109/IEMBS.2010.5626137).

[6] Howard, A., Sandler, M., Chen, B., Wang, W., Chen, L.-C., Tan, M., Chu, G., Vasudevan, V., Zhu, Y., Pang, R., Le, Q. V., and Adam, H., "Searching for MobileNetV3," Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), pp. 1314-1324, Oct. 2019.

[7] de Vente, C., "PyTorch Easy Setup for Glaucoma Detection; 92.6%," Kaggle Notebook, Feb. 2024, (https://www.kaggle.com/code/deathtrooper/pytorch-easy-setup-for-glaucoma-detection-92-6).

[8] Mehta, S., and Rastegari, M., "MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer," Proc. Int. Conf. Learn. Representations (ICLR), Apr. 2022, (https://arxiv.org/abs/2110.02178).

[9] de Vente, C., "Glaucoma Dataset - EyePACS (AIROGS-Light v2)," Kaggle Dataset, Feb. 2024, (https://www.kaggle.com/datasets/deathtrooper/glaucoma-dataset-eyepacs-airogs-light-v2).

[10] ayush02102001, "Glaucoma Classification Datasets (ORIGA)," Kaggle Dataset,(https://www.kaggle.com/datasets/ayush02102001/glaucoma-classification-datasets).

[11] Fang, H., et al., "REFUGE2 Challenge: A Treasure Trove for Multi-Dimension Analysis and Evaluation in Glaucoma Screening," arXiv preprint, 2022. (https://arxiv.org/abs/2202.08994).

[12] jayesh0vasudeva, "REFUGE2-2020 Classification Dataset," Kaggle Dataset, n.d., (https://www.kaggle.com/datasets/jayesh0vasudeva/refuge2020-classificaton).