2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

Cross-modal Adaptation of medical vision-language model for few-shot classification

Jingyi Wu[1,2] and S. Kevin Zhou[1,2,3,4](✉)

[1] School of Biomedical Engineering, Division of Life Sciences and Medicine,
University of Science and Technology of China (USTC), Hefei Anhui, 230026, China
[2] Center for Medical Imaging, Robotics, Analytic Computing & Learning (MIRACLE),
Suzhou Institute for Advance Research, USTC, Suzhou Jiangsu, 215123, China
[3] State Key Laboratory of Precision and Intelligent Chemistry,
USTC, Hefei, Anhui 230026, China
[4] Jiangsu Provincial Key Laboratory of Multimodal Digital Twin Technology,
Suzhou Jiangsu, 215123, China
skevinzhou@ustc.edu.cn

Abstract. Medical Vision-Language Models (VLMs) show significant potential for auxiliary diagnosis, especially given the continuous growth of medical image data. However, adapting these models effectively with limited labeled data remains a challenge. This paper proposes a cross-modal adaptation method for few-shot medical image classification based on pre-trained VLMs. Our approach leverages both image features and corresponding text features extracted from the pre-trained models to train a classifier head. Furthermore, we employ the SHAP interpretability analysis method to select the most informative text features, thereby enhancing classification performance. We evaluated our method on the CheXpert5x200 dataset using MedCLIP and KAD as foundation models, comparing it against zero-shot classification and uni-modal adaptation (using only image features). Results demonstrate that our approach significantly improves few-shot classification performance over the baselines. The SHAP-based feature selection provides additional gains. Ultimately, we present a general, simple, and efficient cross-modal adaptation strategy that enhances medical VLM performance using a small number of image samples, contributing to more reliable AI-powered diagnostic tools.

Keywords: Vision-Language Model, Cross-Modal Adaptation, Few-shot Learning.

1 Introduction

In recent years, VLMs such as CLIP[1] have received increasing attention in medicine. The CLIP model maximizes the similarity between image features and text features of the same category through contrastive learning, introduces text information into the visual model, and improves the model's classification capabilities of medical images. In addition, the increasing amount of medical imaging and pathology report data also provides support for the pre-training of medical visual language large models. More and more medical visual language large models continue to emerge, helping doctors to better diagnose patients.

Based on the ability of the visual language large model to tightly align text and image information into the latent space, the cross-modal adaptive method further improves the classification performance of the model by training a linear classifier with a mixture of aligned text and image features. At the same time, more text features can be added as additional training samples by setting a variety of artificial text prompt methods containing category names, or by using other automatic text prompt methods. This method only requires a small number of image samples to achieve few-sample classification, and quickly improves performance while occupying a small amount of computing resources.

The main contributions of this paper are as follows:

- Introducing cross-modal adaptation methods into the basic model of medical visual language. For the first time, cross-modal adaptation technology is used in the field of medical images to significantly improve the small sample classification performance of the basic model.
- Introducing the SHAP strategy to filter text features. For the same type of disease label, we can generate multiple text prompt words, and their quality varies. We use SHAP interpretability analysis to rank the importance of text features. In this way, we can filter out better text features for classification head training.

2 Related Works

2.1 Vision-Language Model in Medical Imaging

Vision-Language Models (VLMs) such as the CLIP match images with text through text and image encoders, and train the model by comparing and learning the similarity between text and image embeddings. They have good performance in the field of natural images. Compared with natural images, in addition to global visual features, local visual features are also important for medical images. At the same time, the processing of medical reports also requires more professional medical knowledge. The application of CLIP pre-training on medical image text datasets needs to be improved accordingly [2]. ConVIRT [3] first introduced contrastive learning to align medical images and reports. GLoRIA [4] introduced semantically driven multi-scale contrast and considered global and local information for image text feature alignment. MedCLIP[5] further utilized unpaired medical image report data. KAD[6] introduced word-based entity extraction to extract report information, thereby improving the generalization ability of the model. CARZero [7] replaced the original similarity calculation with a cross-attention mechanism to more accurately reflect the similarity of complex relationships in medical semantics. Large models have strong generalization capabilities, but for specific tasks, their performance may not be as good as lightweight models trained for the task. How to enhance the adaptability of large models to specific tasks is one of the current research directions.

In this research, we focus on MedCLIP and KAD as vision-language foundation models.

2.2 Adaptive methods for pre-trained models

With the development of pre-trained models, more and more work has focused on how to better apply large models to downstream tasks. The most basic methods include local or global fine-tuning[8] of the CLIP model, linear probing [9], and other methods. The design of text prompt words greatly affects the model performance and the adaptability of large models to specific tasks. Manually designed prompt words require a lot of attempts to find the best template, while prompt learning such as CoOp [10] and CoCoOp [11] can automatically learn the optimal prompt vector and improve classification performance. CLIP-Adapter [12] and Tip-Adapter[13] increase the fine-tuning speed by adding lightweight multi-layer-perceptrons (MLPs) in parallel and freezing the original parameters during training to only train the Adapter, which not only retains the generalization ability of the large model but also improves its adaptability to specific tasks.

However, the above methods only use information from a single modality. The cross-modal adaptation method[14] is different from the above uni-modality adaptive method. It adds data from other modalities as additional training samples during the adaptive process, making better use of information from both image and text modalities, and further improving the adaptability of the large model on the basis of the above single-modality adaptive method.

3 Methodology

3.1 Uni-modal linear probing

The CLIP is pre-trained on large-scale data using contrastive learning method, and we get pretrained image encoder $E_I(\cdot)$ and text encoder $E_T(\cdot)$, which can extract more general image and text features and have stronger robustness. Linear probing freezes the weights of the $E_I(\cdot)$, extracting features from input images, and training a simple classifier head $H(\cdot)$ (e.g., a single fully connected layer). For n-shots classification, we have n $(I, label)$ pairs, and train the classifier head with the softmax loss:

$$Loss_{uni-modal} = \sum_i softmax(H(E_I(I_i)), label_i) \tag{1}$$

During the uni-modal linear probing, we only use the image features to train the classifier head, while the text features were not used. Although the text and image features are aligned through contrastive learning in the pre-training stage, we can still improve the classification performance by adding text features in the classification head training stage.

3.2 Cross-modal linear probing:

Different from uni-modal linear probing, we incorporate image features and text features together during the training of the classifier head. For n-shots classification, we not only use n $(I, label)$ pairs, but also add m $(T, label)$ pairs into the training. The default value of m is 10.

$$Loss_{cross-modal} = \sum_i softmax\big(H\big(E_I(I_i)\big), label_i\big) + \sum_j softmax(H\left(E_T(T_j)\right), label_j) \qquad (2)$$

Algorithm 1: The framework of Cross-modal Adaptation for medical vision-language foundation model

```
Input:
I, I_label: images and corresponding image labels
T, T_label: texts and corresponding text labels
t: temperature scaling (2.5911 for MedCLIP and 2.6593 for KAD)
Output: logits, logit_head
Features pre-extraction:
I_feature = Image_encoder(I)
T_feature = Text_encoder(T)
# mix the text feature and image feature (optional)
if mix feature :
    # Mix the text feature and image feature which has the same label
    Mix_feature = mix_ratio * I_feature + (1 – mix_ratio) * T_feature
Classifier head training:
train_features = cat((T_feature, I_feature), dim=0)
train_labels = cat((I_label, T_label), dim=0)
train_loader = DataLoader(train_features, train_labels)
# initialize classifier head randomly
linear_head.init()
while train_loader:
    features, labels = train_loader.next()
    #compute the loss and update the params of the classifier head
    logits = linear_head(features)
    loss = SoftmaxLoss(logits / t, labels)
    loss.backward()
    update(linear_head.params)
```

3.3 Features Selection using SHAP analysis

SHAP (SHapley Additive exPlanations) [15] is a model interpretation framework based on cooperative game theory, which aims to unify and optimize existing feature importance analysis methods and provide interpretability for the predictions of complex models (such as deep learning and integrated models).

The Shapley value provides a theoretically sound method to fairly distribute the prediction outcome of a model f for a specific input x among its individual features, calculating the contribution $\varphi_{i(f,x)}$ for each feature i. It operates by considering every possible subset S of features that excludes the feature i being evaluated. For each subset S, it calculates the marginal contribution of adding feature i to that subset, which is the difference in the model's prediction with feature i present ($f_{x(S\cup\{i\})}$) versus absent ($f_{x(S)}$). Finally, the formula computes a weighted average of these marginal contributions across all possible subsets S, using a specific combinatorial weight $\frac{(|S|!*(M-|S|-1)!)}{M!}$ derived from cooperative game theory, ensuring that the feature's contribution is averaged fairly across all possible contexts and orders of feature inclusion. Where F is the set of all input features, M is the total number of features in the set F.

$$\varphi_{i(f,x)} = sum_{S\subseteq F\{i\}} \left(\frac{(|S|!*(M-|S|-1)!)}{M!}\right) * \left[f_{x(S\cup\{i\})} - f_{x(S)}\right] \tag{3}$$

The algorithm 2 outlines a procedure for selecting important text features using SHAP analysis. It requires as input a pre-trained classifier (linear head), the training features (which include both image and text data used to originally train the classifier), and the full set of text features encoded by a text encoder. First, the training features are designated as the background dataset, providing context for the SHAP calculations. A SHAP Explainer is then initialized with the linear head model and this background data. Subsequently, this explainer computes the shapley values for the input text features, quantifying the contribution of each text feature dimension to the linear head's predictions relative to the background. Finally, the algorithm selects the top k (specified as 10 in this example) text features by ranking them based on their calculated shap values (implicitly using a metric like mean absolute SHAP value), outputting these as the selected features.

Algorithm 2: Selecting text features by SHAP analysis

Input:
Linear_head: well trained classifier head
Training features: the features training classifier head including 16 image features and 15 text features per class
Text_features: all the text features encoded by text encoder
Output:
Selected_features: selected text features

```
# select training features as background
background = training_features
explainer = shap.Explainer (linear_head, background)
shap_values = explainer.shap_values (Text_features)
selected_features = top_k (Text_features, sort_by = shap_values, k = 10)
```

4 Experiments and Evaluation

4.1 Experiment Dataset

We referred to the experiments in the MedCLIP[5] and selected the CheXpert5x200 dataset[16] for the experiment, which contains 200 chest X-ray images of five categories: atelectasis, cardiomegaly, edema, pleura, and effusion, and each image contains only one positive label of a specific category. Since both the foundation models MedCLIP and KAD are trained only on X-ray datasets, our experiments are also analyzed and evaluated on X-ray datasets. These images have been checked by radiologists to ensure that they do not contain other abnormalities besides the label of that category, and that the image quality and shooting posture are normal. The image dataset was divided, and 800 images were selected as the test set, and the other 200 images were used for training and validation. The training and validation sets were further divided into small sample sets, and small sample training sets of [1, 2, 4, 8, 16] images were divided respectively, and randomly sampled 10 times.

4.2 Text prompts generation:

For each disease type, we used the text prompt word generation method in the MedCLIP paper. For each disease type, Atelectasis, Cardiomegaly, Consolidation, Edema, and Pleural Effusion, we used the following template to generate prompt words: [‘severity’] [‘subtype’] [‘location’], which is used to generate richer disease descriptions rather than simple disease names. At the same time, more diverse prompt words can also better align with chest X-ray image features. The total number of text prompt words for each disease type is shown in Table 1.

Table 1. Text prompts numbers for each disease class

Class Name	Text prompts numbers
Atelectasis	210
Cardiomegaly	15
Consolidation	192
Edema	18
Pleural Effusion	54

4.3 Experimental Settings

The optimizer used is AdamW, and for each single seed, we try different hyperparameters including learning rate, weight decay and batch size to achieve best results on the validation set. We search the best learning rate in [1e-4, 1e-5], weight decay in [0.0, 0.01, 0.0001], and batch size in [1, 2, 4, 8, 16]. In the first 50 iterations we warmup the learning rate with linear type, and we set the maximal iteration of each examination to

12800. To achieve better results, we evaluate on the validation set per 10 iterations, and retain the model which has the best AUC result on validation set.

4.4 Results and Discussions

Table 2 shows the experimental results obtained with MedCLIP as the pre-training model. The zero-shot row is the zero-shot classification result of the MedCLIP model. The Uni row indicates that only the image features are added during the training of classifier head. N-shots of images are added to perform a few-shot classification experiment, which corresponds to the uni-modal linear probing experiment. The Cross row indicates that in addition to adding a small number of pictures, 10 text features are added to the cross-modal training classification head, which corresponds to the cross-modal linear probing experiment.

Table 2. The ACC and AUC results based on MedCLIP

	Shots	Zero	1	2	4	8	16
acc	Zero-shot	0.56750					
	Uni		0.42150	0.47888	0.54388	0.59938	0.63750
	Cross		0.60863	0.62075	0.63013	0.64025	0.64500
	Shap		**0.60875**	**0.62525**	**0.63513**	**0.64250**	**0.64525**
auc	Zero-shot	0.77569					
	Uni		0.70910	0.75780	0.80465	0.84627	0.86082
	Cross		**0.86291**	0.86448	0.86563	0.87001	0.86952
	Shap		0.86236	**0.86682**	**0.86900**	**0.87322**	**0.87029**

In subsequent experiments, we also tested the impact of different numbers of text features on the results. The detailed results are shown in Table 4. The Shap row shows the results obtained by training after using SHAP to select text features. Here, we select the best 10 features selected by each type of label for training. We also experimented with the number of selected features. The detailed results are shown in Table 5.

It can be seen that the cross-modal linear probing result is significantly improved compared with the uni-modal linear probing, indicating that the introduction of text modality can further improve the adaptive task.

After using SHAP to select text features, the classification effect is also improved, indicating that the feature selecting method is effective. This method can also automatically select the best text features when there are more text inputs in the future, which facilitates feature selection.

The results based on KAD modal is shown in Table 3. Compared with the MedCLIP model, the improvement on the KAD model is not so significant. This is because the KAD model has a step of fusing text and image features during pre-training stage. Even so, it still has a significant improvement over zero-shot classification and still has some advantages compare to single-modal adaptive classification results.

The result shows that, our method can be applied to a variety of pre-trained models, and the training is very fast because text and image features are extracted in advance, and the classification head has very few parameters.

Table 3. Results based on KAD

	Shots	Zero	1	2	4	8	16
acc	Zero-shot	0.52470					
	Uni		0.56813	0.58213	0.58850	0.60213	0.60100
	Cross		0.58250	0.58750	0.59000	0.60236	0.60663
	Shap		**0.59125**	**0.59438**	**0.60100**	**0.60750**	**0.61163**
auc	Zero-shot	0.79160					
	Uni		0.82148	0.83022	0.83846	0.84646	0.85201
	Cross		0.83170	0.83272	0.84035	**0.84709**	0.85176
	Shap		**0.83315**	**0.83374**	**0.84167**	0.84689	**0.85195**

4.5 Ablation study

In the experiments in Table 4, we conducted experiments on different text numbers. Intuitively, we can expect that the performance of the experiment should improve with the increase of text numbers. But comparing the results of 10 and 15 text numbers, extra text features do not significantly improve classification performance, but instead harms it. We set the maximal text number to 15 because the Cardiomegaly class has only15 prompts, as shown in Table 1.

If low-quality text features are added to the Cross-modal linear probing training, the final classification performance will be reduced. Therefore, we need to select the best text features, especially as more and more work uses LLM to generate text prompts. Faced with more and more text features, features selection becomes more important.

Meanwhile, we compared different number of text features selected by SHAP in Table 5. As we add the number of selected features from 10 to 12, the improvement of acc and auc is not that obvious. This may be because the differences between text prompts are limited and a small number of text features are representative enough, and we choose 10 as the default value of k.

Table 4. Ablation text number

		Text num						
	Image num	Uni(0)	1	2	4	8	10	15
acc	1	0.4215	0.5609	0.5821	0.6048	0.6050	**0.6086**	**0.6086**
	2	0.4789	0.5820	0.6031	**0.6245**	0.6174	0.6208	0.6186
	4	0.5439	0.5974	0.6113	0.6261	0.6259	**0.6301**	0.6286
	8	0.5994	0.6188	0.6246	0.6394	**0.6403**	**0.6403**	0.6395
	16	0.6375	0.6423	0.6419	0.6458	**0.6460**	0.6450	0.6458
auc	1	0.7091	0.8280	0.8475	0.8564	0.8599	**0.8629**	0.8617
	2	0.7578	0.8362	0.8473	0.8607	0.8612	0.8645	**0.8654**
	4	0.8047	0.8349	0.8482	0.8617	0.8634	0.8656	**0.8684**
	8	0.8463	0.8571	0.8617	0.8654	0.8690	0.8700	**0.8726**
	16	0.8608	0.8631	0.8644	0.8659	0.8682	0.8695	**0.8707**

Table 5. Ablation SHAP

	Shots	1	2	4	8	16
acc	Cross	0.60863	0.62075	0.63013	0.64025	0.64500
	SHAP (n=10)	0.60875	**0.62525**	**0.63513**	0.64250	0.64525
	SHAP (n=12)	**0.60925**	0.62475	0.63413	**0.64300**	**0.64988**
auc	Cross	**0.86291**	0.86448	0.86563	0.87001	0.86952
	SHAP (n=10)	0.86236	**0.86682**	0.86900	0.87322	0.87029
	SHAP (n=12)	0.86205	0.86516	**0.86915**	**0.87382**	**0.87052**

5 Conclusion and Future Work

In this paper, we presented an approach for enhancing the few-shot classification performance of pre-trained medical vision-language models (VLMs) through cross-modal adaptation. By incorporating aligned text features alongside limited image features during the training of a simple linear classifier head, we demonstrated significant improvements in ACC and AUC compared to both zero-shot classification and traditional unimodal linear probing on the CheXpert5x200 dataset, using MedCLIP and KAD as foundation models. Furthermore, we successfully integrated SHAP analysis as an effective method for selecting high-quality text features from a pool of generated prompts, leading to additional performance gains and providing a strategy to manage potentially noisy or redundant text information. Our results highlight that leveraging the inherent cross-modal alignment within VLMs during adaptation is a highly effective and generalizable strategy for boosting performance on specific downstream tasks even with very few labeled image samples.

Acknowledgments. Supported by Natural Science Foundation of China under Grant 62271465 and Suzhou Basic Research Program under Grant SYG202338.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

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.: Learning Transferable Visual Models From Natural Language Supervision. In: Proceedings of the 38th International Conference on Machine Learning. pp. 8748–8763. PMLR (2021).
2. Zhao, Z., Liu, Y., Wu, H., Li, Y., Wang, S., Teng, L., Liu, D., Cui, Z., Wang, Q., Shen, D.: CLIP in Medical Imaging: A Comprehensive Survey, http://arxiv.org/abs/2312.07353, (2023).
3. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive Learning of Medical Visual Representations from Paired Images and Text. In: Proceedings of the 7th Machine Learning for Healthcare Conference. pp. 2–25. PMLR (2022).
4. Huang, S.-C., Shen, L., Lungren, M.P., Yeung, S.: GLoRIA: A Multimodal Global-Local Representation Learning Framework for Label-Efficient Medical Image Recognition. Presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision (2021).
5. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: MedCLIP: Contrastive Learning from Unpaired Medical Images and Text. Proc Conf Empir Methods Nat Lang Process. 2022, 3876–3887 (2022). https://doi.org/10.18653/v1/2022.emnlp-main.256.
6. Zhang, X., Wu, C., Zhang, Y., Xie, W., Wang, Y.: Knowledge-enhanced visual-language pre-training on chest radiology images. Nat Commun. 14, 4542 (2023). https://doi.org/10.1038/s41467-023-40260-7.
7. Lai, H., Yao, Q., Jiang, Z., Wang, R., He, Z., Tao, X., Zhou, S.K.: CARZero: Cross-Attention Alignment for Radiology Zero-Shot Classification. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11137–11146. IEEE, Seattle, WA, USA (2024). https://doi.org/10.1109/CVPR52733.2024.01059.
8. He, K., Chen, X., Xie, S., Li, Y., Dollar, P., Girshick, R.: Masked Autoencoders Are Scalable Vision Learners. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15979–15988. IEEE, New Orleans, LA, USA (2022). https://doi.org/10.1109/CVPR52688.2022.01553.
9. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A Simple Framework for Contrastive Learning of Visual Representations. In: Proceedings of the 37th International Conference on Machine Learning. pp. 1597–1607. PMLR (2020).
10. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to Prompt for Vision-Language Models. Int J Comput Vis. 130, 2337–2348 (2022). https://doi.org/10.1007/s11263-022-01653-1.

11. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional Prompt Learning for Vision-Language Models. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16795–16804. IEEE, New Orleans, LA, USA (2022). https://doi.org/10.1109/CVPR52688.2022.01631.
12. Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y.: CLIP-Adapter: Better Vision-Language Models with Feature Adapters. Int J Comput Vis. 132, 581–595 (2024). https://doi.org/10.1007/s11263-023-01891-x.
13. Zhang, R., Zhang, W., Fang, R., Gao, P., Li, K., Dai, J., Qiao, Y., Li, H.: Tip-Adapter: Training-Free Adaption of CLIP for Few-Shot Classification. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., and Hassner, T. (eds.) Computer Vision – ECCV 2022. pp. 493–510. Springer Nature Switzerland, Cham (2022). https://doi.org/10.1007/978-3-031-19833-5_29.
14. Lin, Z., Yu, S., Kuang, Z., Pathak, D., Ramanan, D.: Multimodality Helps Unimodality: Cross-Modal Few-Shot Learning with Multimodal Models. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19325–19337. IEEE, Vancouver, BC, Canada (2023). https://doi.org/10.1109/CVPR52729.2023.01852.
15. Lundberg, S.M., Lee, S.-I.: A Unified Approach to Interpreting Model Predictions. In: Advances in Neural Information Processing Systems. Curran Associates, Inc. (2017).
16. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D.A., Halabi, S.S., Sandberg, J.K., Jones, R., Larson, D.B., Langlotz, C.P., Patel, B.N., Lungren, M.P., Ng, A.Y.: CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison, http://arxiv.org/abs/1901.07031, (2019). https://doi.org/10.48550/arXiv.1901.07031.