International Journal of Information Technology Vol. 30 No. 2 2024

Enhancing Context Optimization Architecture in Medical Domain

Xiao Fang[a], Yi Lin[a], Yibo Hu[a], Xiaoyu Fu[a], Yi Gu[a], Dong Zhang[b], Kwang-Ting Cheng[b] and Hao Chen[a,c,d]

[a]Department of Computer Science and Engineering, HKUST, Hong Kong, China
[b]Department of Electronic and Computer Engineering, HKUST, Hong Kong, China
[c]Department of Chemical and Biological Engineering, HKUST, Hong Kong, China
[d]HKUST Shenzhen-Hong Kong Collaborative Innovation Research Institute, Futian, Shenzhen, China

jhc@cse.ust.hk

Abstract

Previous advancements in the Context Optimization method successfully investigate ways to adopt powerful pre-trained CLIP-like vision-language models to suitable downstream tasks. In this work, we propose a novel visual symptom-guided prompt learning framework for medical image analysis. It provides convenience in transferring general knowledge from CLIP [13]. There are two key components: a visual symptom generator (VSG) and a dual-prompt network. The VSG targets figuring out visual symptoms from pre-trained large language models such as GPT or Deepseek, while the dual-prompt network will lead the training of two learnable prompts, context prompt (CoP) and merge prompt (MeP). These two prompt modules will benefit the adaption process of our framework to the medical domain, *i.e.* medical image analysis. Experiments show that ViP can obtain an SOTA performance on two datasets.

Keywords: Large-Language Model, Medical Image Analysis, Prompt Tuning, Vision-Language Models.

I. Introduction

Medical image analysis is of considerable importance in healthcare and medical domain tasks, as it enables non-invasive diagnosis and various medical conditions. Current methods have revolutionized zero-shot and few-shot learning by aligning the visual and textual embeddings through contrastive pertaining. However, adopting these supervised learning models for various downstream tasks requires extensive manual prompt-tuning methods, which are labor intensive and suboptimal. More importantly, no previous work has been found that transfers the domain of context optimization from natural domain datasets to the medical domain.

The emergence of large vision language models (VLMs) makes it possible for us to transfer knowledge from large-scale pre-trained models to specific medical tasks that obtain limited data. The data related limitation is more serious in medical domain compared to the natural dataset. There is no available large dataset with the similar size of ImageNet in medical domain as a result of modality, privacy, and amount. Vision language models based on CLIP [13] can learn from language supervision that can leverage text descriptions to enable zero-shot transfer, which allows the model to classify images into unseen categories using flexible text prompts such as *"a photo of a {}"*. Inspired by recent work [10, 12], we propose to address the interpreting challenge by translating abstract medical lexicons to visual symptoms that can be shared across natural and medical domain, which contains color, shape, and texture.

Based on the model of context optimization (CoOp), conditional context optimization (CoCoOp), and Knowledge-guided Context Optimization (KgCoOp). We try to hold the pros of static context optimization that shows great performance on the base class while at the same time outperform great generalization ability on downstream tasks.

The main contributions of our work are as fellows: 1) We show that large language models can benefit greatly on prompt engineering, providing better performance and enhancing interpretability. 2) We propose ViP that use LLM to generate natural language visual symptoms for medical images, and use CLIP based models to facilitate knowledge transfer. 3) We conduct experiments on two datasets, and the result demonstrates the strong generalization ability of ViP in the domain of medical image analysis.

II. Method

A. Pipeline

The pipeline of our method is presented in Fig 1. We consider an input image x and a set of disease labels $C = \{c_1, c_2, ..., c_n\}$, where n is the total number of disease categories. The pipeline begins by passing x through a pre-trained vision encoder in the dual-prompt network to compute a feature vector f. In parallel, visual symptoms are generated by the visual sympton generator (VSG) for each disease category. These symptons are then processed to transformation in the context prompt (CoP) module in order to create textual input embeddings for the dual-prompt network. These textual embeddings are then processed by merge prompt (MeP) throgh the pre-trained text encoder to compute the textual features to obtain a set of aggregated visual descriptive features $S = \{s^{c_1}, s^{c_2}, ..., s^{c_n}\}$, where s^c is the representative feature for disease category c. Finally we predict the disease category with the cosine similarity score $f * s^c, c \in C$.

We propose ViP to help us with the process, a **Vi**sual symptom-guided **P**rompt learning framework consists a visual symptom generator (VSG) and a dual-prompt network to help us.

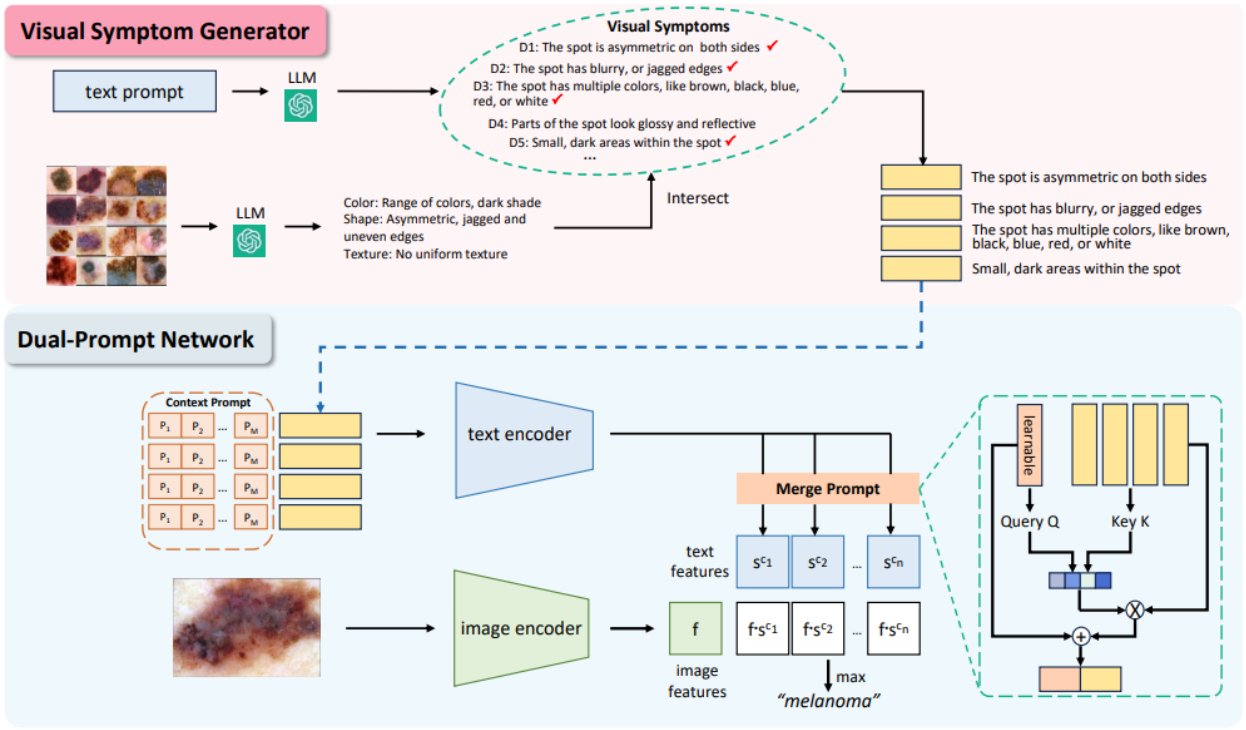

Fig 1 Overview of ViP, which consists of a visual symptom generator (VSG) and a dual-prompt network. The visual symptoms predicted by VSG are used as inputs for downstream networks (marked by the blue dashed line).

A-1. Visual Symptom Generator (VSG)

In VSG, we query pre-trained large language models to generate the visual symptoms, which serve as text inputs for the dual-prompt network. It aims to generate a comprehensive set of visual symptoms specific to each disease category and modality. *Q: I am going to use CLIP, a vision-language model to detect* {*category*} *in* {*modality*}. *What are useful medical visual features for diagnosing*, where {*category*} is substituted for a given category $c \in C$ and $\{modality\}$ is substituted for the imaging modality of the dataset. Next, we refine the coarse set by leveraging the visual-question-answering function of GPT-4 [1]. We prompt it with multiple images for each disease category using the following query: *Q: Please provide visual features regarding color, shape, and texture of this category image, which contains 16 sub-images.* Fig 2 shows the result we obtained an answer containing the visual symptom through the visual-question-answering function. As expected, the

generated result cover descriptions of color and shape of lesions, presence or absence of certain structures, and other relevant visual features in text.

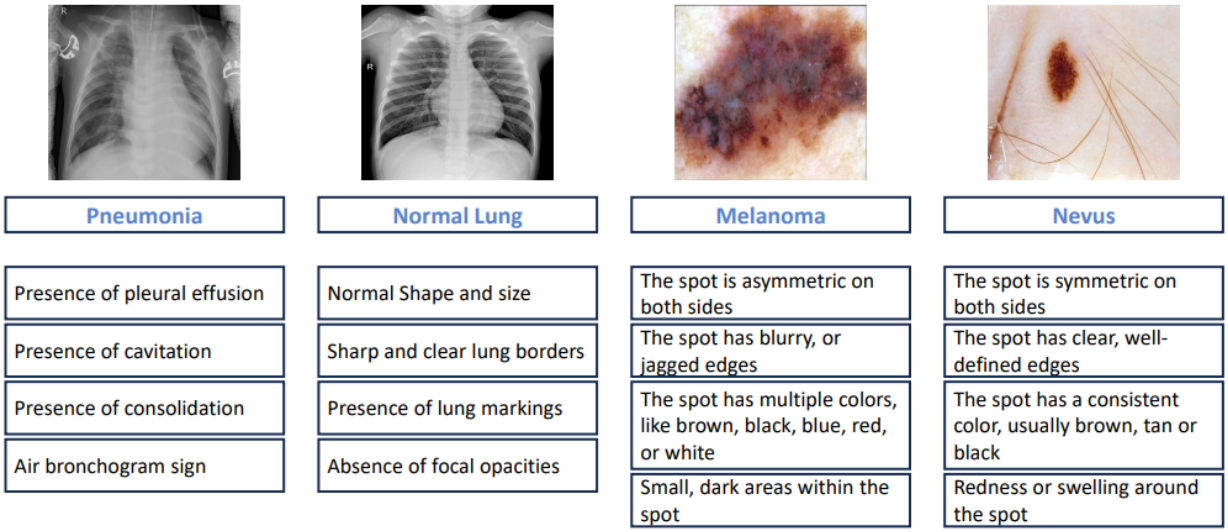

Fig 2 Example visual symptoms generated by GPT-4 [1].

A-2. Dual-Prompt Network

The dual-prompt network, which bases on CLIP-based approachses such as context optimization, enhances the generalization ability by proposing *context prompt* and *merge prompt*. We freeze the image encoder and text encoder of CLIP and visual symptoms generated from the VSG to facilitate the alignment of medical image features. However, the generalization ability is still limited due to CLIP text input format in the response from LLMs and inherent challenge of effectively aggregating visual symptoms into a disease representation without explicit training [9, 4], which leads us to propose the dual-prompt network.

CoP. Other than the category names, context words help to form a complete sentence that plays a crucial rule in the textual input of CLIP with the form of *a photo of a* {} . However, it is not understandable for natural pretrained LLM to understand medical imaging without any

prompt. So other than the original category name, we introduce a set of learnable tokens $\{p\}_{i=1}^{M}$, where $p_i \in R^d, i = 1, 2, ...M$, and d is the text embedding dimension, prior to visual symptoms in which context of medical tasks are automatically learned. Specifically, given a category $c \in C$, and a visual symptom word embedding e_d, the final input word embedding T for the text encoder is concatenation of the learnable tokens and e_d, which can be formulated as $T = Concat(p_1, p_2, ...p_M, e_d)$.

MeP. After processing visual symptoms via text encoder, the next step is to merge visual symptoms into a single representation. Previous methods [8, 10, 2] include average function, which treats all visual symptoms equally, or the maximum function, which diagnoses based on the most prominent feature. In reality, in contrast, not all visual symptoms contribute equally to a diagnosis process of a disease. Therefore, a learnable token for each disease categories are introduced to learn the features of the disease. Given a category $c \in C$, text features matrix $T = [T_1^c, T_2^c, ..., T_k^c]^T$, where $T \in R^{k \times d}$ and d is the text embedding dimension, which is obtained by processing related visual symptoms through the text encoder, and a learnable group token $g \in R^d$, we first project g and T into query $Q \in R^d$ and key $K \in R^{k \times d}$ with different weights $W_q \in R^{d \times d}$ and $W_k \in R^{d \times d}$, which can be formulated as:

$$Q = gW_q, K = TW_k \tag{1}$$

Then we calculated the aggregated feature s^c by combining grouping g and weighted text features matrix T, which can be formulated as:

$$s^c = g + Softmax(\frac{QK^T}{\sqrt{d}})T \tag{2}$$

After aggregating visual features of all disease categories, CoP and MeP are jointly optimized by Cross-Entropy Loss, which can be formulated as:

$$L_{ce} = -log\frac{exp(f \cdot s^{c_y}/\gamma)}{\sum_{i=1}^{n} exp(f \cdot s^{c_i}/\gamma)} \tag{3}$$

where c_y denotes the ground truth disease category and γ is a learned temperature.

III. Experiments

Dataset. The experiment is held on two available public datasets: Pneumonia [7] and Derm7pt [6]. Pneumonia is a two catogories dataset containing normal lung and pneumonia. The official split of this dataset contains 5232 images for training and 624 images for testing. We further randomly divide the training set with a 9:1 ratio for training and validation. Derm7pt consists of over 2000 clinical and dermoscopic images. We follow [11] to filter the dataset to obtain 827 images belonging to "melanoma" and "nevus" classes, and split the dataset into 346, 161, and 320 images for training, validation, and testing, respectively. For both datasets, we adopt Accuracy (ACC) and Macro F1-score (F1) as evaluation matrics.

Settings. Through the whole experiments, we average the results of three vision backbones in CLIP [13], $i.e.$, ViT-B/16 [3], ViT-L/14 [3], and ResNet-50 [5] We follow CLIP [13] to set the text embeddings dimension d to 512. We follow CoOp [14, 15] to learn a unified task context and set the length M of the context prompt (CoP) to 4. Training is done with SGD and an initial learning rate is 0.001. The training epoch is set to 50. We follow CLIP [13] to set the temperature γ in the cross-entropy loss to 0.01.

Results. We further compare ViP with several SOTA prompt-based models to evaluate the generalization ability. As shown in Table 1, ViP achieves highest accuracy of 86.69%, 81.11%,

Table 1 Result comparisons with SOTAs. The mean and standard deviation is computed across three vision backbones.

Method	Pneumonia		Derm7pt	
	ACC	F1	ACC	F1
CoOp	$0.8337_{0.019}$	$0.8148_{0.017}$	$0.7823_{0.005}$	$0.7328_{0.017}$
CoCoOp	$0.8440_{0.025}$	$0.8217_{0.032}$	$0.7668_{0.014}$	$0.6647_{0.057}$
KgCoOp	$0.8303_{0.022}$	$0.8010_{0.027}$	$0.7726_{0.009}$	$0.7093_{0.033}$
Bayesian	$0.8301_{0.041}$	$0.8081_{0.048}$	$0.6921_{0.014}$	$0.5561_{0.054}$
MaPLe	$0.8553_{0.034}$	$0.8393_{0.036}$	$0.7903_{0.038}$	$0.7250_{0.073}$
Supervised	$0.8660_{0.025}$	$0.8530_{0.025}$	$0.7277_{0.044}$	$0.6236_{0.093}$
ViP_{ours}	$\mathbf{0.8669}_{0.031}$	$\mathbf{0.8494}_{0.036}$	$\mathbf{0.8111}_{0.007}$	$\mathbf{0.7730}_{0.015}$

and F1-score of 84.94%, 77.30% on Pneumonia [7] and Derm7pt [6] respectively, indicating strong generalization ability of our method. Especially on Derm7pt in which there is less training data, demonstrating powerful generalization ability.

IV. Conclusion

This paper presents ViP, a visual-guided prompt learning pipeline, effectively transfers knowledge from VLMs to medical domain. By leveraging pretrained LLMs, ViP generates useful visual symptoms to guide CLIP [13] in aligning medical image with natural language visual symptoms. Compared with context optimization methods, we additionally apply context prompt and merge prompt, to further enhance the generalization ability. Experiments show that we perform a effective tool for the module to understand medical image in natural language. The effectiveness of each module and the superior performance of our pipeline to state-of-the-art methods.

References

[1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F.L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, and et al. *Gpt-4 technical report*. arXiv, 2023.

[2] M. Byra, M.F. Rachmadi, and H. Skibbe. *Few-shot medical image classification with simple shape and texture text descriptors using vision-language models*. arXiv, 2024.

[3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, and et al. *An image is worth 16x16 words: Transformers for image recognition at scale*. In: international Conference on Learning Representations, 2020.

[4] T. Franquet. *Imaging of pneumonia: trends and algorithms*. European Respiratory Journal 18(1), 196–208, 2001.

[5] K. He, X. Zhang, S. Ren, and J. Sun. *Deep residual learning for image recognition*. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778, 2016.

[6] J. Kawahara, S. Daneshvar, G. Argenziano, and G. Hamarneh. *Seven-point checklist and skin lesion classification using multitask multimodal neural nets*. IEEE journal of biomedical and health informatics 23(2), 538–546, 2018.

[7] D.S. Kermany, M. Goldbaum, W. Cai, C.C. Valentim, H. Liang, S.L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, and et al. *Identifying medical diagnoses and treatable diseases by image-based deep learning*. cell 172(5), 1122–1131, 2018.

[8] J. Liu, T. Hu, Y. Zhang, X. Gai, Y. Feng, and Z. Liu. *A chatgpt aided explainable framework for zero-shot medical image diagnosis*. arXiv, 2023.

[9] S.N. Markovic, L.A. Erickson, R.D. Rao, R.R. McWilliams, L.A. Kottschade, E.T. Creagan, R.H. Weenig, J.L. Hand, M.R. Pittelkow, B.A. Pockaj, and et al. *Malignant melanoma in the 21st century, part 1: epidemiology, risk factors, screening, prevention, and diagnosis*. In: Mayo Clinic Proceedings. vol. 82, pp. 364–380, 2007.

[10] S. Menon and C. Vondrick. *Visual classification via description from large language models*. In: The Eleventh International Conference on Learning Representations, 2022.

[11] C. Patrício, J.C. Neves, and L.F. Teixeira. *Coherent concept-based explanations in medical image and its application to skin lesion diagnosis*. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3798–3807, 2023.

[12] Yi Qin, Z., H.H., Q. Lao, and K. Li. *Medical image understanding with pretrained vision language models: A comprehensive study*. In: The Eleventh International Conference on Learning Representations, 2022.

[13] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, P. Askell, A. amd Mishkin, J. Clark, and et al. *Learning transferable visual models from natural language supervision.* In: International conference on machine learning pp. 8748–8763. PMLR, 2021.

[14] K. Zhou, J. Yang, C.C. Loy, and Z. Liu. *Conditional prompt learning for visionlanguage models*. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16816–16825, 2022.

[15] K. Zhou, J. Yang, C.C. Loy, and Z. Liu. *Learning to prompt for vision-language models*. International Journal of Computer Vision Vision 130(9), 2337–2348, 2022.