대조적 학습과 생성적 학습을 활용한 안저 이미지 분석을 위한 자가 지도 다중 모달 학습

Nguyen Duc Toan[1], 손소영[2], 추현승[1,3]
[1]성균관대학교 AI 시스템공학과 박사과정
[2]성균관대학교 전자전기컴퓨터공학과 박사과정
[3]성균관대학교 전자전기컴퓨터공학과 교수

austin47@g.skku.edu, alvy.sun@g.skku.edu, choo@skku.edu

Self-Supervised Multi-Modal Learning for Fundus Image Analysis Using Contrastive and Generative Learning

Toan Duc Nguyen[1], Sun Xiaoying[2], Hyunseung Choo[1,2]
[1]Dept. of AI Systems Engineering, Sungkyunkwan University
[2]Dept. of Electrical and Computer Engineering, Sungkyunkwan University

Abstract

In this study, we propose a self-supervised learning framework for fundus image processing, utilizing both contrastive and generative learning techniques for pre-training. Our contrastive learning approach integrates both image and text modalities through cross-attention mechanisms, allowing the model to learn more informative and semantically rich representations. After pre-training, the model is fine-tuned for downstream tasks, including zero-shot, few-shot, and full fine-tuning. Experimental results show that our method significantly outperforms existing approaches, achieving 15% higher performance in zero-shot, 4.5% in few-shot, and 10.1% in fine-tuning scenarios. The proposed method demonstrates its potential in the medical imaging field, where access to large annotated datasets is often limited. By efficiently leveraging both image and textual information, our approach contributes to improving the accuracy and generalizability of models in fundus image analysis, highlighting its broader applicability in medical diagnostics and healthcare.

1. Introduction

Medical imaging plays a crucial role in modern healthcare, facilitating the diagnosis and monitoring of various diseases [1]. Techniques such as magnetic resonance imaging (MRI), computed tomography (CT), and fundus photography are widely used for detecting and tracking the progression of diseases. In particular, conventional fundus imaging (CFI) is an essential tool for capturing detailed images of the retina, aiding in the diagnosis of ophthalmic conditions such as diabetic retinopathy, glaucoma, and age-related macular degeneration. Despite its significance, automatic analysis of CFI still presents challenges, particularly when large annotated datasets are unavailable, which is common in medical imaging domains.

In recent years, self-supervised learning has emerged as a promising approach to addressing the data scarcity problem by leveraging unlabelled data for pre-training models [2]. Two key methods in self-supervised learning are contrastive learning and generative learning. Contrastive learning works by bringing similar data points closer together in the latent space while pushing dissimilar points apart, allowing models to learn effective representations without supervision. Generative learning, on the other hand, focuses on learning the underlying data distribution to generate new samples, which helps in understanding the structure of the input data. Both methods have shown success in various fields, but their application in medical imaging, particularly in fundus analysis, remains underexplored.

In this work, we propose a self-supervised learning approach that incorporates multi-modal learning, specifically cross-modal learning between image and text. By using cross-attention mechanisms, we fuse information from both modalities, allowing the model to learn more robust and semantically enriched representations of fundus images. The image-text interaction enhances the model's ability to interpret visual data in context, which is particularly useful in medical imaging where textual descriptions such as clinical notes or labels can provide critical insights. The main

contributions of this paper are as follows:

1. We introduce a novel self-supervised learning framework that combines contrastive learning and generative learning for fundus image analysis.
2. We integrate cross-modal learning, utilizing both image and text modalities via cross-attention to enhance representation learning.
3. Our approach is evaluated on downstream tasks including zero-shot, few-shot, and fine-tuning, demonstrating superior performance over current state-of-the-art methods.
4. We show the applicability and usefulness of this approach in medical imaging, addressing the challenges posed by limited annotated data.

The rest of this paper is organized as follows: Section 2 reviews related work in self-supervised learning, multi-modal learning, and fundus image analysis. Section 3 outlines the methodology behind our proposed approach. Section 4 presents the experimental setup and datasets used for evaluation. Section 5 discusses the results, and finally, Section 6 concludes the paper with insights and directions for future research.

2. Methodology

In this work, we propose a novel self-supervised framework for fundus image processing that leverages multi-modal learning with both contrastive and generative objectives. Our approach combines image and text modalities through cross-attention mechanisms, enabling the model to learn rich, multi-modal representations that enhance downstream task performance. Our model, shown in Figure 1, consists of two primary branches: an image branch and a text branch. Each branch is composed of an encoder and a decoder network, allowing for both contrastive learning between the two modalities and generative learning within each modality. Specifically, the image branch processes fundus images using an image encoder that extracts feature embeddings from the input data. A portion of the input image is masked, similar to masked image modeling strategies, and the goal of the image decoder is to reconstruct the original image, imposing a generation loss. The encoder-decoder setup enables the model to learn visual representations of the retinal structure while understanding the context of the entire image through reconstruction.

2.1 Contrastive learning

In the contrastive learning phase, the model learns to align image and text embeddings using a contrastive loss function. Given a batch of paired images and textual descriptions, the model generates embeddings for both modalities using separate encoders: E_{img} for images and E_{text} for text. Let z_{img} be the image embedding and z_{text} be the text embedding:

$$z_{img} = E_{img}(x_{img}), z_{text} = E_{text}(x_{text}) \#(1)$$

where x_{img} is a fundus image, and x_{text} is the associated textual description. The goal is to minimize the distance between embeddings of matching image-text pairs while maximizing the distance between non-matching pairs. We use a contrastive loss, defined as:

$$\mathcal{L}_{contrastive} = -\frac{1}{N}\sum_{i=1}^{N} log\frac{exp(sim(z_{img}^{i}, z_{text}^{i})/\tau)}{\sum_{j=1}^{N} exp(sim(z_{img}^{i}, z_{text}^{j})/\tau)} \#(2)$$

where *sim* represents cosine similarity, τ is a temperature hyperparameter, and *N* is the number of samples in the batch.

2.2 Generative learning

The generative learning component consists of three tasks: image generation, image captioning, and cross-modal learning. Each of these tasks is handled through encoder-decoder architectures and aims to improve the model's ability to reconstruct masked data from the input.

2.2.1 Image generation

For image generation, we follow a masked image modeling approach, where parts of the input image are masked, and the model is tasked with reconstructing the masked regions. The image encoder, E_{img} processes the input image x_{img} and generates an embedding. The masked input image is denoted as $x_{img-mask}$. The decoder, D_{img} then reconstructs the original image, producing $\hat{x}_{img}$:

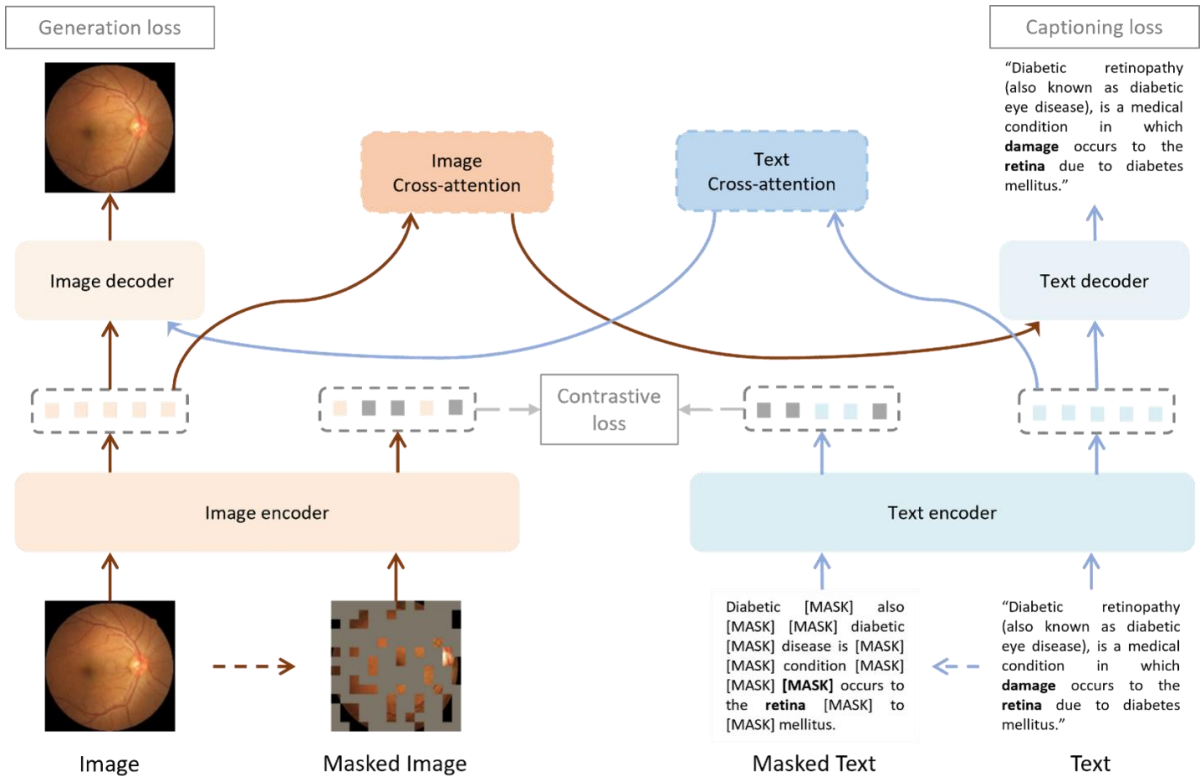
$$\hat{x}_{img} = D_{img}\left(E_{img}(x_{img-mask})\right) \#(3)$$

The model is then trained to minimize the image generation loss $\mathcal{L}_{img-gen}$ which is defined as the mean squared error (MSE) between the original and reconstructed images:

$$\mathcal{L}_{img-gen} = \frac{1}{N}\sum_{i=1}^{N} \left\| x_{img}^{i} - \hat{x}_{img}^{i} \right\|^2 \#(4)$$

2.2.2 Image captioning

For the text modality, the task is to generate captions from masked text inputs. The text encoder, E_{text} processes the masked clinical descriptions or captions. The masked input text is denoted as $x_{text-mask}$. The text decoder, D_{text}



(Figure 1) Overview of the proposed self-supervised learning framework for fundus image processing. The model consists of two branches: an image branch and a text branch, each with its own encoder-decoder architecture.

reconstructs the original text sequence, producing $\hat{x}_{text}$:

$$\hat{x}_{text} = D_{text}\left(E_{text}(x_{text-mask})\right)\#(5)$$

The captioning loss $\mathcal{L}_{caption}$ is computed as the cross-entropy between the predicted and true word tokens:

$$\mathcal{L}_{\text{caption}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} y_{\text{text}}^{i} \log(\hat{y}_{\text{text}}^{t})\#(6)$$

where y_{text}^{t} is the true token at time step t, and $\hat{y}_{text}^{t}$ is the predicted probability of that token.

2.2.3 Cross-modal learning

Cross-modal learning is achieved through cross-attention mechanisms that link the image and text modalities. The image embedding z_{img} is passed into the text cross-attention block, while the text embedding z_{text} is passed into the image cross-attention block. This allows the model to exchange information between the two modalities, enhancing its understanding of the relationships between visual and textual data. Let the cross-attention functions be denoted as CA_{img} for the image and CA_{text} for the text. The updated embeddings are:

$$z_{img-update} = CA_{img}\left(z_{img}, z_{text}\right),$$

$$z_{text-update} = CA_{text}\left(z_{img}, z_{text}\right)\#(7)$$

The total loss $\mathcal{L}_{total}$ is a combination of the contrastive loss, image generation loss, and captioning loss:

$$\mathcal{L}_{total} = \mathcal{L}_{contrastive} + \mathcal{L}_{img-gen} + \mathcal{L}_{caption}\#(8)$$

3. Experiments

To evaluate the effectiveness of our proposed self-supervised learning approach, we collected dataset of fundus images from multiple publicly available datasets. These datasets contain fundus images that capture a variety of retinal conditions, including diabetic retinopathy, glaucoma, and age-related macular degeneration, among others. The dataset includes a wide range of image resolutions and qualities, ensuring that the model generalizes well across different clinical environments. For the text modality, we associated each image with disease names and their corresponding definitions, using descriptions from medical literature and clinical guidelines. These textual annotations provide crucial context for the fundus images and serve as the input for the text branch of our model. By linking the visual and textual data, our model learns to associate retinal conditions with their descriptions, improving the robustness of the learned representations.

For the image branch, we utilize a Vision Transformer (ViT) [3] as the backbone architecture. The ViT model has

shown state-of-the-art performance in several vision tasks due to its ability to capture long-range dependencies in the image via self-attention mechanisms. The input to the image encoder is a fundus image of size 224×224, which is resized from the original resolution to match the model's requirements. The ViT architecture divides the input image into non-overlapping patches, each of size 16×16, and processes them through several transformer layers to obtain the image embeddings. The text branch uses a standard Transformer architecture to process textual descriptions of diseases. The input to the text encoder consists of disease names and their corresponding definitions.

4. Results

In the zero-shot setting, the model is evaluated on unseen tasks without further training or fine-tuning. As shown in Table 1, our model outperforms the baselines on the ODIR dataset, achieving an accuracy of 36.24%. Our model surpasses Clip (25.67%), Coca (33.56%), and Eva (32.43%), demonstrating a notable performance gain. In the few-shot setting, the model is fine-tuned with a small number of labeled samples. On the FIVES dataset, our model again shows superior performance compared to the baselines. Our method achieves an accuracy of 46.83%, outperforming Clip (42.33%), Coca (40%), and Eva (40%).

For fine-tuning, the model is fully trained on labeled data from the target task, and its performance is evaluated on the Aptos datasets. In this setting, where the model is trained on a fully labeled dataset, our proposed method achieves state-of-the-art performance on the Aptos dataset. The Kappa score of our model reaches 85.07, significantly outperforming Clip (74.05), Coca (74.96), and Eva (72.48). Detailed results are shown in Table 1.

<Table 1> Comparison of performance across zero-shot, few-shot, and fine-tuning tasks

Method	Zero-shot	Few-shot	Fine-tuning
Clip [4]	25.67	42.33	74.05
Coca[5]	33.56	40	74.96
Eva [6]	32.43	40	72.48
Ours	**36.24**	**46.83**	**85.07**

5. Conclusion

In this work, we presented a self-supervised learning framework for fundus image processing that leverages both contrastive and generative learning techniques. By incorporating multi-modal learning through cross-attention between image and text modalities, our approach effectively enhances the model's ability to learn rich representations from both visual and textual data. We demonstrated the effectiveness of our model across various tasks, including zero-shot, few-shot, and fine-tuning, achieving superior performance compared to state-of-the-art methods.

Our model's success in zero-shot learning highlights its strong generalization capability, which is crucial for medical imaging tasks where labeled data is scarce. Additionally, the few-shot and fine-tuning results further validate the robustness and adaptability of our approach, making it suitable for real-world clinical applications, particularly in retinal disease diagnosis. The key contributions of our work include a new self-supervised learning paradigm for fundus image analysis, the integration of cross-modal learning, and the demonstration of its effectiveness on multiple publicly available datasets. Moving forward, this framework can be extended to other medical imaging domains where multi-modal data is available, opening new possibilities for more accurate and efficient diagnostic systems in healthcare.

Acknowledgements

This work was partly supported by the BK21 FOUR Project, Korea government (MSIT), IITP, Korea, under the ICT Creative Consilience program (RS-2020-II201821, 50%), AI Innovation Hub (RS-2021-II212068, 25%), and AI Graduate School Program (Sungkyunkwan University, (RS-2019-II190421, 25%).

References

[1] Litjens, Geert, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I. Sánchez. "A survey on deep learning in medical image analysis." *Medical image analysis* 42 (2017): 60-88.

[2] Azizi, Shekoofeh, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh et al. "Big self-supervised models advance medical image classification." In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3478-3488. 2021.

[3] Dosovitskiy, Alexey. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv preprint arXiv:2010.11929* (2020).

[4] Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry et al. "Learning transferable visual models from natural language supervision." In *International conference on machine learning*, pp. 8748-8763. PMLR, 2021.

[5] Yu, Jiahui, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. "Coca: Contrastive captioners are image-text foundation models." *arXiv preprint arXiv:2205.01917* (2022).

[6] Fang, Yuxin, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. "Eva-02: A visual representation for neon genesis." *Image and Vision Computing* 149 (2024): 105171