# 망막질환에서의 멀티모달 설명을 위한 객체 검출 라벨 활용 연구

조정래 [1], 박동호 [2], 정성문 [1,3*]
[1] 경북대학교병원 의료인공지능연구센터
[2] 경북대학교 의과대학 안과학교실
[3] 경북대학교 의과대학 의료정보학교실
{zzemb6, jeongsm00}@gmail.com

# Study on the Application of Object Detection Labels for Multimodal Explanations in Retinal Disease

Jungrae Cho[1], Dong Ho Park[2] and Sungmoon Jeong[1,3*]
[1]Research Center for Artificial Intelligence in Medicine, Kyungpook National University Hospital, Daegu, South Korea
[2]Department of Ophthalmology, School of Medicine, Kyungpook National University, Kyungpook National University Hospital, Daegu, South Korea
[3]Department of Medical Informatics, School of Medicine, Kyungpook National University, Daegu, South Korea

요 약

최근 몇 년간 딥러닝을 활용한 광간섭단층촬영 영상에서의 망막질환의 분류 연구가 활발히 진행되었다. 일부 연구는 속성 방법 기반의 설명을 제공하여 증거 기반 의학자들의 인공지능 분류 모델에 대한 신뢰도 향상을 도모했다. 그러나 단순 히트맵 형태의 설명은 활성화 위치와 강도만 제공하기에 그 표현력에 한계가 있었다. 본 연구는 망막 병변 바운딩 박스를 텍스트로 변환하여 이미지-텍스트 은닉 공간을 통한 멀티모달 설명 방법을 제안한다. 사전학습된 이미지와 텍스트 인코더를 망막질환 도메인 데이터셋에 미세 조정한 후, 매칭된 이미지 임베딩과 텍스트 임베딩을 입력 받은 각 분류기가 동일한 분류 결과를 도출하도록 일관성 손실을 분류 학습에 도입한다. 학습이 완료된 분류기의 출력으로부터 각각 이미지 속성맵과 텍스트 속성맵을 동시에 시각화한다. 본 연구는 망막질환의 지역적 병변 정보를 텍스트로 변환하여 이를 멀티모달 학습 패러다임에 편입할 뿐만 아니라 단순 종단학습 심층합성곱신경망보다 더 해석가능한 설명을 제공함을 실험을 통해 검증한다.

## Ⅰ. 서 론

최근 몇 년간 딥러닝을 활용한 광간섭단층촬영 (optical coherence tomography, OCT) 영상에서의 황반변성과 당뇨망막질환을 포함한 망막질환의 분류 연구가 활발히 진행되었다.[1] 일부 연구는 사용자인 망막전문의의 인공지능 모델에 대한 신뢰도를 높이기 위해 속성 방법(attribution method)을 통해 모델의 판단 근거를 히트맵 형태로 시각화했다.[2] 그러나 이러한 속성맵(attribution map)은 위치적인 활성화 수치만 표시하기 때문에 표현에 한계가 있었다.

대조 언어-이미지 사전학습 (Contrastive Language-Image Pretraining, CLIP)은 이미지와 텍스트를 같은 은닉 공간에 매칭되도록 학습하는 방법을 제시함으로써 이미지와 텍스트 임베딩 간의 호환성을 높였다.[3] 본 연구는 CLIP 의 이러한 이미지-텍스트 매칭 능력을 활용하여 위치적인 속성맵 뿐만 아니라 텍스트 속성 정보를 함께 제공하는 방법을 제공하여 모델 설명의 표현력을 높이는 방법을 제안한다. OCT 이미지 데이터셋에 포함된 황반변성 병변 라벨을 규칙 기반으로 텍스트로 변환한 후 사전학습된 (pre-trained) CLIP 을 OCT 이미지-텍스트 쌍에 미세 조정 (fine-tuning) 한다. 미세 조정된 CLIP 인코더로부터 이미지, 텍스트 임베딩을 추출한 후 이미지, 텍스트 계층을 각각 학습한다. 이미지, 텍스트 계층의 출력의 일관성을 위해 이미지, 텍스트 로짓 (logit) 사이의 크로스 엔트로피 (cross-entropy) 손실을 도입한다. 이렇게 학습된 분류기로부터 이미지 속성맵과 텍스트 속성 정보를 동시에 도출하여 기존 이미지 분류기보다 개선된 멀티모달 설명을 제공한다. 본 연구는 단순 종단학습 심층합성곱신경망과 제안 방법의 분류 정확도와 설명을 비교함으로써 제안 방법의 우수성을 검증한다.

## Ⅱ. 본론

**CLIP 도메인 미세 조정.** ResNet50[4]과 트랜스포머[5]로 구성된 사전학습된 CLIP 을 OCT 이미지 데이터셋에 미세 조정했다. Adam 옵티마이저를 사용하여 학습률(lr)은 5e-5 로, 베타 값은 각각 0.9 와 0.98 로 설정했다. 또한, epsilon 값은 1e-5 로, 가중치 감소율(weight decay)은 0.2 로 지정하여 모델의 파라미터를 최적화했다. 총 300 에폭 동안 미세 조정을 실시했다.

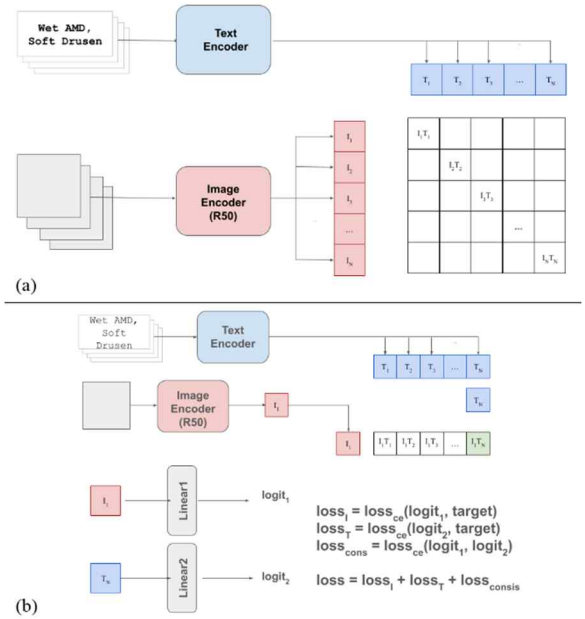


그림 1. 제안 모델 개요. (a) CLIP 도메인 미세 조정, (b) 분류 일관성 손실

**분류 일관성 손실**. 미세 조정된 CLIP 의 임베딩을 입력으로 받는 이미지, 텍스트 피드포워드 계층을 추가하여 황반변성 이진 분류 태스크 훈련을 수행했다. 훈련 손실은 이미지 출력 $\hat{y}_i$, 텍스트 출력 $\hat{y}_t$ 에 대한 분류 손실 $L_{ce}(\hat{y}_i, y)$ 과 $L_{ce}(\hat{y}_t, y)$, 그리고 이미지 출력과 텍스트 출력을 비교하는 일관성 손실 $L_{ce}(\hat{y}_i, \hat{y}_t)$ 을 추가하여 두개의 모달리티에 대해 분류 결과가 일정하게 최적화되도록 하였다. 모든 손실은 크로스 엔트로피 함수로 구성되었다. 최종 훈련 손실은 다음 수식과 같다.

$$L_{total} = \frac{1}{3}\left(L_{ce}(\hat{y}_i, y) + L_{ce}(\hat{y}_t, y) + L_{ce}(\hat{y}_i, \hat{y}_t)\right) \quad (1)$$

**멀티모달 설명.** 학습된 분류 계층에 대해 이미지 인코더의 경우 GradCAM [6] 기법을 사용하여 분류 로짓의 활성화 영역을 시각화 했다. 텍스트의 경우 GradCAM 과 유사한 방식으로 중요도가 높은 단어를 강조하도록 속성 정보를 시각화 했다.

**데이터셋.** OCT5K [7]는 1,672 장의 OCT 이미지로 구성된 공공 데이터셋이다. 이 중 정상 (healthy) 이미지 530 장, 황반변성 이미지 462 장, 총합 992 장의 이미지를 9:1:1 비율로 랜덤 분할하여 훈련 792 장, 검증 99 장, 테스트 99 장의 데이터셋을 구축했다. 황반변성의 경우 9 종의 병변 객체 라벨을 포함하고 있고, '<분류 클래스>, <병변 1>, <병변 2>, …'의 형태로 규칙 기반으로 텍스트를 생성했다. 텍스트의 종류는 총 78 종이 형성되었다.

**실험 결과**. 베이스라인 모델인 ResNet50 은 제안 모델과 같은 조건으로 백본 계층의 가중치는 동결한 채로 피드포워드 계층만 총 50 에폭, 학습률 1e-4 로 설정하여 훈련했다. 테스트셋에 대한 성능 평가는 표 1 과 같다. CLIP 의 이미지, 텍스트 분류기 모두 베이스라인 모델인 ResNet50 보다 개선된 정확도를 달성했다. 성능 비교는 표 1 과 같다.

표 1. 분류 테스트 성능 평가

| Model | Accuracy | Sensitivity | Specificity |
|---|---|---|---|
| ResNet50 | 0.74 | 0.58 | 0.87 |
| CLIP(Image) | 0.79 | 0.89 | 0.70 |
| CLIP(Text) | 0.75 | 0.89 | 0.62 |

그림 2 는 모델의 설명가능성에 대한 정성적인 평가 결과를 비교한 것으로, 베이스라인 (a)는 지역적 위치만 표시하는 반면 제안 모델 (b)는 지역적 위치에 더하여 텍스트 설명까지 제시하여 표현력을 향상시킨 것을 확인할 수 있다.

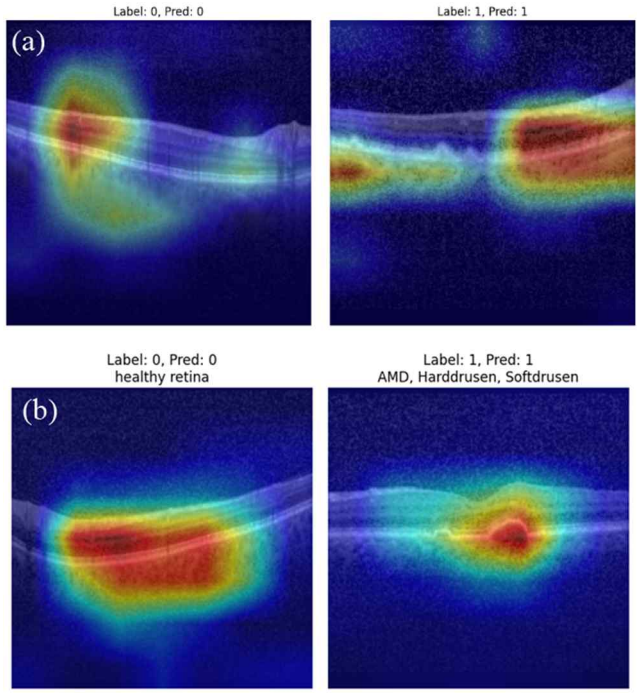


그림 2. 설명가능성 정성적 평가. (a) RssNet50 의 GradCAM 속성맵 (b) 제안 모델의 속성맵과 텍스트 설명

## Ⅲ. 결론

본 연구는 객체 검출 라벨을 텍스트로 변환하여 황반변성 분류를 위한 멀티모달 설명 방법을 제안했다. 제안된 방법은 도메인 적응된 CLIP 에 각 모달리티에 특화된 분류기와 분류 일관성 손실을 도입하여 OCT 이미지에 대한 이미지와 텍스트에 대한 일관성 있는 이미지 분류 학습 방법을 제시했다. 학습된 분류기로부터 이미지 속성맵과 텍스트 속성정보를 동시에 제공함으로써 기존의 이미지 기반의 분류 모델에 비해 좀더 다양한 정보를 제공하는 설명을 출력할 수 있게 되었다. 이러한 객체 검출 라벨의 활용방법은 현재 활발히 도입중인 거대 언어 모델(large language model)이나 기초 모델 (foundation model)에 이미지 데이터셋을 멀티모달 패러다임에 편입할 수 있는 가능성을 보여줬다. 향후 이미지-텍스트 대조 학습을 포함하여 거대 모델의 프롬프트로 활용되어 의료인공지능 분야에서 설명가능성을 확장할 수 있을 것으로 기대된다.


## ACKNOWLEDGMENT

This research was partly supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2024-2020-0-01808) supervised by the IITP(Institute of

Information & Communications Technology Planning & Evaluation) and a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number : HR22C1832).

## 참 고 문 헌

[1] Kurmann, T., Yu, S., Márquez-Neila, P., Ebneter, A., Zinkernagel, M., Munk, M. R., ... & Sznitman, R. (2019). Expert-level automated biomarker identification in optical coherence tomography scans. Scientific reports, 9(1), 13605.

[2] Singh, A., Sengupta, S., & Lakshminarayanan, V. (2020). Explainable deep learning models in medical image analysis. *Journal of imaging*, *6*(6), 52.

[3] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PMLR.

[4] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[5] Vaswani, A. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.

[6] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618-626).

[7] Arikan, M., Willoughby, J., Ongun, S., Sallo, F., Montesel, A., Ahmed, H., ... & Dubis, A. M. (2023). OCT5k: A dataset of multi-disease and multi-graded annotations for retinal layers. *bioRxiv*, 2023-03.