서울대학교 대학원 융합전공 혁신의과학 전공

학부연구교육생 연구 보고서

연 구 기 간: 2024. 12. 23.(월) ~ 2025. 2. 14.(금)

연 구 제 목: Conditional Prompt Learning for Anomaly Detection in Medical Image Slices

성 명: 김도훈 (인 또는 서명)

지도교수 성 명: 이재성 (인 또는 서명)

(※ A4용지 10장 이내, 바탕체, 11포인트, 줄간격 160%. 아래 항목대로 기술하기 힘든 경우에는 해당 항목만을 기술함)

1. Abstract (영문의 경우 : 150단어 이내, 국문의 경우 : 750자 이내)

Explaining anomalies in medical images is a crucial task, not only for accurate medical diagnosis, but also for reliability between doctors and patients. Recently, CLIP based anomaly detection methods showed that prompt learning methods can represent abnormality and make segmentation map. However, existing methods did not focus on diverse adaptation, specifically in medical imaging modalities. In this work, following previous frameworks, we propose a conditional prompt learning approach to perform same tasks in diverse medical images.

2. Introduction

Zero-Shot Anomaly Detection (ZSAD) refers to a task to identify anomalies in unseen images during training[4], using auxiliary datasets. Recent advancements have demonstrated remarkable performance not only in industrial applications but also in medical imaging[4, 5], highlighting its potential for diverse domains. However, existing methods focus on an object-agnostic prompt learning approach, which can be inadequate for medical images since it cannot effectively handle domain shifts that frequently occur in medical imaging environments.

Many medical images, such as MRI, are inherently 3D. Existing studies often adopt a strategy of selecting specific slices from 3D images to utilize pretrained 2D CLIP[1] models. However, this approach has a notable limitation: even if all slices have same category name, they differ a lot. For instance, even when the same class label, such as "brain", is used, features visible in

axial slices may not appear in sagittal or coronal slices. Similarly, features evident in specific slices may not be present in other slices. This highlights a fundamental limitation of applying 3D images to 2D CLIP models.

To capture information from diverse slices without compromising CLIP's generalizability, it is essential to employ an object-conditioned prompt to represent this diverse shift. In classification task, CoOp[6] and CoCoOp[7] are studied, but conditional prompt learning is not reported in anomaly detection task. For that, we propose an approach that incorporates image-specific conditions into the prompt learning process. By tailoring the detection mechanism to the unique characteristics of medical images, our method achieves robust performance across multiple MRI modalities. Experimental results demonstrate that the proposed model outperforms conventional approaches, offering significant improvements in accuracy and adaptability. This research underscores the importance of condition-based anomaly detection in addressing the intricacies of medical imaging and sets a foundation for further exploration in this critical domain.

In summary, our contributions are as follows:

- We demonstrate that CLIP, through conditional prompt learning, can effectively understand and extract semantic details from diverse medical images without altering its pretrained parameters.
- Our approach eliminates the need for extensive fine-tuning, reducing computational costs while maintaining robust performance across various domains.

3. Method

3.1 Problem Definition

CLIP consists of two encoders: a Text encoder $T : R^{l} \rightarrow R^{d}$ with Transformer[2] architecture, and an Image encoder $I : R^{H \times W \times C} \rightarrow R^{d}$ with ViT[3] architecture, where l is the number of tokens in the text input, $H \times W \times C$ represent the height, width, and channels of the image, and d is the dimension of the shared embedding space. CLIP is trained via contrastive learning to align image and text representations in this shared multimodal embedding space.

Our main objective is to adapt CLIP to understand the concept of abnormalities in medical images, without altering its pretrained parameters. By leveraging prompt learning, we aim to enable CLIP to generalize across diverse medical imaging modalities, bypassing the need for costly fine-tuning or extensive labeled data.

3.2 Approach

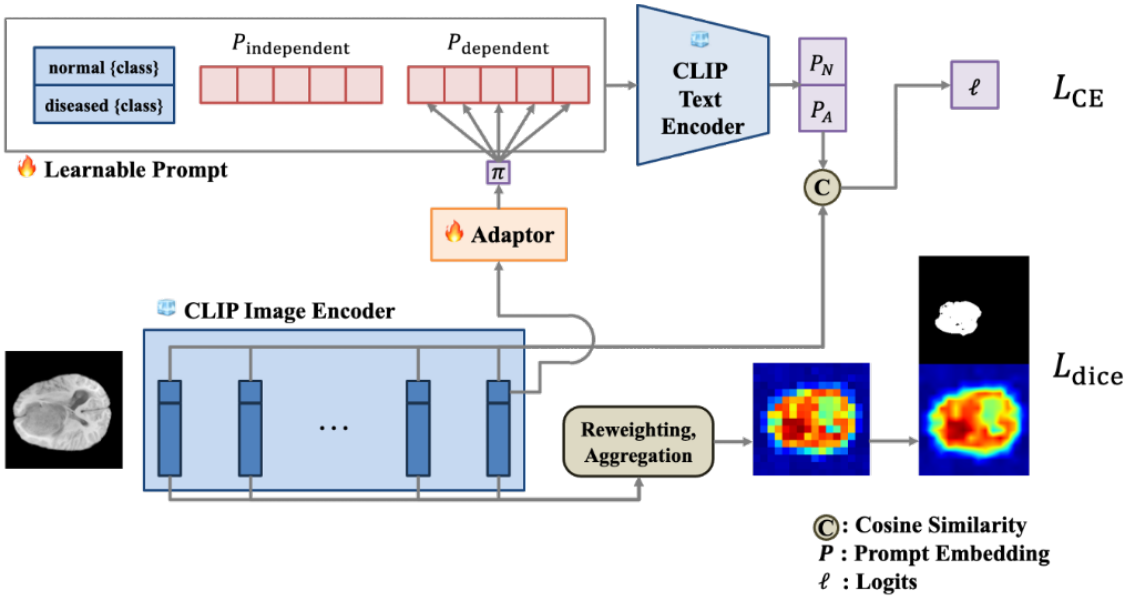

Figure 1. Overall architecture of the model.

Figure 1 shows overall pipeline of the model. The prompt consists of an image dependent prompt and an independent prompt, where the dependent prompt is transformed through an adaptor. Since CLIP is trained to align images with their corresponding textual meanings for classification tasks, using raw CLIP can show poor performance at some tasks that need to focus on semantic details.

Prompt Learning. The learnable prompt P tries to optimize the input of the text encoder. Specifically, the learnable prompt P is composed as follows.

$$P_{\text{independent}} = [\mathrm{V}_1][\mathrm{V}_2]\ldots[\mathrm{V}_\mathrm{M}]$$

Here, each $[V_m](m \in \{1,2,\ldots,\mathrm{M}\})$ is learnable vector with same dimension of embedding space. In addition to this prompt, we also adopt image-conditioned prompt template. Given the image x_I image-dependent prompt template is composed as follows.

$$P_{\text{dependent}} = [\mathrm{W}_1(\mathrm{x}_\mathrm{I})][\mathrm{W}_2(\mathrm{x}_\mathrm{I})]\ldots[\mathrm{W}_\mathrm{M}(\mathrm{x}_\mathrm{I})]$$

Finally, the prompt is composed as follows.

$$P = concat([\text{state}],[\text{CLS}],\mathrm{P}_{\text{independent}},\mathrm{P}_{\text{dependent}})$$

[state] is a vector embedding either "normal" or "diseased". And [cls] is the name of class (ex, brain). This makes the input prompt image-dependent, allowing the prompt to represent diverse meaning depend on target domain and

slices.

By leveraging both prompts simultaneously, the model generates prompts that are robust across various domains and anatomical regions.

Anomaly Score and Localization. Anomaly scores are computed based on similarity between normal and abnormal class embeddings. The global anomaly score is derived from the CLS token. To incorporate features from multiple layers, an attention-based weighting mechanism is applied. The importance weight for each layer is computed as:

$$w_l = \mathrm{softmax}\left(\frac{1}{HW}\sum_{h,w} M_l(h,w)\right)$$

The final anomaly map is then obtained as:

$$M_{\text{final}} = \sum_{l=1}^{L} w_l M_l$$

The anomaly map is upscaled to match the input resolution, using bicubic interpolation.

Same strategy is applied to S_l, the cls token of l-th image encoder layer. Finally the final anomaly score is computed as the average across all layers.

$$S_{\text{final}} = \frac{1}{L}\sum_{l=1}^{L} S_l$$

Detailed algorithm for anomaly map computation is as follows.

Algorithm 1 Anomaly Map Computation

Require: Image Feature I, Prompt Feature P
  Initialize M as an empty list
  for $l = 1$ to L **do**
    Extract patch features F_l from $I[l]$
    Compute anomaly score for patch: $M_l = \cos(F_l, P_A) - \cos(F_l, P_N)$
  end for
  Compute weighted anomaly scores: $w = \mathrm{softmax}\left(\frac{1}{HW}\sum_l M_l\right)$
  Compute final anomaly map: $M_{final} = \sum w_l M_l$
  Upscale M_{final} to size $(224, 224)$ and normalize
  Return M_{final}

Logit is calculated by multiplying logit scale and cosine similarity between image feature and prompt feature. Model is optimized via cross entropy loss and dice loss.

3.3 Datasets

Brain dataset. For brain data, the BraTS2021 dataset is utilized, a publicly available collection of MRI scans for brain tumor segmentation. For each sample, we selected 10 slices centered on the middle slice of the tumor region, with 5 slices above and 5 slices below to capture relevant contextual

information. If there was no segmented tumor area, the sample was labeled as normal; otherwise, it was marked as abnormal. We used MRI scans from 400 patients for training and the remaining data for testing.

Chest dataset. Following BMAD benchmark[9], Chest X-ray dataset is organized utilizing [10].

Retinal dataset. Also following BMAD benchmark, Optical Coherence Tomography (OCT) dataset is organized utilizing two dataset, Retinal Edema Segmentation Challenge dataset and Retinal-OCT dataset.

4. Experiment

4.1 Results

Experiment on diverse viewpoints and sequences. To validate generalizability to diverse view and MRI sequences, we trained model on axial slice with T1 weighted MRI, and tested on sagittal and coronal slice with other MRI sequences. Pixel-level AUROC is reported as below.

	Viewpoint		
MRI Sequences	Axial	Sagittal	Coronal
T1	93.19	87.43	87.16
T1CE	85.97	83.52	84.74
T2	90.26	85.92	87.42
FLAIR	88.43	84.78	85.20

Table 1. Results of experiment on diverse viewpoints and sequences. Pixel-level AUOROC (%) is reported.

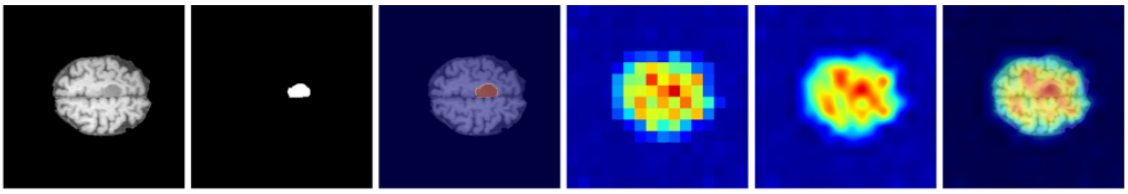

Figure 2. Sample of In-Domain (T1w MRI, axial slice) data. Input, GT, Input+GT, patch embedding, anomaly map after interpolation, Input+Prediction.

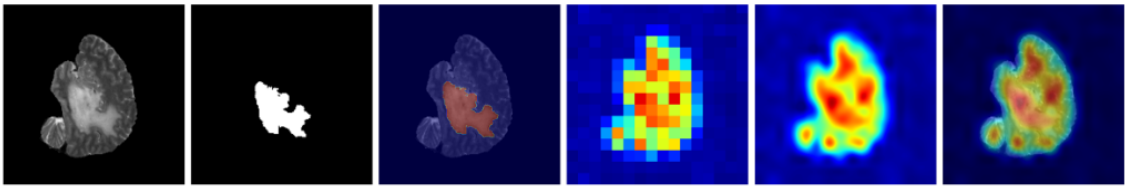

Figure 3. Sample of Domain Shift (T2w MRI, sagittal slice) data. Input, GT, Input+GT, patch embedding, anomaly map after interpolation, Input+Prediction.

Experiment on diverse organs. To validate generalizability to severe domain

shifts, we tested model on BMAD dataset, as mentioned above. Here we used zero-shot transfer setting, which means there was no extra train to target domain.

	Image-AUROC	Pixel-AUROC
Chest	55.47	N/A
Retina	55.83	53.15

Table 2 Results of experiment on diverse viewpoints and sequences. Image-level AUROC (%) and pixel-level AUOROC (%) is reported.

AUROC values close to 0.5 indicate performance nearly equivalent to random classification, suggesting that the model did not perform well on Chest and Retina datasets. However, the model demonstrated meaningful performance in pixel-level AUROC. Given that it was trained solely on T1w brain MRI and evaluated in a zero-shot transfer setting, we expect even better performance with few-shot learning on the target domain or by incorporating auxiliary data from other domains.

5. Discussion

By utilizing not only the CLS token but also the patch embeddings in ViT, computational efficiency was improved. However, this approach is not suitable for precise segmentation tasks. Due to the absence of a decoder, pixel-wise predictions are not feasible, and an optimal threshold must be specified. Additionally, it was observed that optimizing the patch tokens and the CLS token simultaneously was not effective. As the influence of classification loss increased, segmentation performance deteriorated, and vice versa. It remains unclear whether this issue stems from the reweighting method or the design of the loss function, but resolving this is crucial for improving the performance of a multi-task learning approach.

Another issue is that the model was trained solely on MRI data. Previous studies[4, 5] pre-trained models on datasets with diverse classes, such as MVTec-AD and VisA, or fine-tuned on target domain[6]. However, this study only used the T1w MRI data in BraTS dataset. As a result, the model did not learn from a variety of classes, leading to lower performance on new categories like lung and retina. It is expected that increasing the size of the support set or performing few-shot fine-tuning on the target domain would yield better performance.

6. Conclusion

Although previous studies have emphasized object-agnostic prompt learning, its effectiveness has not been demonstrated in medical datasets due to inappropriate processing methods and its application being limited to specific slices. In this study, we demonstrate that object-conditioned prompts are effective for CLIP-based anomaly detection. We expect that the proposed approach can be applied not only to brain but also other organs.

Reference

[1] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2103.00020

[2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is All you Need. *arXiv (Cornell University)*, 30, 5998-6008. https://arxiv.org/pdf/1706.03762v5

[3] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *International Conference on Learning Representations*. https://openreview.net/pdf?id=YicbFdNTTy

[4] Zhou, Q., Pang, G., Tian, Y., He, S., & Chen, J. (2023). AnomalyCLIP: Object-agnostic prompt learning for zero-shot anomaly detection. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2310.18961

[5] Cao, Y., Zhang, J., Frittoli, L., Cheng, Y., Shen, W., & Boracchi, G. (2024). AdaCLIP: Adapting CLIP with Hybrid Learnable Prompts for Zero-Shot Anomaly Detection. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2407.15795

[6] Zhang, X., Xu, M., Qiu, D., Yan, R., Lang, N., & Zhou, X. (2024). MediCLIP: Adapting CLIP for few-shot medical image anomaly detection. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2405.11315

[7] Zhou, K., Yang, J., Loy, C. C., & Liu, Z. (2022). Learning to prompt for Vision-Language models. *International Journal of Computer Vision*, *130*(9), 2337-2348. https://doi.org/10.1007/s11263-022-01653-1

[8] Zhou, K., Yang, J., Loy, C. C., & Liu, Z. (2022a). Conditional Prompt Learning for Vision-Language models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2203.05557

[9] Bao, J., Sun, H., Deng, H., He, Y., Zhang, Z., & Li, X. (2023). BMAD: Benchmarks for Medical Anomaly Detection. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2306.11876

[10] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017). ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *arXiv preprint arXiv:1705.02315.*