

Exploring Low-Resource Medical Image Classification with Weakly Supervised Prompt Learning

Fudan Zheng¹, Jindong Cao¹, Weijiang Yu², Zhiguang Chen¹ & Yutong Lu¹

¹*Sun Yat-Sen University, No. 132 Waihuan Road, Guangzhou Higher Education Mega Center, Guangzhou, 510006, China;*

²*Huawei Technologies Co., Ltd., Huawei Industrial Park, Bantian, Longgang District, Shenzhen, 518129, China*

Abstract Most advances in medical images supporting clinical auxiliary diagnosis meet a challenge due to the low-resource medical data with highly-expensive and professional annotations. This low-resource problem can be alleviated by leveraging the transferable representation capabilities of large-scale pre-trained vision-language models like CLIP. After pre-trained using large-scale unlabeled medical images and texts (such as medical reports), the vision-language models can learn transferable representations and support flexible downstream clinical tasks such as medical image classification via relevant medical text prompts. However, existing pre-trained vision-language models require domain experts (clinicians) to carefully design the medical text prompts based on different datasets when applied to specific medical image tasks, which is extremely time-consuming and greatly increases the burden on clinicians. To address this problem, we propose a weakly supervised prompt learning method *MedPrompt*, which only employs class labels for inexact supervised learning to enable the model to adaptively and automatically generate effective medical text prompts according to different datasets. Benefiting from these prompts, the pre-trained vision-language model can be freed from the strong expert dependency of manual annotation and manual prompts design, thus truly achieving end-to-end, low-cost medical image classification. Experimental results show that in all four public datasets, the model using our automatically generated prompts outperforms its hand-crafted prompts counterpart which is fine-tuned on all labeled samples with only a minimal number of labeled samples (at most 16) for few-shot learning, and reaches superior or comparable accuracy on zero-shot classification prediction. In addition, the proposed training method can be seamlessly embedded into any other network structures.

Keywords medical image classification, weakly supervised learning, prompt engineering, few-shot learning, zero-shot learning

Citation

1 Introduction

Medical imaging techniques, such as computed tomography (CT), magnetic resonance imaging (MRI) and X-rays, are often used in clinical practice for monitoring, diagnosis and treatment. With the massive growth of medical imaging data and the rapid development of deep learning technology, researchers have developed various deep learning models applied in the field of medical images to support clinical decision-making, and these models have been proven to have high accuracy and generalization ability [1–3]. Existing models generally adopt the supervised learning mode, requiring medical images to have corresponding complete, exact, and accurate annotations. Unlike natural image annotation, which can be done by ordinary people, medical image annotation needs to be carried out by experienced domain experts (clinicians), with a high annotation threshold and cost. This directly leads to the situation of low-resource medical data - limited available annotated samples with highly-expensive and professional annotations, leaving a vast number of unannotated medical images and text (such as medical reports) samples unused.

The emergence of large-scale pre-trained vision-language models such as CLIP [4] makes it possible for unlabeled samples to be fully utilized. There have also been some studies utilizing such models to make

* Corresponding author (email: weijiangyu8@gmail.com, yutong.lu@nscg-gz.cn)

† Fudan Zheng and Jindong Cao have the same contribution to this work.

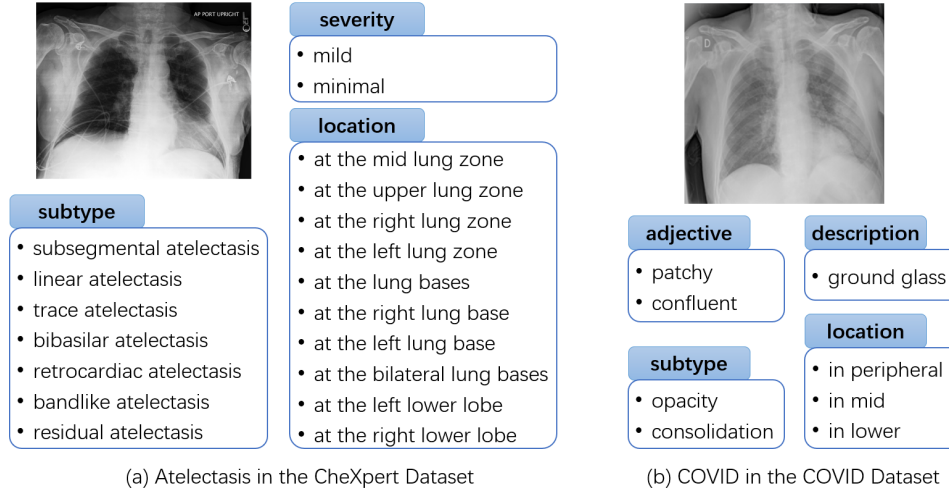


Figure 1 Examples of prompts manually designed by clinicians according to different datasets. (a) Atelectasis in the CheXpert Dataset. (b) COVID in the COVID Dataset. It can be seen that these text prompts are highly related to the characteristics of images in the dataset, and strongly dependent on domain experts and domain knowledge.

the most of the huge volume of unlabeled medical images and texts [5–7]. The large-scale vision-language models are pre-trained by predicting the correct matching between image and text pairs, and learn transferable image and text representations to support flexible downstream clinical task. Transferring the pre-trained models to specific downstream clinical tasks is often done with the help of relevant medical text prompts. Taking the large-scale medical pre-trained vision-language model MedCLIP [7] as an example, specifically, when using the pre-trained MedCLIP model for medical image classification, image and relevant medical text prompts are fed into the model’s image encoder and text encoder respectively to obtain the corresponding embeddings. Then, the model calculates the similarity between the image embeddings and the text embeddings, and takes the category corresponding to the text embeddings that leads to the highest similarity as the category of the image. The quality of the input medical text prompts, even a slight change in wording, has been proven to have a huge impact on model performance [8]. Therefore, there is currently a significant reliance on domain experts (clinicians) when generating these medical text prompts.

Figure 1 shows some examples of prompts carefully designed by clinicians according to different datasets. It can be seen that the prompts designed, such as disease subtypes, severity, and location, are highly related to the characteristics of images in the dataset, have significant differences in category and specific content, and are highly targeted and professional. Thus, manually generating text prompts is extremely time-consuming and highly dependent on domain experts and domain knowledge. The original intention of employing large-scale pre-trained vision-language models in the field of low-resource medical images is to minimize the pressure of extra annotations on clinicians. However, it is obviously contrary to the original intention to bring about additional work burden of manual prompts design. Therefore, it is necessary to design a model that can automatically generate high-quality medical text prompts to reduce the burden and dependence on clinicians.

To this end, we propose a weakly supervised prompt learning framework *MedPrompt*, which enables the model to directly learn from the datasets and automatically generate high-quality medical text prompts. The framework includes an unsupervised pre-trained vision-language model and a weakly supervised prompt learning model. The unsupervised pre-trained vision-language model adopts large-scale medical images and medical texts for pre-training, utilizing the natural correlation between medical images and corresponding medical texts without any manual annotation. The weakly supervised prompt learning model only utilizes the classes of images in the dataset to guide the learning of the specific class vector in the prompt, while other context vectors in prompt do not require any manually labeled guided signals. The framework can be trained to automatically generate high-quality prompts. Benefiting from these prompts, the pre-trained vision-language model utilizes the similarity between images embeddings and prompts embeddings to determine the categories of the images. Throughout the entire process, the framework is almost completely free from the expert reliance on manual annotation and manual prompts design, thus enabling truly end-to-end, low labor cost medical image learning. Experimental results of

few-shot image classification on four medical benchmark datasets with different categories show that for unseen classes classification, our model outperforms models with hand-crafted prompts, even when all samples are used for fine-tuning in these models. In addition, our model performs better in zero-shot image classification on most datasets than those with expert manual prompts. Besides, we find that prompts generated for different datasets and different categories have high semantic similarity, which reflects that our model can generate a common prompt that is generally applicable to all cases to some extent. In other words, the model extracts some kind of common knowledge when generating the prompt, which make it have the potential to be extended to a general multimodal medical recognition model. Moreover, the module used to automatically generate prompts is a lightweight module that does not impose too much additional network parameters and computational burden on the model, and can be seamlessly embedded into any network architecture. In summary, our work has three main contributions:

1. We propose a weakly supervised prompt learning framework *MedPrompt*, where the model can directly learn from the datasets and automatically generate effective prompts by using only class labels for inexact supervised learning, saving the cost and effort of experts to manually design prompt;
2. The performance of our model in few-shot and zero-shot classification on four datasets outperform that of the models with experts hand-crafted prompts;
3. The module used for automatically generating prompts lightweight that can be seamlessly embedded into any network architecture.

2 Related work

2.1 Medical vision-language models

Due to the small amount of image and text data available for pre-training in the medical field, compared to the already mature and widespread large-scale pre-trained vision-language models in general domains, the pre-trained vision-language models in the medical domain are still under exploration.

Following the general pre-trained vision-language models, existing medical image-text representation learning also adopts the framework of contrast learning. Among them, most of the work used strictly paired medical images and texts for contrast learning [5,6,9], which not only reduced the number of paired images and texts available for pre-training, but also introduced false negative noise during the training process. To solve these problems, Wang et al. [7] proposed to decouple the strong pairing relationship between image and text to obtain more usable training data and eliminate false negatives. Accordingly, a semantic matching loss was used to replace the commonly used InfoNCE loss in the original vision-language models. Following Wang's work, we also extend the pre-trained medical image and text pairs to enable the model to be pre-trained on more abundant large-scale data.

2.2 Prompt learning

Prompt learning was initially an important research hotspot in natural language processing. The motivation is to leverage the pre-trained language models (such as BERT [10] or GPT [11–14]) as knowledge bases from which prompt templates can be used to elicit information useful for downstream tasks [15]. After pre-training the model using a large amount of original text, a prompt function can be designed to adapt the pre-trained model to small or unlabeled data in other scenarios for few-shot learning or zero-shot inference.

Manual template design is commonly used in previous work, and the designed templates have achieved good performance in cloze test, question answering, translation, text classification, and conditional text generation tasks [13,16–19]. However, creating and verifying these prompts require time and experience, and even the most experienced prompt designers may not be able to manually discover the best prompts [20,21]. To solve this problem, many approaches have emerged for automatic prompt design. For example, Jiang et al. propose mining-based and paraphrasing-based methods to automatically generate high-quality and diverse prompts, which more accurately estimate the knowledge contained in language models [20]. Shin et al. develop an automated method based on a gradient-guided search to create prompts for a diverse set of tasks [22]. Other studies transform the prompt into a set of continuous vectors and optimize the objective function end to end [23–25].

CoOp [8] and CoCoOp [26] are the first to apply prompt learning to the adaptation of large vision-language models in computer vision. CoCoOp is a continuous prompt learning method that can auto-

matically construct conditional contextual prompts. It consists of a set of context prompt vectors and a lightweight neural network (Meta-Net), which is used to generate an input-conditional token (vector) for each image. CoCoOp has been proven to achieve excellent zero-shot inference performance on previously unseen categories under low resource conditions.

2.3 Weakly supervised learning

In many tasks, it is difficult to obtain strong supervision information due to the high cost of manual data labeling. Weakly supervised learning is proposed to be applied to such low resource scenarios which lack enough correct manual labeling, aiming at building prediction models through weak supervision signals. There are three typical types of weak supervision: incomplete supervision, inexact supervision, and inaccurate supervision [27]. Among them, inexact supervision means that the training data only has coarse-grained labels, without exactly the exact labels. In the prompt learning process of this work, we also do not have exact labels (exact natural language or semantic labels) for the training of context and class embeddings and the generation of the prompts. All we have for supervising model training are only the class labels. So we explore the feasibility and efficiency of adopting this weakly supervised learning approach in the low resource medicine field in this work.

2.4 Zero-shot learning and few-shot learning

Unlike the recognition pattern of traditional machine learning methods, zero-shot learning enables the AI model to mimic human reasoning by only training on base classes to recognize new classes that have never been seen before [28–31]. Zero-shot learning often requires auxiliary information (such as attributes of objects or related descriptions of objects, in our research problem, prompts describing pathological symptoms) to learn the semantic space. However, due to the “seen class bias” problem [30], zero-shot learning often fails to achieve satisfactory results. Few-shot learning provides another sample-efficient learning method. By utilizing prior knowledge, few-shot learning can quickly generalize to new tasks by only learning from a small number of samples [32]. This learning method, which learns from only a small amount of supervised samples, is more similar to that of human beings [33]. Zero-shot learning and few-shot learning are suitable for the medical field where the cost and threshold of manual labeling are high. In this study, we explore the zero-shot and few-shot learning performance of the model.

3 Method

3.1 Overall architecture

The overall architecture of our proposed model *MedPrompt* is shown in Figure 2. The whole model mainly includes two training stages: pre-training stage (dark gray line) and prompt learning stage (orange line). In the pre-training stage, a total of 600,526 X-ray images and 201,063 reports from the CheXpert and MIMIC-CXR datasets are extracted by a knowledge extractor (e.g. Negbio [34]) to obtain a ground-truth (GT) similarity, which is used to supervise the learning of the predicted similarity of embeddings obtained after the text and image encoders of the model. After such pre-training, the text encoder and image encoder of the model learn transferable representations, which can then be transferred to downstream task of image classification (green line). In the prompt learning stage, the model trains an instance-adaptive prompt generator with the help of the image embeddings output from the image encoder, and ultimately learns context embeddings and class embeddings on its own to form auto prompts embeddings. The auto prompts are the prompts required for the pre-trained model to perform downstream task, which can help the model in zero-shot image classification on images of completely unseen category. When performing few-shot image classification, the model fine-tunes the class embeddings of the prompt generator with very few images in the unseen category, with the rest of the prompt generator remaining fixed.

3.2 Pre-training

We employ images and reports from two large-scale medical imaging report datasets to pre-train a large-scale vision-language model. Following MedCLIP [7], we decouple each pair of images and their reports. Instead of taking the pairing of images and their reports as the objective of pre-training, we measure

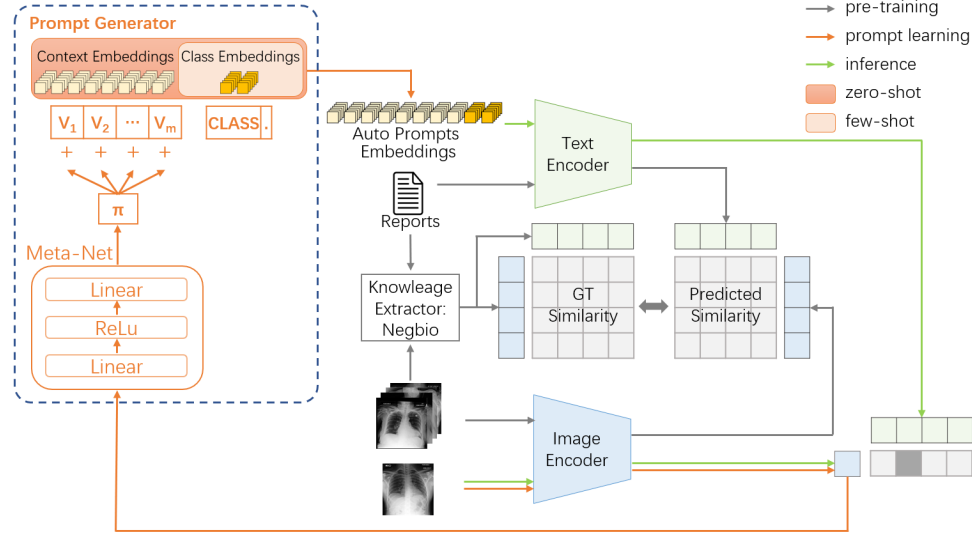


Figure 2 The overall architecture of our proposed model *MedPrompt*. The whole model mainly includes pre-training stage (dark gray line) and prompt learning stage (orange line). In the pre-training phase, the model learns transferable representations by matching the similarity between the paired images and the reports. In the prompt learning stage, the model trains an instance-adaptive prompt generator with the help of the embeddings of the images themselves and finally learns auto prompts embeddings for zero-shot image classification. When performing few-shot image classification, the model fine-tunes the class embeddings of the prompt generator with very few images in the unseen category, with the rest of the prompt generator remaining fixed. Best viewed in color.

the similarity of each image and each report in the datasets, and use the similarity matching as the pre-training objective. Specifically, for images, their diagnostic categories in the datasets are directly used as labels; for text, we leverage the Negbio [34] tool as the knowledge extractor to extract medical entities defined in the Unified Medical Language System [35] from the reports to construct text labels. Subsequently, the cosine similarity of each image label and each text label is calculated as the ground truth to guide the training of the pre-trained vision-language model.

Image encoder and text encoder. The original image I and report T are encoded into image embeddings and text embeddings by an image encoder E_I (e.g. ResNet or ViT) and a text encoder E_T (e.g. Transformer), respectively. A projection operation $proj(\cdot)$ is used for dimensional alignment to facilitate subsequent semantic similarity calculation. Thus, the multimodal embeddings I_e and T_e can be obtained:

$$I_e = proj(E_I(I)) \quad (1)$$

$$T_e = proj(E_T(T)) \quad (2)$$

Semantic similarity matching. The cosine similarity between image label I_{GT} and text label T_{GT} in the pre-trained datasets will be used as the ground-truth semantic S :

$$S = \frac{I_{GT} \cdot T_{GT}}{\|I_{GT}\| \|T_{GT}\|} \quad (3)$$

By calculating the similarity, for an image i , we can obtain a normalized similarity set, where the similarity between image i and prompt j can be represented as y_{ij} :

$$y_{ij} = \frac{\exp(S_{ij})}{\sum_{j=1}^{N_{text}} \exp(S_{ij})} \quad (4)$$

Similarly, the similarity \hat{S} between image embeddings I_e and report embeddings T_e can be calculated using the following formula:

$$\hat{S} = \frac{I_e \cdot T_e}{\|I_e\| \|T_e\|} \quad (5)$$

And the predicted similarity can be represented as \hat{y}_{ij} :

$$\hat{y}_{ij} = \frac{\exp(\hat{S}_{ij}/\tau)}{\sum_{j=1}^{N_{text}} \exp(\hat{S}_{ij}/\tau)} \quad (6)$$

, where τ is a learnable temperature parameter initialized at 0.07.

Loss. We aim to minimize the semantic loss between image and text. The semantic loss is calculated by the cross entropy loss, including an image-to-text semantic loss and a text-to-image semantic loss, which are averaged:

$$L = -\frac{1}{2} \left(\frac{1}{N_{img}} \sum_{i=1}^{N_{img}} \sum_{j=1}^{N_{text}} y_{ij} \log \hat{y}_{ij} + \frac{1}{N_{text}} \sum_{j=1}^{N_{text}} \sum_{i=1}^{N_{img}} y_{ji} \log \hat{y}_{ji} \right) \quad (7)$$

3.3 Prompt learning

After conducting the above pre-training, we fix the parameters of the text encoder and image encoder and proceed with prompt learning, aiming to learn a prompt generator that can generate auto prompts embeddings for each image.

The auto prompts generated by the model consist of several learnable words and one learnable class, which are finally presented in the form of embeddings. Inspired by CoCoOp [26], the prompt generator is trained to generate prompts based on specific instances rather than a general class, to learn richer transferable representations that can generalize better on unseen classes. Thus, the prompt generator consists of a learnable Meta-Net and the projection operations of context embeddings and class embeddings. Specifically, as shown in Figure 2, the features extracted by the image encoder for each instance are processed through a Meta-Net (a two-layer bottleneck structure with Linear-ReLU-Linear) to obtain a unique conditional token π for that instance. The conditional token is then combined with each context vector to form the final context vectors.

Therefore, for an image i and a class k , the auto prompt generated by the model can be denoted as:

$$p_{ik}^{auto} = \{MetaNet(proj(E_I(i)) + V, C_k)\} \quad (8)$$

, where $V = \{v_1, v_2, \dots, v_m\}$ are m word embeddings, each with a dimension of 512, which are randomly initialized. m is the number of context tokens in the prompt, which is set to 16 in this study.

Then, the generated auto prompt embeddings are feed to the text encoder of the model, and the similarity S_{ik}^{auto} between the generated new text embeddings $proj(E_T(p_{ik}^{auto}))$ and the embeddings $proj(E_I(i))$ of image i can be represented as:

$$S_{ik}^{auto} = \frac{proj(E_T(p_{ik}^{auto})) \cdot proj(E_I(i))}{\|proj(E_T(p_{ik}^{auto}))\| \|proj(E_I(i))\|} \quad (9)$$

Therefore, the probability of image i being predicted as class k can be expressed as:

$$\hat{y}_{ik} = \frac{\exp(S_{ik}^{auto}/\tau)}{\sum_{j=1}^{N_{class}} \exp(S_{ij}^{auto}/\tau)} \quad (10)$$

Then, cross entropy loss was calculated between the prediction probabilities of N_{class} classes and the GT class of each image, using formula (7).

Prompt learning is conducted before zero-shot inference and few-shot learning. Before performing zero-shot inference, the prompt generator is trained with the base classes, and then conducts zero-shot inference directly on the unseen classes. In this process, the parameters of text encoder and image encoder of the model are fixed, and Meta-Net and the linear layer of the projection operation mapped from v_1, v_2, \dots, v_m and *class* to the embeddings are trained. When performing few-shot learning, the linear layer of the projection operation mapped from *class* to the embeddings in the prompt generator are trained and fine-tuned directly on a few images (e.g. 1,2,4,8,16). In this process, the parameters of the text encoder and image encoder of the model are also fixed, and the base classes are also used to train the Meta-Net and the linear layer of the projection operation mapped from v_1, v_2, \dots, v_m and *class* to the embeddings. Then, a few samples from unseen classes are used to fine-tune the model. Note that only the linear layer corresponding to the *class* embedding is fine-tuned at this time.

It is worth mentioning that in the process of prompt learning, for the auto prompts to be generated, we do not have exact labels that can be used to train the context and class embeddings. What we can use to supervise model training are only the class labels, without any natural language or semantic labels. This is precisely where the low resource problem in the medical field lies - due to the high cost and threshold of labeling, there is too little available exact labels. However, in the absence of sufficient exact labels, our weakly supervised learning method, which only uses class labels for training, has achieved good results, generating auto prompts that are comparable to or even surpassing manual prompts. This effectively alleviates the problem of expensive and limited manual labeling and excessive reliance on domain experts in the low resource medical field.

4 Experiments

In this section, we evaluate *MedPrompt* on four different datasets on a downstream task of medical image classification. We validate the model's performance under zero-shot inference and few-shot learning. We also provide ablation studies to analyze the contribution of key factors of *MedPrompt*. In addition, we visualize and analyze the generated prompts.

4.1 Datasets

CheXpert [36] is a dataset containing 224,316 chest X-ray images, covering 65,240 patients who underwent radiologic examinations at Stanford Medical Center. The dataset includes training, validation and test sets, containing 14 observed labels. Among them, the training set includes three sets of labels automatically extracted from relevant radiology reports using various automatic labelers (CheXpert [36], CheXbert [37] and VisualCheXbert [38]). The labels for the validation and test sets are provided by certified radiologists. This dataset is of great significance for the automatic analysis and diagnosis of chest X-rays. The entire dataset participates in the unlabeled pre-training phase. To evaluate the performance of the model on zero-shot inference and few-shot learning, we follow Huang et al. [6] and sample a multi-class subset [7], namely CheXpert 5×200 , which contains five categories: Atelectasis, Cardiomegaly, Consolidation, Edema and Pleural Effusion, each with 200 positive samples.

MIMIC-CXR [39] is a large medical imaging dataset that includes 377,111 chest X-ray images and 201,063 corresponding radiological reports, as well as 14 observed labels. The dataset aims to support the research of image understanding, natural language processing and decision support in the medical domain. Similarly, the entire dataset is used for pre-training. In addition, for evaluation, we also sample a MIMIC-5 \times 200 subset in the same way as CheXpert-5 \times 200, with the same five categories.

COVID [40] is a publicly available X-ray image dataset on COVID-19 released by Rahman et al. in 2021. This dataset contains both COVID and non-COVID labels, with a positive to negative sample ratio of approximately 1:1. This dataset does not participate in pre-training, but only participates in prompt learning and is used for evaluation. There are 2,162 and 3,000 images in the training and test sets, respectively.

RSNA [41] is a chest X-ray image dataset publicly provided by the National Institutes of Health in the United States, which consists of pneumonia and normal samples. Following Wang et al. [7], we extracted a balanced subset with a 1:1 ratio of positive and negative samples. This dataset also does not participate in pre-training, but only participates in prompt learning and is used for evaluation. There are 8,486 and 3,538 images in the training and test sets, respectively.

4.2 Baselines

We compared our model with previous studies, including CLIP [4], ConVIRT [5], GLoRIA [6] and Med-CLIP [7].

CLIP is a pretrained model for matching images and text released by OpenAI in early 2021. It is pretrained directly on 4 million image-text pairs collected from the Internet and achieves state-of-the-art performance on many tasks on natural image datasets.

ConVIRT is dedicated to visual-text contrastive learning in medicine, learning medical visual representations directly from naturally paired medical image and text data. It is the pioneering work of multimodal contrastive learning in the medical field.

GLoRIA is also committed to visual-text contrastive learning in medicine, and uses attention mechanisms to learn the global-local representation of images by matching words and image subregions in radiology reports. For the first time, it realizes the prompt based zero-shot prediction for medical images after image-text pre-training.

MedCLIP employs an unsupervised contrastive learning approach on unpaired medical image-text data. It utilizes BioClinicalBERT [42] as text encoder and ViT as image encoder to construct the model, and employs soft semantic loss to supervise the training of the model.

4.3 Implementation details

We adopted the pre-trained text transformer from CLIP [4] as our text encoder. For vision encoder, we compared the CNN-based ResNet and the transformer-based Swin Transformer [43], and finally adopted Swin Transformer as the vision encoder.

We used the CheXpert and MIMIC-CXR datasets for pre-training, and for their text data, we split them into sentences and removed all sentences with length less than 4.

In the prompt learning stage, we held out five classes Atelectasis, Cardiomegaly, Consolidation, Edema and Pleural Effusion in the CheXpert and MIMIC datasets as unseen classes (to validate the zero-shot and few-shot performance of the models), and used the remaining nine classes as base classes to train the prompt generator.

We performed the following data augmentation on the images: scaling the original image to 256×256 , applying random crop with a size of 224×224 , horizontal flipping with a probability of 0.5, color jittering the brightness of the image to a random value within [80%, 120%], random affine transformation with degree sampled from [-10, 10].

We set the learning rate as $5e-5$ with a learning rate warmup ratio of 0.1, weight decay $1e-4$, batch size 400, and train the model for 20 epochs. A single half-precision pre-training on 4 V100 GPUs took about 10 hours.

4.4 Comparison to state-of-the-art

Zero-shot learning After training the prompt generator with the base classes, the model directly performs zero-shot inference on classes it has never seen before. Specifically, for the CheXpert and MIMIC-CXR datasets, there are 5 unseen categories including Atelectasis, Cardiomegaly, Consolidation, Edema and Pleural Effusion. For COVID and RSNA, the two categories in the datasets are completely new to the model. The process of performing zero-shot classification on an image is as follows: For each predicted class, the image will generate an auto prompt, which will be processed by a text encoder to obtain text embeddings of the image about each class. Then, the similarities between these text embeddings and the image embeddings of the image obtained from the image encoder are calculated. The class corresponding to the text embeddings with the highest similarity is the class of the image predicted by the model.

We compare the zero-shot classification performance of our auto prompts-based model with the previous models based on manual designed prompts. As can be seen from Table 1, our model achieves better results than all previous models on CheXpert- 5×200 , MIMIC- 5×200 and COVID, but performs slightly worse than MedCLIP on the RSNA dataset. We speculate that the unsatisfactory results on RSNA may be due to the fact that the labels of the base classes of CheXpert and MIMIC-CXR used to train the prompt generator are annotated by Negbio, while RSNA uses manual annotation labels, which may have gaps with Negbio's annotation results. It is worth mentioning that the model has never seen the COVID and RSNA datasets at all during the large-scale image-text pre-training stage and the prompt generator training stage, but still achieves satisfactory performance on these two datasets. Overall, it can be concluded that the prompts automatically generated by our model achieve better or comparable results in assisting zero-shot image classification than manual prompts.

To further analyze the performance of zero-shot classification, we explore the classification performance of its fully supervised counterpart, which is trained with all examples on the training set. As shown in Table 2, full supervision is naturally better than direct zero-shot inference, which is undoubtedly reflected in the experimental results. On all the four datasets, our model significantly outperforms the existing models, with the most significant improvement on the COVID dataset with a 16.63% improvement. In addition, by comparing the results of our model in Table 1 with those of the existing models in Table 2, we are pleasantly surprised to find that our model's performance in zero-shot classification without any

Method	CheXpert	MIMIC-CXR	COVID	RSNA
CLIP [4]	0.2036	0.2254	0.5090	0.5055
ConVIRT [5]	0.4224	0.4010	0.6647	0.4647
GLoRIA [6]	0.4210	0.3382	0.5702	0.4752
MedCLIP-ViT [7]	0.5942	0.5024	0.7943	0.7682
MedPrompt-ViT (Ours)	0.6220	0.5720	0.7997	0.7284

Table 1 Performance of zero-shot image classification on four datasets. For models with manually designed prompts, we only report the results of prompt ensemble. Best performance are in bold.

samples for fine-tuning has surpassed the fully supervised results of the existing models on CheXpert-5×200, MIMIC-5×200 and COVID, which reflects that the model has learned transferable representations and has strong generalization ability.

Method	CheXpert	MIMIC-CXR	COVID	RSNA
CLIP [4]	0.3020	0.2780	0.5866	0.7303
ConVIRT [5]	0.4770	0.4040	0.6983	0.7846
GLoRIA [6]	0.5370	0.3590	0.7623	0.7981
MedCLIP-ViT [7]	0.5960	0.5650	0.7890	0.8075
MedPrompt-ViT (Ours)	0.6580	0.6160	0.9553	0.8304

Table 2 Performance of fully supervised image classification on four datasets. Best performance are in bold.

In addition, we compare the accuracy of zero-shot classification of different manual prompts and auto prompts generated by our model on all four datasets, as shown in Figure 6 in Appendix A.

Few-shot learning In the few-shot learning phase, we first fix the pre-trained image encoder and text encoder, and then train the whole prompt generator with the base classes. Subsequently, we follow the few-shot evaluation protocol in CLIP by taking a few samples from the training set of unseen classes (e.g. 1,2,4,8,16) to fine-tune on class embeddings in the prompt generator. The model is deployed on the test set and the results are shown in Table 3.

Method	CheXpert	MIMIC-CXR	COVID	RSNA
MedPrompt-ViT 0-shot	0.6220	0.5720	0.7997	0.7284
MedPrompt-ViT 1-shot	0.6315	0.5895	0.8020	0.7538
MedPrompt-ViT 2-shot	0.6360	0.5875	0.8290	0.7665
MedPrompt-ViT 4-shot	0.6400	0.5870	0.8627	0.7761
MedPrompt-ViT 8-shot	0.6320	0.5815	0.8693	0.7778
MedPrompt-ViT 16-shot	0.6500	0.6000	0.8700	0.8013
MedPrompt-ViT full-shot	0.6580	0.6160	0.9553	0.8304

Table 3 Performance of few-shot learning of our model on four datasets. Best performance are in bold.

Comparing the results in Table 3 with the that of MedCLIP-ViT [7] in Table 1, it can be seen that with at most only 4 additional samples for fine-tuning, our auto prompts model outperforms the state-of-the-art manual prompts model on all datasets. This is very exciting news in the field of low-resource medical data, as it means that even in the absence of sufficient available and reliable expert annotations, we can still achieve exceptionally good performance on downstream tasks with the help of large-scale pre-training and the automatically generated prompts.

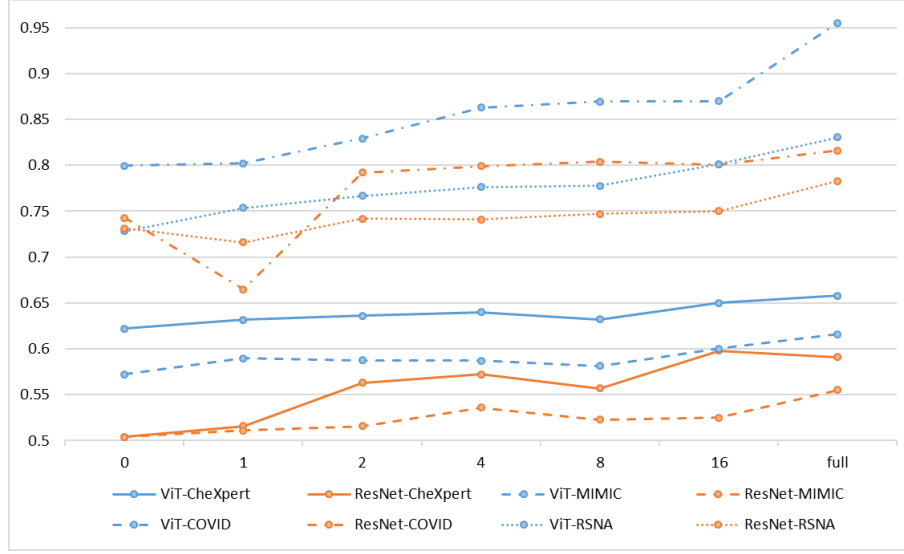


Figure 3 Performance comparison of our model adopting the traditional CNN-based architecture ResNet-50 and the Transformer-based architecture Swin Transformer as the image encoder on four datasets. The blue lines represent the ViT architecture and the orange lines represent the ResNet architecture. It can be seen that the models with ResNet-50 as the image encoder is overall inferior to its ViT counterparts. Best viewed in color.

4.5 Ablation studies

CNN-based architecture vs. transformer-based architecture We also verify the effects of adopting the classical CNN-based architecture ResNet50 as the image encoder. As shown in Figure 3, the effect of using ResNet50 as an image encoder is overall inferior to its ViT counterparts. We speculate that the main reasons are that, on the one hand, the transformer architecture has a good adaptation ability to big data, and its attention mechanism enables the model to fully learn the relationship between features. On the other hand, the shifted windows of swin transformer can better connect with the windows of the previous layer, which significantly enhances the modeling ability.

The effect of Meta-Net and context embeddings In addition, we verify the contribution of each component of the prompt generator. The learnable components of our prompt generator include Meta-Net, context embeddings and class embeddings. Since the downstream task is image classification, the class embeddings need to be learned in any case. We validate the effectiveness of learning only class embeddings, learning only context embeddings and class embeddings, learning only Meta-Net and class embeddings, and learning a complete prompt generator, respectively. Table 4 presents the average results of zero-shot learning, few-shot learning and full supervision on these four ablation versions over four datasets. We can conclude from the table that both the Meta-Net and the context embeddings contribute significantly to the performance improvement of the model. When the two are combined into the prompt generator, the performance of the model has made a great leap, with zero-shot performance improving by 3% and full supervision performance improving by 4.6%. We speculate that the main reason is that MetaNet has learned richer transferable representations from each specific instance, which helps to better generalize the model to unseen categories. The combination of these representations and context embeddings makes the generated auto prompts more adaptable to instances in previously unseen classes.

Meta-Net	Context Embeddings	Class Embeddings	0	1	2	4	8	16	full
		✓	0.6508	0.6877	0.6894	0.6886	0.6936	0.6994	0.7190
	✓	✓	0.6566	0.6738	0.6810	0.6943	0.7011	0.7021	0.7158
✓		✓	0.6675	0.6757	0.6793	0.6946	0.7005	0.7027	0.7193
✓	✓	✓	0.6805	0.6942	0.7048	0.7165	0.7152	0.7303	0.7649

Table 4 The average results of zero-shot learning, few-shot learning and full supervision for these components and their combinations over four datasets. The best results are highlighted in bold.

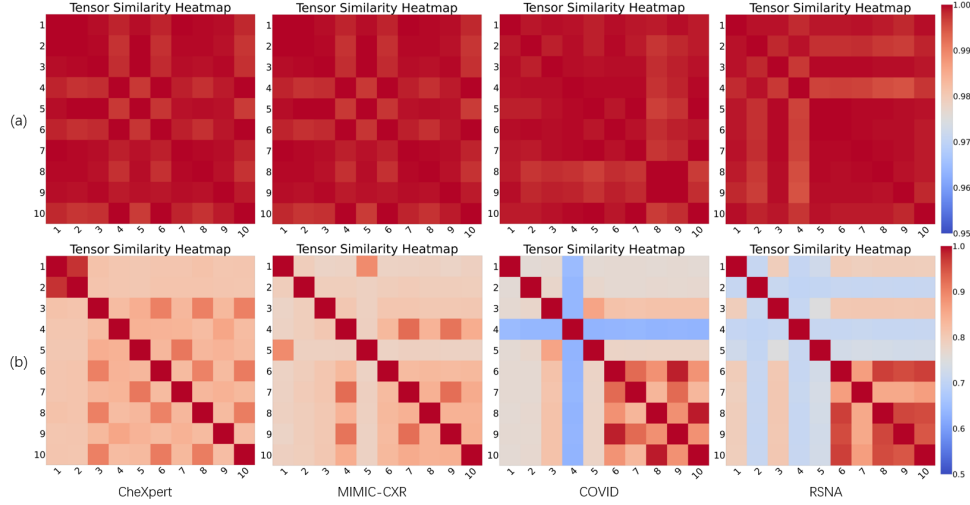


Figure 4 Visualization of prompts similarities on four datasets. (a) The similarities of prompts automatically generated by the model. (b) The similarities of manually designed prompts. The 10 images based on which the prompts are automatically generated by the model and the 10 manual prompts are randomly selected. Best viewed in color.

4.6 Further analysis

4.6.1 What do the generated auto prompts look like?

We randomly selected 10 images in each dataset and visualized the cosine similarity of the contexts embeddings (without the class embeddings) in the corresponding 10 prompts automatically generated by the model, as shown in Figure 4(a). We were surprised to find that the contexts generated by the model were surprisingly similar for different categories of images from different datasets. This indicates that our model has learned some common knowledge to some extent when generating auto prompts. This is very enlightening, indicating that our model may have the potential to be expanded to a general multimodal medical recognition model.

As a comparison, we also randomly selected 10 manually designed prompts for each dataset and calculated the cosine similarity between these prompts. As shown in Figure 4(b), the overall similarity of the manual prompts is relatively low, indicating that the manual prompts need to be carefully designed according to the characteristics of different categories of images in different datasets. Such a process undoubtedly relies excessively on the professional knowledge of domain experts and requires a lot of effort and time.

4.6.2 Limitation

Our approach can automatically generate high-quality prompts according to different datasets, thus relieving the pressure on clinicians to design prompts manually. However, it may still encounter failure when confronted with a completely new disease dataset. For example, the model does not perform particularly well in zero-shot classification on the COVID dataset that has not been seen at all during pre-training. This is mainly because the pre-trained model has not seen COVID samples, nor does it know the category of COVID. Therefore, similarly, it can be inferred that if there are other completely new diseases, the model may still not be able to identify them very well without sufficient samples for pre-training. The solution to this problem is to use more real-time updated datasets on the Internet for pre-training to improve the representation transferability of the model.

In addition, our method has only been validated on the medical image classification task, and lacks validation on other downstream clinical tasks such as medical image retrieval, segmentation, detection, and prognosis prediction. In future studies, we will verify the accuracy and generalization ability of the model in more tasks and datasets.

5 Conclusion

In this work, we propose a weakly supervised prompt learning method that can automatically generate medical text prompts for large-scale pre-trained vision-language models. On the one hand, the generated medical text prompts can assist in effectively transferring the pre-trained vision-language models to downstream clinical tasks such as medical image classification; On the other hand, the automatic generation of medical text prompts has freed clinicians from the pressure of manually designing prompts. Abundant experimental results indicate that our proposed method achieves state-of-the-art performance in few-shot and zero-shot classification of unseen classes. Moreover, the prompt learning method has the potential to be embedded in any network architecture seamlessly.

6 Appendix

A. Comparison of zero-shot classification accuracy of different human prompts and automatic prompts generated by our model on four datasets

Figure 5 presents the comparison of zero-shot classification accuracy of different manual prompts and auto prompts generated by our model on four datasets.





Dataset	Class	Prompt	Accuracy	Dataset	Class	Prompt	Accuracy
<div>CheXpert</div> 	Atelectasis	minimal retrocardiac atelectasis at the bilateral lung bases	0.6000	<div>MIMIC-CXR</div> 	Atelectasis	linear atelectasis at the left lung zone	0.5390
	Cardiomegaly	cardiac silhouette size is upper limits of normal			Cardiomegaly	ap erect chest radiograph demonstrates the heart size is the upper limits of normal	
	Consolidation	increased reticular consolidation at the right lower lobe			Consolidation	improved bilateral consolidation at the left upper lobe	
	Edema	mild pulmonary edema			Edema	moderate pulmonary interstitial edema	
	Pleural Effusion	small right bilateral pleural effusion			Pleural Effusion	right bilateral pleural effusion	
	Atelectasis	trace atelectasis at the right lung base	0.6030		Atelectasis	minimal bandlike atelectasis at the upper lung zone	0.5420
	Cardiomegaly	cardiomegaly which is unchanged			Cardiomegaly	heart size is borderline enlarged	
	Consolidation	improved partial consolidation at the left upper lobe			Consolidation	retrocardiac consolidation at the right lower lobe	
	Edema	improvement in pulmonary edema			Edema	moderate pulmonary edema	
	Pleural Effusion	stable right bilateral pleural effusion			Pleural Effusion	decreased right bilateral pleural effusion	
	Atelectasis	mild trace atelectasis at the left lung zone	0.6070		Atelectasis	residual atelectasis at the bilateral lung bases	0.5460
	Cardiomegaly	redemonstration of cardiomegaly			Cardiomegaly	cardiac silhouette size is mildly enlarged	
	Consolidation	partial consolidation at the left lower lobe			Consolidation	increased reticular consolidation at the lower lung zone	
	Edema	improvement in pulmonary edema			Edema	moderate trace interstitial edema	
	Pleural Effusion	decreased left subpulmonic pleural effusion			Pleural Effusion	left bilateral pleural effusion	
—	[V] ₁ [V] ₂ ... [V] _m [CLASS]	0.6220	—	[V] ₁ [V] ₂ ... [V] _m [CLASS]	0.5720		
Dataset	Class	Prompt	Accuracy	Dataset	Class	Prompt	Accuracy
<div>COVID</div> 	Normal	cardiomediastinal silhouette is stable given differences in positioning	0.5680	<div>RSNA</div> 	Normal	the cardiomediastinal silhouette is stable	0.4350
	COVID	patchy ground glass consolidation in peripheral			Pneumonia	mycoplasma at the left lower lobe	
	Normal	in addition , there is thickening along the right pleura at the anterior junction line	0.7540		Normal	there are a probable trace left pleural effusion	0.4859
	COVID	confluent ground glass consolidation in lower			Pneumonia	mycoplasma at the mid lung zone	
	Normal	multilevel degenerative changes with marginal osteophyte formation of the thoracic spine are noted with mild dextroconvex curvature of the mid-to-lower thoracic spine	0.7583		Normal	compared with the prior studies , mild pulmonary vascular congestion is new	0.7462
	COVID	patchy ground glass consolidation in mid			Pneumonia	viral at the lung bases	
—	[V] ₁ [V] ₂ ... [V] _m [CLASS]	0.7997	—	[V] ₁ [V] ₂ ... [V] _m [CLASS]	0.7284		

Figure 5 Comparison of zero-shot classification accuracy of different human prompts and automatic prompts generated by our model on four datasets

Acknowledgements This work was supported in part by the National Natural Science Foundation of China (Grant No. U1811461), the Key Areas Research and Development Program of Guangdong (Grant No. 2018B010109006), and Guangdong Introducing Innovative and Entrepreneurial Teams Program (Grant No. 2016ZT06D211).

Supporting information Appendix A. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

- 1 Esteva A, Chou K, Yeung S, et al. Deep learning-enabled medical computer vision. *NPJ Digit Med*, 2021, 4, (1): 5
- 2 Chen X X, Wang X M, Zhang K, et al. Recent advances and clinical applications of deep learning in medical image analysis. *Med Image Anal*, 2022, 79: 102444
- 3 Jiang H Y, Diao Z S, Shi T Y, et al. A review of deep learning-based multiple-lesion recognition from medical images: classification, detection and segmentation. *Comput Biol Med*, 2023, 157: 106726

- 4 Radford A, Kim J W, Hallacy C, et al. Learning Transferable Visual Models From Natural Language Supervision. In: Proceedings of International Conference on Machine Learning, 2021
- 5 Zhang Y H, Jiang H, Miura Y, et al. Contrastive Learning of Medical Visual Representations from Paired Images and Text. In: Proceedings of Machine Learning for Healthcare, 2022
- 6 Huang S C, Shen L, Lungren M P, et al. GLoRIA: A Multimodal Global-Local Representation Learning Framework for Label-efficient Medical Image Recognition. In: Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021: 3922-3931
- 7 Wang Z F, Wu Z B, Agarwal D, et al. MedCLIP: Contrastive Learning from Unpaired Medical Images and Text. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Abu Dhabi, United Arab Emirates, 2022: 3876-3887
- 8 Zhou K Y, Yang J K, Loy C C, et al. Learning to prompt for vision-language models. *Int J Comput Vis*, 2022, 130: 2337-2348
- 9 Wang X S, Xu Z Y, Tam L K, et al. Self-supervised image-text pre-training with mixed data in chest X-rays. *arXiv preprint arXiv:2103.16022*.
- 10 Devlin J, Chang M-W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, Minnesota, 2019: 4171-4186
- 11 Radford A, Narasimhan K. Improving Language Understanding by Generative Pre-Training, 2018
- 12 Radford A, Wu J, Child R, et al. Language Models are Unsupervised Multitask Learners, 2019
- 13 Brown T B, Mann B, Ryder N, et al. Language Models are Few-Shot Learners. *arXiv preprint arXiv:2005.14165*.
- 14 OpenAI. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.
- 15 Petroni F, Rocktäschel T, Lewis P, et al. Language Models as Knowledge Bases? In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019: 2463-2473
- 16 Chen M, Tworek J, Jun H, et al. Evaluating Large Language Models Trained on Code. *arXiv preprint arXiv:2107.03374*.
- 17 Schick T, Schütze H. Few-Shot Text Generation with Pattern-Exploiting Training. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021
- 18 Schick T, Schütze H. Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference. In: Proceedings of Conference of the European Chapter of the Association for Computational Linguistics, 2021: 255-269
- 19 Schick T, Schütze H. It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021: 2339-2352
- 20 Jiang Z B, Xu F F, Araki J, et al. How can we know what language models know? *Trans Assoc Comput Linguist.*, 2019, 8: 423-438
- 21 Liu P F, Yuan W Z, Fu J L, et al. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. *ACM Comput Surv*, 2021,55,(9): 1-35
- 22 Shin T, Razeghi Y, Logan I, et al. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 2020
- 23 Lester B, Al-Rfou R, Constant N. The Power of Scale for Parameter-Efficient Prompt Tuning. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021: 3045-3059
- 24 Li X L, Liang P. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In: Proceedings of Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021:
- 25 Zhong Z X, Friedman D, Chen D Q. Factual Probing Is [MASK]: Learning vs. Learning to Recall. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021: 5017-5033
- 26 Zhou K Y, Yang J K, Loy C C, et al. Conditional Prompt Learning for Vision-Language Models. In: Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022: 16795-16804
- 27 Zhou Z H. A brief introduction to weakly supervised learning. *Natl Sci Rev*, 2018, 5: 44-53
- 28 Chao W-L, Changpinyo S, Gong B, et al. An Empirical Study and Analysis of Generalized Zero-Shot Learning for Object Recognition in the Wild. In: Proceedings of European Conference on Computer Vision, 2016
- 29 Wang W, Zheng V W, Yu H, et al. A survey of zero-shot learning: settings, methods, and applications. *ACM Trans Intell Syst Technol*, 2019, 10, (2): 1-37
- 30 Xian Y Q, Schiele B, Akata Z. Zero-Shot Learning — The Good, the Bad and the Ugly. In: Proceedings of the IEEE Conference On Computer Vision And Pattern Recognition (CVPR), 2017: 3077-3086
- 31 Yi K, Shen X Q, Gou Y H, et al. Exploring Hierarchical Graph Representation for Large-Scale Zero-Shot Image Classification. In: Proceedings of European Conference on Computer Vision, Cham, 2022: 116-132
- 32 Wang Y Q, Yao Q M, Kwok J T-Y, et al. Generalizing from a few examples: a survey on few-shot learning. *ACM Comput Surv*, 2020, 1, (1): 1-34
- 33 Lake B M, Ullman T D, Tenenbaum J B, et al. Building machines that learn and think like people. *Behav Brain Sci*, 2016, 40
- 34 Peng Y F, Wang X S, Lu L, et al. NegBio: a high-performance tool for negation and uncertainty detection in radiology reports. *AMIA Jt Summits Transl Sci Proc*, 2018: 188-196
- 35 Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res*, 2004, 32(Database issue): D267-270
- 36 Irvin J A, Rajpurkar P, Ko M, et al. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. In: Proceedings of AAAI Conference on Artificial Intelligence, 2019
- 37 Smit A, Jain S, Rajpurkar P, et al. CheXbert: Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT. In: Proceedings of Conference on Empirical Methods in Natural Language Processing, 2020
- 38 Jain S, Smit A, Truong S, et al. VisualCheXbert: Addressing the Discrepancy between Radiology Report Labels and Image Labels. In: Proceedings of Proceedings of the Conference on Health, Inference, and Learning, 2021
- 39 Johnson A E W, Pollard T J, Berkowitz S J, et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 2019, 6, (1): 317
- 40 Rahman T, Khandakar A, Qiblawey Y, et al. Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images. *Comput Biol Med*, 2020, 132: 104319
- 41 Shih G, Wu C C, Halabi S S, et al. Augmenting the national institutes of health chest radiograph dataset with expert

- annotations of possible pneumonia. *Radiol Artif Intell*, 2019, 1, (1): e180041
- 42 Alsentzer E, Murphy J R, Boag W, et al. Publicly Available Clinical BERT Embeddings. In: *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, 2019: 72-78
- 43 Liu Z, Lin Y, Cao Y, et al. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In: *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021: 9992-10002