Medical Image Analysis 91 (2024) 103018

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Mapping medical image-text to a joint space via masked modeling

Zhihong Chen [a,b], Yuhao Du [a,b], Jinpeng Hu [a,b], Yang Liu [a,b], Guanbin Li [c,*], Xiang Wan [a,b], Tsung-Hui Chang [a,b]

[a] The Chinese University of Hong Kong, Shenzhen, 518172, China
[b] Shenzhen Research Institute of Big Data, Shenzhen, 518172, China
[c] Sun Yat-sen University, Guangzhou, 510275, China

ARTICLE INFO

MSC:
41A05
41A10
65D05
65D17

Keywords:
Multi-modal pre-training
Masked autoencoders
Medical vision-and-language analysis

ABSTRACT

Recently, masked autoencoders have demonstrated their feasibility in extracting effective image and text features (e.g., BERT for natural language processing (NLP) and MAE in computer vision (CV)). This study investigates the potential of applying these techniques to vision-and-language representation learning in the medical domain. To this end, we introduce a self-supervised learning paradigm, multi-modal masked autoencoders (M^3AE). It learns to map medical images and texts to a joint space by reconstructing pixels and tokens from randomly masked images and texts. Specifically, we design this approach from three aspects: First, taking into account the varying information densities of vision and language, we employ distinct masking ratios for input images and text, with a notably higher masking ratio for images; Second, we utilize visual and textual features from different layers for reconstruction to address varying levels of abstraction in vision and language; Third, we develop different designs for vision and language decoders. We establish a medical vision-and-language benchmark to conduct an extensive evaluation. Our experimental results exhibit the effectiveness of the proposed method, achieving state-of-the-art results on all downstream tasks. Further analyses validate the effectiveness of the various components and discuss the limitations of the proposed approach. The source code is available at https://github.com/zhjohnchan/M3AE.

1. Introduction

Medical data is inherently multi-modal, including tabular data, time-series data, imaging data, text data, and structured data (Acosta et al., 2022; Moor et al., 2023). Among them, imaging and text data are two critical ones, where for the former, radiography, magnetic resonance imaging, and computed tomography are crucial for understanding the structural and functional aspects of the human body; for the latter, radiology reports and medical texts provide critical insights into the patient's medical history, symptoms, and diagnoses. Mapping the data to the joint space can lead to a holistic understanding of medical images and texts. Yet, it is challenging due to the heterogeneity of data of different modalities.

To address the challenge of understanding medical data, medical vision-and-language pre-training (Med-VLP) has emerged as a crucial technique. Med-VLP aims to learn generic representations from large-scale medical image-text data, which can be transferred to various medical vision-and-language tasks. These tasks include medical visual question answering (Med-VQA), which requires answering questions based on visual and textual information from medical image-text pairs; medical image-text classification, which involves categorizing images and associated texts based on their medical conditions; and medical image-text retrieval, which involves retrieving relevant medical images and texts based on given queries. Med-VLP has become essential for jointly understanding medical images and texts, especially given the limited availability of large-scale labeled data and domain knowledge.

While vision-and-language pre-training (VLP) has received sustained attention (Chen et al., 2020b; Huang et al., 2020; Kim et al., 2021; Su et al., 2019; Tan and Bansal, 2019), the application of this technique to the medical domain has been limited to a few studies. For example, Li et al. (2020a) applied four VLP models, namely LXMERT (Tan and Bansal, 2019), VisualBERT (Li et al., 2019), UNITER (Chen et al., 2020b), and PixelBERT (Huang et al., 2020), to a medical image-text classification task, but found that the models performed worse than in the general domain without incorporating domain-specific information. To address this issue, Khare et al. (2021) proposed to perform pre-training on medical image-text pairs to capture medical knowledge. However, they only evaluated the approach on Med-VQA and failed to explore its promising improvement. The most related work to ours is Moon et al. (2021), which performed pre-training of a Med-VLP model and demonstrated its effectiveness on

* Corresponding author.
E-mail addresses: liguanbin@mail.sysu.edu.cn (G. Li), wanxiang@sribd.cn (X. Wan).

https://doi.org/10.1016/j.media.2023.103018
Received 10 April 2023; Received in revised form 1 October 2023; Accepted 30 October 2023
Available online 4 November 2023

various downstream tasks, but the pre-training was limited to chest X-rays and was not performed in a self-supervised manner using diagnosis labels. Previous studies have mainly used convolutional neural networks (CNNs) as their visual backbones, limiting their simplicity and effectiveness and ignoring purely Transformer-based models (Vaswani et al., 2017). Therefore, it is essential to develop an appropriate Med-VLP approach, considering four perspectives: data (e.g., pre-training corpus), models (e.g., purely Transformer-based models), objectives (e.g., more suitable pre-training objectives), and evaluation (e.g., designs of the downstream benchmark), to promote the application of Med-VLP to medical data.

In this paper, we propose to map medical images and texts to a joint space using a multi-modal masked autoencoder (M^3AE) based on purely Transformer-based models. Our approach is designed to learn cross-modal domain-specific knowledge from large-scale medical image-text datasets in a self-supervised manner without requiring fine-grained annotations on either images or texts, making it highly applicable in the medical domain. The M^3AE works by randomly masking patches of the input image and tokens of the input text and reconstructing the missing pixels and tokens. We develop the M^3AE design from three perspectives. First, we use different masking ratios for the input images and texts, considering the different information densities of vision and language. Second, we select visual and textual features from distinct layers to perform the construction, taking into account the different levels of abstraction in vision and language. Third, we use two different decoder designs for vision and language, where a Transformer model and a multi-layer perceptron (MLP) are used for vision and language decoding, respectively. We perform pre-training on two large-scale medical image-text datasets, namely ROCO (Pelka et al., 2018) and MedICaT (Subramanian et al., 2020), and evaluate the effectiveness of our approach on a medical vision-and-language understanding benchmark, which includes three tasks: Med-VQA, medical image-text classification, and medical image-text retrieval. Experimental results show that our approach outperforms previous studies on all downstream tasks. Furthermore, we conduct several analyses to examine the effectiveness of different components and various pre-training settings and to discuss the limitations of the proposed approach.

The contributions of this work are from three perspectives:

- **Problem to be solved**: We focus on a more general problem in our paper, where the goal is to learn generic vision-and-language representations for medical images and texts that can be transferred to many downstream tasks. To this end, we design the entire pipeline in the medical domain, including the pre-training data, the model/algorithm, and the evaluation benchmark.
- **Technical novelty**: Technically, we propose a simple yet effective approach for medical vision-and-language understanding through several designs using a purely Transformer-based architecture.
- **Effectiveness**: Our proposed approach outperforms existing studies on all downstream tasks. Besides, we analyze different components of the approach.

2. Related work

In this section, we review the literature related to three topics: (i) medical vision-and-language tasks, (ii) general-domain vision-and-language representation learning, (iii) masked language/image modeling, and (iv) medical-domain vision-and-language representation learning.

2.1. Medical vision-and-language tasks

There are many tasks in the medical domain involving both vision and language modalities. In general, such tasks could be divided into two types: vision-and-language understanding and vision-and-language generation. For vision-and-language understanding, the most straightforward one is text-assisted image classification, which improves the performance of medical image classification with extra information (e.g., patient history and previous studies). For example, Li et al. (2020a) and Monajatipoor et al. (2022) testified the performance of existing vision-and-language models on the Chest X-ray disease diagnosis and showed that the textual modality improved the performance significantly. Another typical task is medical visual question answering (Nguyen et al., 2019; Do et al., 2021; Seenivasan et al., 2022). This task requires models to answer a medical question related to the image, the application of which can improve the interaction between machines and patients. Besides, medical image-text retrieval (Subramanian et al., 2020) is also an important vision-and-language understanding task, where connections between the text and figures are useful to enable the retrieval of figures via textual queries and to produce systems that are capable of analyzing and understanding medical images. For vision-and-language generation, medical report generation (Shin et al., 2016; Jing et al., 2018; Chen et al., 2020d; Liu et al., 2021a; Wang et al., 2021; Yan et al., 2021; Najdenkoska et al., 2022) and text-to-image synthesis (Chambon et al., 2022) are two popular tasks, where the former aims to automatically generate a report for a given radiology image and the latter targets at synthesizing medical images given a description text.

2.2. Masked language/image modeling

Masked Language Modeling (MLM) aims to curate training samples for text models from a large-scale unannotated corpus. One of the early studies is Word2Vec (Mikolov et al., 2013), which is trained to predict the center or neighboring words given the context. Then, BERT (Devlin et al., 2019) with the Transformer architecture and the MIM pretext task achieves great improvement in various tasks by scaling up the pre-training scale (including the parameters of the models and the number of training samples). After that, the paradigm of natural language processing (NLP) was shifted to the pretrain-then-finetune paradigm. Following BERT, different strategies (including data and pretext tasks) for the improvement of MLM were proposed (Liu et al., 2019; Yang et al., 2019; Clark et al., 2019; Dong et al., 2019). While the masking ratio of 15% is the choice of most of the work, there is also a study (Wettig et al., 2023) indicating that a larger masking ratio is beneficial for larger models.

Following MLM, Masked Image Modeling (MIM) aims to learn effective visual representations via predicting the masked visual content in a self-supervised manner, which can be traced back to ViT (Dosovitskiy et al., 2020) and iGPT (Chen et al., 2020c). Subsequently, the BEIT series (BEIT-1 (Bao et al., 2021) and BEIT-2 (Peng et al., 2022)) greatly improve the performance of MIM in the downstream tasks by discreting the continous image pixels to a sequence of visual tokens. Another research line in MIM is to explore the reconstruction of the pixels directly. The most classical work is MAE (He et al., 2021), where the authors found that the key to pixel-target MIM is to use a large masking ratio for images (i.e., 75% in their study). Afterwards, many studies are starting to dive into the pixel/feature regression (Zhang et al., 2022b; Wei et al., 2022c; Zhou et al., 2021; Xie et al., 2022; Wei et al., 2022a,b; Huang et al., 2022; Gao et al., 2022; Dong et al., 2023; Chen et al., 2023; Fang et al., 2023; Li et al., 2022; Wang et al., 2023).

2.3. Vision-and-language pre-training

Driven by the effectiveness of self-supervised pre-training approaches in NLP (e.g., BERT Devlin et al., 2019) and computer vision (CV) (e.g., SimCLR Chen et al., 2020a and MoCo He et al., 2020), there is a growing interest in creating VLP techniques to tackle a broad array of vision-and-language-related challenges. Generally, VLP methods fall into two groups based on the vision-and-language interplay: dual-encoder and fusion-encoder. Present dual-encoder strategies can be outlined by the following factors: (i) utilizing medium-scale curated image-text data (Radford et al., 2021), (ii) employing large-scale

noisy image-text data (Jia et al., 2021), (iii) devising more intricate image-text contrasts (Yao et al., 2022), (iv) implementing additional single-modal contrastive learning (Mu et al., 2022). As for fusion-encoder approaches, existing research can be divided according to these three viewpoints: (i) Uni-modal encoders: various methods employ different image features (for example, the region features Li et al., 2019; Lu et al., 2019, patch embeddings Kim et al., 2021, and grid features Huang et al., 2020) and unique text features (for instance, statistic embeddings Kim et al., 2021, and dynamic embeddings Dou et al., 2021); (ii) Multi-modal fusion modules: existing research utilized the single-stream fusion scheme (Su et al., 2019; Li et al., 2020b) or dual-stream fusion scheme (Tan and Bansal, 2019; Yu et al., 2021); (iii) Pretext tasks: existing research investigates a range of pre-training tasks, including masked language modeling (Li et al., 2019), masked image modeling (Lu et al., 2019; Chen et al., 2020b), image-text matching (Zhang et al., 2021).

2.4. Medical vision-and-language pre-training

As an application and extension of VLP to the medical field, Med-VLP focuses on comprehending the content of medical images and texts. This can be traced back to Zhang et al. (2022a) for dual-encoders and Li et al. (2020a) for fusion-encoders. In the case of dual-encoders, subsequent research (Huang et al., 2021; Müller et al., 2021; Wang et al., 2022b) delved into global-local image-text contrastive learning to obtain more detailed information from medical images and texts, achieving top results in the medical image classification task. Concerning fusion-encoders, Li et al. (2020a) explored the performance of four vision-and-language models pre-trained in the general domain on a disease classification task. Then MMBERT (Khare et al., 2021) and MedViLL (Moon et al., 2021) performed pre-training on medical image-text data before fine-tuning models on the downstream tasks. Moreover, Chen et al. (2022) incorporated medical knowledge into the pre-training process to boost performance on downstream medical tasks.

2.5. The relationship to existing studies

The work most related to ours is BEIT-3 (Wang et al., 2022a). In BEIT-3, the authors used the image tokenizer from BEIT-2 to discrete the images so that a simple masked token modeling could be applied to learning the joint space of images and texts. Besides, scaling up the data and model scales is another important contribution to making the model achieve great performance on a broad range of datasets/tasks. We are different from BEIT-3 from three perspectives: (i) BEIT-3 used an image tokenizer to discrete the image to a sequence of tokens as the reconstruction target, whereas ours recovers the pixel directly; (ii) BEIT-3 also recovers the masked image-only and text-only inputs, which makes it good at encoding images and texts separately, whereas ours only targets at the joint representations of images and texts; (iii) BEIT-3 used large-scale data for representation learning, whereas we can only gather a much smaller dataset in the medical domain. Therefore, it could be seen that there is a huge difference between the general and medical domains, and our study could be a proof-of-concept of vision-and-language pre-training in the medical domain that builds up the pipeline for medical vision-and-language pre-training.

In addition to the idea of the approach, we borrow the experience from the general domain to develop our detailed implementation. In specific, we decided on the masking ratios of our approach largely based on the experimental findings from BERT in NLP and MAE in CV (i.e., a masking ratio of 75% for images and 15% for texts). Besides, we referred to their studies to choose the types of decoders for images and texts.

3. The proposed approach

In this section, we first formulate the problem to be solved in Section 3.1. Then, the backbone model architecture is detailed in Section 3.2. Finally, we introduce the multi-modal masked modeling for mapping medical image-text to a joint space in Section 3.3.

3.1. Problem formulation

We employ the pre-training-and-fine-tuning methodology in the context of medical vision-and-language comprehension. During the pre-training phase, the framework establishes various pretext tasks to train the model utilizing medical image-text pairs. Formally, given a medical image I and its associated descriptive text T, the model's training aims to minimize the objective via

$$\theta^*, \theta_1^*, \ldots, \theta_S^* = \arg\min_{\theta, \theta_1, \ldots, \theta_S} \sum_{s=1}^{S} L_s(Y_s, \mathcal{D}_{\theta_s}(\mathcal{M}_\theta(I, T))), \tag{1}$$

where S is the number of pretext tasks, L_s are the loss functions of pretext tasks, $\mathcal{D}_{\theta_s}$ are the decoders with their parameters $\theta_1, \ldots, \theta_S$, and $\mathcal{M}_\theta$ is the backbone model with its parameters θ. In the following subsections, we detail the designs of $\mathcal{M}_\theta$ in Section 3.2 and pretext tasks[1] in Section 3.3. During the fine-tuning phase, the learned model is applied to performing different downstream tasks by exploiting the pre-trained weights. A comprehensive overview of the proposed methodology is depicted in Fig. 1.

3.2. The backbone model architecture $\mathcal{M}_\theta$

Our backbone model can be divided into three key components: the vision encoder for encoding the input image, the language encoder for encoding the input text, and the multi-modal fusion module for interacting with the extracted visual and textual features.

Vision encoder. For simplicity and effectiveness, we focus on purely Transformer-based models and study the use of a vision Transformer (ViT) for the vision encoder in this paper. Specifically, in ViT, an image $I \in \mathbb{R}^{H\times W\times C}$ is first segmented into patches $\{p_1, p_2, \ldots, p_N\}$, where $H \times W$ is the image resolution, C is the number of channels, $p_n \in \mathbb{R}^{P^2\times C}$ and $P \times P$ is the patch resolution. Subsequently, the patches are flattened and linearly projected into patch embeddings via a linear transformation $E^v \in \mathbb{R}^{P^2C\times D}$, with an additional learnable token embedding $p_I \in \mathbb{R}^D$ introduced for visual information aggregation. Afterwards, the input representations are obtained by summing up the patch embeddings and learnable 1D position embeddings $E^v_{pos} \in \mathbb{R}^{(N+1)\times D}$:

$$X^v = [p_I; p_1E^v; p_2E^v; \ldots; p_NE^v] + E^v_{pos}. \tag{2}$$

Finally, X^v is input into a transformer model comprising Nv Transformer layers to acquire the contextualized image representations $H^v = [h^v_I; h^v_1; h^v_2; \ldots; h^v_N]$.

Language encoder. In the language encoder, we follow BERT (Devlin et al., 2019) to tokenize the input text to subword tokens $\{w_1, w_2, \ldots, w_M\}$ by WordPiece (Wu et al., 2016), where the tokens $w_m \in \mathbb{R}^V$ are represented in one-hot form and V is the vocabulary size. Then the tokens are linearly projected into embeddings through a linear transformation $E^l \in \mathbb{R}^{V\times D}$. Afterwards, a start-of-sequence token embedding $w_T \in \mathbb{R}^D$ and a special boundary token embedding $w_{SEP} \in \mathbb{R}^D$ are added to the text sequence. Therefore, the text input representations

[1] We mainly discuss the design of masked autoencoders and omit the description of image-text matching that commonly adopted in previous VLP studies.

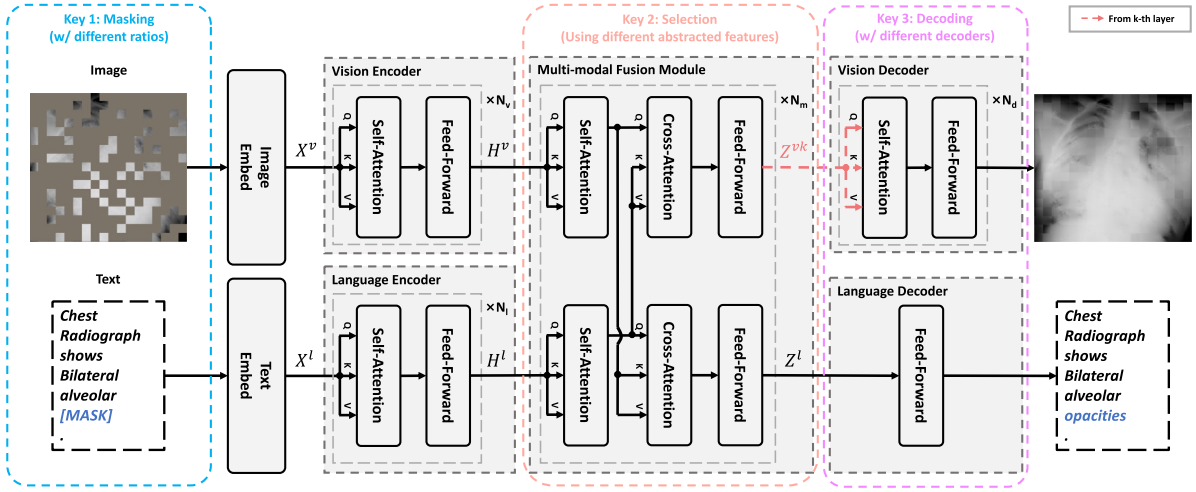

Fig. 1. The overall architecture of our proposed approach, where the masked inputs, uni-modal encoders, the multi-modal fusion module, and decoders are shown in gray dash boxes. Three key components are shown in three colored dash boxes. Note that the input image and text are masked separately in different forward processes.

are computed via summing up the token embeddings and text position embeddings $E^l_{pos} \in \mathbb{R}^{(M+2)\times D}$:

$$X^l = [w_T; w_1 E^l; ...; w_M E^l; w_{SEP}] + E^l_{pos}. \tag{3}$$

Similarly, X^l is fed into a transformer model with N_l Transformer layers to obtain the contextualized text representations $H^l = [h^l_T; h^l_1; h^l_2; ...; h^l_M; h^l_{SEP}]$.

Multi-modal fusion module. We adopt the co-attention mechanism in the multi-modal fusion module to fuse the contextualized representations from images and texts. In detail, the multi-modal fusion module consists of two Transformer models, each of which is a stack of N_m Transformer layers. In each Transformer layer, there are three sub-layers, i.e., a self-attention sub-layer, a cross-attention sub-layer, and a feedforward sub-layer. The attention mechanism is applied in the self-attention and cross-attention sub-layers, and it is defined as

$$\text{ATTN}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) \cdot V, \tag{4}$$

where d_k is the dimension of K. In the self-attention sub-layer, the representations interact within modalities:

$$\begin{aligned} H^{vs} &= \text{ATTN}(H^v, H^v, H^v), \\ H^{ls} &= \text{ATTN}(H^l, H^l, H^l). \end{aligned} \tag{5}$$

In the cross-attention sub-layer, the representations interact across modalities to integrate cross-modal information into their representations:

$$\begin{aligned} H^{vc} &= \text{ATTN}(H^{vs}, H^{ls}, H^{ls}), \\ H^{lc} &= \text{ATTN}(H^{ls}, H^{vs}, H^{vs}). \end{aligned} \tag{6}$$

Finally, H^{vc} and H^{lc} are input to the feedforward sub-layer (i.e., an MLP) to obtain the multi-modal representations $Z^v = [z^v_I; z^v_1; z^v_2; ...; z^v_N]$ for vision and $Z^l = [z^l_T; z^l_1; z^l_2; ...; z^l_M; z^l_{SEP}]$ for language.

3.3. Multi-modal masked autoencoders

The concept of masked autoencoders has seen tremendous success in natural language processing, exemplified by models like BERT and, more recently, in computer vision with models such as MAE (He et al., 2021). In the broader domain of vision and language pre-training (VLP), existing research (Dou et al., 2021; Kim et al., 2021) has primarily focused on recovering the original tokens of masked text, a process known as masked language modeling (MLM). However, it has been shown that attempting to reconstruct the original signals of masked images, referred to as masked image modeling (MIM), can negatively impact pre-training performance. This disparity in performance can be attributed to the differing characteristics of vision and language, necessitating specific design adaptations for masked autoencoders to function effectively in a multi-modal context. To address these challenges, we propose three essential and straightforward design modifications that can help bridge the gap between the two modalities and optimize the performance of masked autoencoders in both vision and language tasks.

Masking strategy. The information density between vision and language differs significantly. Languages, which are information-dense messages created by humans, can present a sophisticated language understanding task by predicting just a few held-out tokens. In contrast, images exhibit spatial redundancy, meaning that a missing patch can often be easily reconstructed from visible neighboring patches. Consequently, we employ a random sampling approach with a much higher masking ratio for images (75%) than texts (15%), where we decide the value of masking ratios mainly according to the experimental findings in BERT in NLP and MAE in CV. This strategy helps remove image redundancy and enables the model to extract valuable features from images and texts effectively.

Representation selection for reconstruction. Images and texts are abstracted at different semantic levels; image pixels are at a lower semantic level than text tokens. In our model, we aggregate their representations layer-by-layer using a hierarchical approach. To ensure that the final learned representations of images are semantically rich, we use the intermediate outputs of the multi-modal fusion module (specifically, the visual outputs from the kth Transformer layer, denoted as Z^{vk}) for the low-level construction task, MIM. This allows the model to focus on capturing more abstract features from the images. For MLM, we retain the final output Z^l for predicting tokens since reconstructing the missing words requires a higher semantic information level.

Decoder designs. The vision and language decoders' primary function is to map the high-level semantic representations Z^{vk} and Z^l back to their original input forms (image and text, respectively). For the vision decoder, the output is required to be in pixel space, which inherently has a lower semantic level. To achieve this, we introduce a Transformer model as the decoder, designed to map the high-level Z^{vk} representations to lower semantic representations. This enables the vision decoder to perform low-level reconstruction, effectively recreating the original images. The language decoder's targets (words) are abstracted at a higher semantic level, making the design more straightforward by using an MLP. The MIM loss is calculated using the mean squared error (MSE) between the reconstructed and original images in pixel space, while the

MLM loss is computed as the negative log-likelihood loss for the masked tokens.

It is essential to note that the MLM and MIM tasks are performed separately in different forward procedures to ensure the model's effectiveness in learning from both modalities. Specifically, instead of feeding masked images and masked texts (i.e., *Input: Image (masked) and Texts (masked)*) to the model in one forward process, we feed masked images and texts (i.e., *Input: Image (masked) and Texts (unmasked)*) to the model to obtain the MIM loss and feed images and masked texts to (i.e., *Input: Image (unmasked) and Texts (masked)*) to get the MLM loss. The advantage is that we can force the model to learn the relationship between the visual and textual spaces and map them to the joint spaces. For example, if a phrase '*pleural effusion*' is masked in the text, the model can learn to "look at" the image to find the clue to recover the masked texts. We have revised the corresponding description to make it more clear.

4. Experimental settings

In this section, we detail the pre-training setup in Section 4.1 and the downstream evaluation in Section 4.2.

4.1. Pre-training setup

Our experiments were conducted on two publicly available datasets, i.e., ROCO (Pelka et al., 2018) and MedICaT (Subramanian et al., 2020). The ROCO dataset consists of over 81,000 medical image-text pairs, where each image is accompanied by a corresponding textual description. The MedICaT dataset, on the other hand, contains over 217,000 medical images with captions and inline textual references. To train and evaluate our models, we used ROCO's official splits, while in MedICaT, we randomly sampled 1000 images for validation and 1000 images for testing and used the remaining images for training. For pre-training, we used the training set of both datasets to train our models with the pre-training tasks presented in Section 3 in addition to the common image-text matching task (Chen et al., 2020b), which aims to predict whether a given image and its textual description are semantically aligned.

The architecture we used for the vision encoder was CLIP-ViT-B (Radford et al., 2021), while RoBERTa-base (Liu et al., 2019) was used for the language encoder. The multi-modal fusion module was composed of 6 Transformer layers with a hidden state dimension of 768 and 12 heads. For all pre-training experiments, we trained the models with the AdamW optimizer (Loshchilov and Hutter, 2018) for 100,000 steps. The learning rates for uni-modal encoders (i.e., the vision encoder and the language encoder) were set to 1e-5, while the learning rate for the multi-modal fusion module was set to 5e-5. We set the warm-up ratio to 10% and used a linear learning rate scheduler after warm-up. We used center-crop to resize each image to 288×288.

4.2. Vision-and-language transfer tasks

The evaluation of our models was conducted on three medical image-text understanding tasks: medical visual question answering (Med-VQA), medical image-text classification, and medical image-text retrieval. These tasks evaluate the ability of our models to understand the relationship between medical images and their associated textual descriptions and to perform different downstream tasks based on this understanding.

Medical visual question answering. This task evaluates the ability of our models to answer clinical-related questions according to medical images. To train and evaluate our models, we used three publicly available datasets: VQA-RAD (Lau et al., 2018), SLAKE (Liu et al., 2021c), and MedVQA-2019 (Abacha et al., 2019), and we adopted their official dataset splits. In the VQA-RAD and SLAKE datasets, the questions are categorized into two types: closed-ended and open-ended. Closed-ended questions have a fixed set of answer choices, while open-ended questions require the model to generate an answer from scratch. To fine-tune the models on this task, we regard it as a multi-label classification task and feed the concatenation of the image and text representations to a two-layer MLP to predict the corresponding answer. The models are trained with a binary cross-entropy loss with a batch size of 64.

Medical image-text classification. This task requires our models to predict the label associated with the given medical image and its corresponding text. To train and evaluate our models on this task, we used the MELINDA dataset (Wu et al., 2021), a biomedical experiment method classification dataset with the official split. To fine-tune the models on this task, we learn a two-layer MLP on top of the concatenation of the image and text representations. We train the models with a cross-entropy loss with a batch size of 16 over a maximum of 20 epochs.

Medical image-text retrieval. The medical image-text retrieval task consists of two subtasks: image-to-text (I2T) retrieval and text-to-image (T2I) retrieval. In the I2T subtask, the goal is to retrieve the most relevant texts from a large pool of texts given an image. Conversely, the T2I subtask aims to retrieve the most relevant images given a text query. We trained and evaluated our models on the official split of the ROCO dataset. To fine-tune the models on this task, we initialize the similarity score head from the pre-trained ITM head. The model is tuned with cross-entropy loss to maximize the scores on positive pairs with 15 random texts sampled as negative samples with a batch size of 256 over a maximum of 10 epochs. During the evaluation, we sample 2,000 image-text pairs from the ROCO test set and report the results on the sampled 2,000 image-text pairs due to the large time complexity of the ranking process.[2] For this task, we adopt two settings for evaluating the models (similar to the studies (Kim et al., 2021; Dou et al., 2021) in the general domain). Specifically, in the zero-shot setting, we directly applied the pre-trained models to perform the image-to-text/text-to-image retrieval task without further fine-tuning; In the fine-tuning setting, we train the model specific for the retrieval task with its corresponding training samples and then evaluate the fine-tuned model.

For the metrics, we selected them based on the types of the tasks. For the Med-VQA and medical text-image classification tasks, we evaluated the models based on their accuracy. On the other hand, for the retrieval task, we used Recall@K (K = 1, 5, 10) as our evaluation metric. We ran each experiment three times with different random seeds and reported the mean value of the corresponding metric(s).

To better demonstrate the effectiveness of the proposed approach, we compare our approach with the following approaches:

- MFB (Yu et al., 2017) is the classical approach in general-domain visual question answering through Multi-modal Factorized Bilinear (MFB) pooling to combine multi-modal features.
- SAN (Yang et al., 2016) is another classical architecture in general-domain visual question answering that introduced the stacked attention mechanism.
- BAN (Kim et al., 2018) proposed bilinear interactions among visual and textual information and achieved promising results on the multi-modal tasks.

[2] The time complexity of the ranking process is $O(N^2)$, where N is the sample number.

Table 1
Comparisons of our proposed approach with previous studies on the test sets of three Med-VQA datasets with respect to the accuracy metric, where we detail the close-set and open-set results for VQA-RAD and SLAKE.

Methods	VQA-RAD			SLAKE			MedVQA-2019
	Open	Closed	Overall	Open	Closed	Overall	Overall
MFB (Yu et al., 2017)	14.50	74.30	50.60	72.20	75.00	73.30	–
SAN (Yang et al., 2016)	31.30	69.50	54.30	74.00	79.10	76.00	–
BAN (Kim et al., 2018)	37.40	72.10	58.30	74.60	79.10	76.30	–
MEVF-SAN (Nguyen et al., 2019)	49.20	73.90	64.10	75.30	78.40	76.50	68.90
MEVF-BAN (Nguyen et al., 2019)	49.20	77.20	66.10	77.80	79.80	78.60	77.86
CPRD-BAN (Liu et al., 2021b)	52.50	77.90	67.80	79.50	83.40	81.10	–
M^3AE (Ours)	**67.23**$_{\pm 1.41}$	**83.46**$_{\pm 1.32}$	**77.01**$_{\pm 0.33}$	**80.31**$_{\pm 1.23}$	**87.82**$_{\pm 0.50}$	**83.25**$_{\pm 0.92}$	**79.87**$_{\pm 0.19}$

Table 2
Results on the Med-ITC task (i.e., the MELINDA dataset) to compare with the state-of-the-art methods.

Method	Accuracy
ResNet-101 (He et al., 2016)	63.84
LSTM (Hochreiter and Schmidhuber, 1997)	59.20
RoBERTa (Liu et al., 2019)	75.40
SciBERT (Beltagy et al., 2019)	77.70
NLF (Wu et al., 2021)	76.60
SAN (Yang et al., 2016)	72.30
M^3AE (Ours)	**78.50**$_{\pm 1.04}$

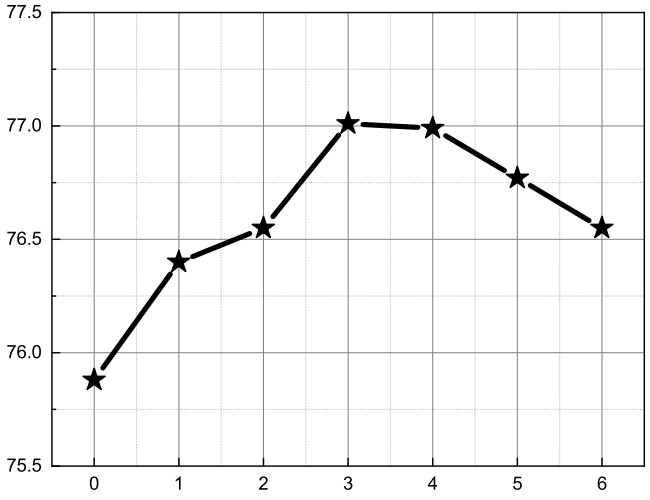

Fig. 2. Accuracy v.s. Different layers of representations for performing masked image modeling (MIM) on the VQA-RAD test set.

- MEVF (Nguyen et al., 2019) targeted medical visual question answering and proposed to use the unsupervised denoising auto-encoder and the supervised meta-learning to overcome the data scarcity problem in the medical domain.
- CPRD (Liu et al., 2021b) is the state-of-the-art approach in medical visual question answering, which proposed to use both contrastive learning and representation distillation to boost the performance of medical visual question answering.
- ViLT (Kim et al., 2021) is a study from general-domain vision-and-language pre-training that proposed to use a simple unified model for representation learning.
- METER (Dou et al., 2021) conducted extensive empirical studies to explore the effects of different components in vision-and-language pre-training and designed a state-of-the-art scheme.

5. Experimental analyses

In this section, we firstly compare the proposed approach with existing studies across several datasets and tasks in Section 5.1. Afterwards, we do in-depth quantitative analyses in Section 5.2 and Section 5.3 and we conduct qualitative analyses in Section 5.4. Finally, we analyze the limitations of the proposed approach in Section 5.5.

5.1. Main results

The main experimental results on all downstream tasks are shown in Table 1, 2, 3, where our proposed approach achieves state-of-the-art results on all datasets. In the Med-VQA task, our approach outperforms the advanced CPRD-BAN approach by 14.7% and 5.5% in terms of accuracy for open-ended and closed-ended questions on the VQA-RAD dataset, respectively. Furthermore, it achieves overall improvements of 2.1% and 2.0% on the SLAKE and VQA-2019 datasets, respectively. For medical image-text classification, our method outperforms previous uni-modal and multi-modal methods under the non-continued pre-training setting, where it outperforms NLF by approximately 1.9%. In the medical image-text retrieval task, the proposed approach outperforms previous studies by a large margin in the zero-shot (ZS) and fine-tuning (FT) settings.

5.2. Ablation study

To evaluate the effectiveness of each component we proposed, we conducted an ablation study on the test set of VQA-RAD without any loss of generation. Our study involved analyzing the contributions of different components of our approach, and the results are presented in Table 4. We observed several key findings from the results. Firstly, we found that using only MIM as the pre-training objective did not lead to any significant improvement. This was evident from the comparison between the first and second rows in Table 4. Secondly, we found that incorporating MLM as one of the pre-training objectives (i.e., the third and fourth rows) resulted in considerably better performance compared to the cases without MLM (i.e., the first and second rows). Finally, our proposed M^3AE, combining both MIM and MLM as objectives, achieved the best performance among all the evaluated approaches. We attribute this success to the fact that M^3AE can implicitly model the critical mappings between medical images and texts by utilizing both MIM and MLM objectives, thereby facilitating the learning of multi-modal representations. Overall, our ablation study demonstrates the effectiveness of all components of M^3AE in addressing the Med-VQA task and highlights the importance of incorporating multiple pre-training objectives for better performance.

5.3. Effects of different MIM layers

We conducted a thorough analysis to evaluate the impact of representations from different layers on MIM. To this end, we pre-trained our model using representations from layers 0 to 6 and analyzed their performance on MIM tasks, as depicted in Fig. 2. Our analysis yielded two key observations that shed light on the effectiveness of different layers in MIM pre-training. First, we found that using representations from layer 0, corresponding to visual features without any textual

Table 3
Results on the image-to-text retrieval and text-to-image retrieval tasks (i.e., the ROCO dataset) to compare with the state-of-the-art methods, where the zero-shot and fine-tuned results are shown.

Methods	Text-to-image Retrieval			Image-to-text Retrieval		
	Recall@1	Recall@5	Recall@10	Recall@1	Recall@5	Recall@10
ViLT (Kim et al., 2021)	9.75	28.95	41.40	11.90	31.90	43.20
METER (Dou et al., 2021)	11.30	27.25	39.60	14.45	33.30	45.10
M^3AE (Ours) (Zero-shot)	19.05	47.75	61.35	19.10	45.60	61.20
M^3AE (Ours) (Fine-tuning)	**22.20**	**52.50**	**66.65**	**22.90**	**51.05**	**65.80**

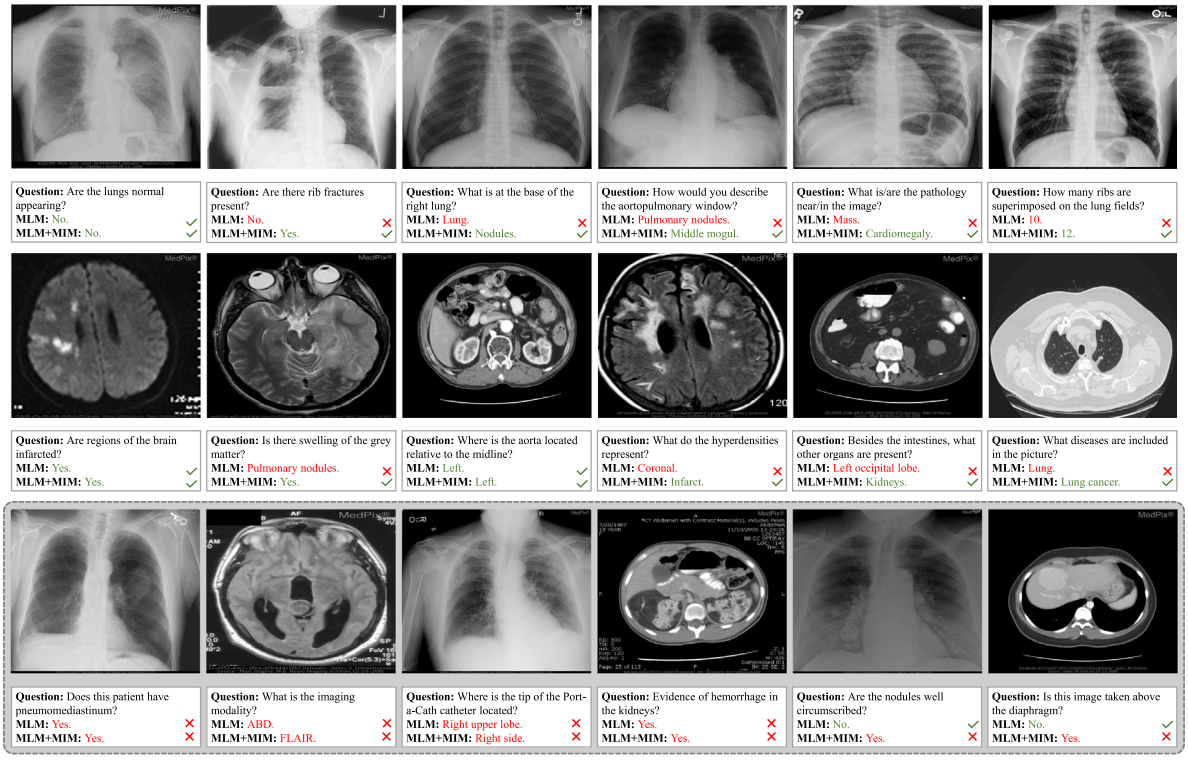

Fig. 3. Illustrations of Med-VQA examples from the models pre-trained with MLM and MLM+MIM on the VQA-RAD test set, where the upper and middle rows refer to chest X-ray and CT cases, respectively. The bottom row refers to some failure cases of MLM+MIM.

Table 4
Ablation study of masked image modeling (MIM) and masked language modeling (MLM) on the VQA-RAD test set.

ID	MIM	MLM	Open	Closed	Overall
1	✗	✗	24.67	80.78	58.48
2	✓	✗	22.41	79.17	56.56
3	✗	✓	67.04	81.99	76.05
4	✓	✓	**67.23**	**83.46**	**77.01**

information, resulted in the worst performance. This observation underscores the vital role of text information in performing MIM effectively. Second, we found that using representations from the intermediate layer (i.e., layer 3) led to the best results. This observation suggests that utilizing lower-level representations for MIM facilitates capturing more hierarchical information in images and texts and generating higher semantic-level representations, which is beneficial for representation learning. Overall, our analysis provides valuable insights into effectively utilizing different layers in MIM pre-training.

5.4. Qualitative analysis

To further validate the efficacy of our approach, we performed a qualitative analysis of several Med-VQA cases on the VQA-RAD test set, as presented in Fig. 3. According to the first two rows, our findings demonstrate that the MLM+MIM model was able to correctly answer all "yes/no" questions presented in the first two columns, whereas the MIM model was only able to answer about half of them correctly. In the middle two columns, where the questions were organ- or region-related, the MIM model was only able to correctly answer one case, while the MLM+MIM model was able to handle all of them. Lastly, in the last two columns, which consisted of more difficult counting or searching-related questions, the MLM+MIM models performed well in correctly answering them, while the MIM model failed to do so. This qualitative analysis illustrates that pre-training with MLM+MIM can enable the model to learn more intricate and nuanced mappings between images and texts, which can enhance the overall performance of the model. In addition, we also show the failure cases in the bottom row of Fig. 3. In these instances, the MLM+MIM setting produced incorrect answers, whereas the MIM setting answered correctly in some cases, indicating the limitations of our model.

Furthermore, we used Grad-CAM (Selvaraju et al., 2017) to generate attention maps that show where our model focuses its attention when answering questions in the VQA-RAD dataset with CT (the top row) and MRI (the bottom row) modalities. The attention maps are shown in Fig. 4, along with corresponding questions for each example. According to the first two rows, the maps illustrate the interpretability of our proposed method. In the first three examples, the questions ask about specific organs or regions, and our model properly focuses on those parts. For example, the model identifies the "costophrenic angles",

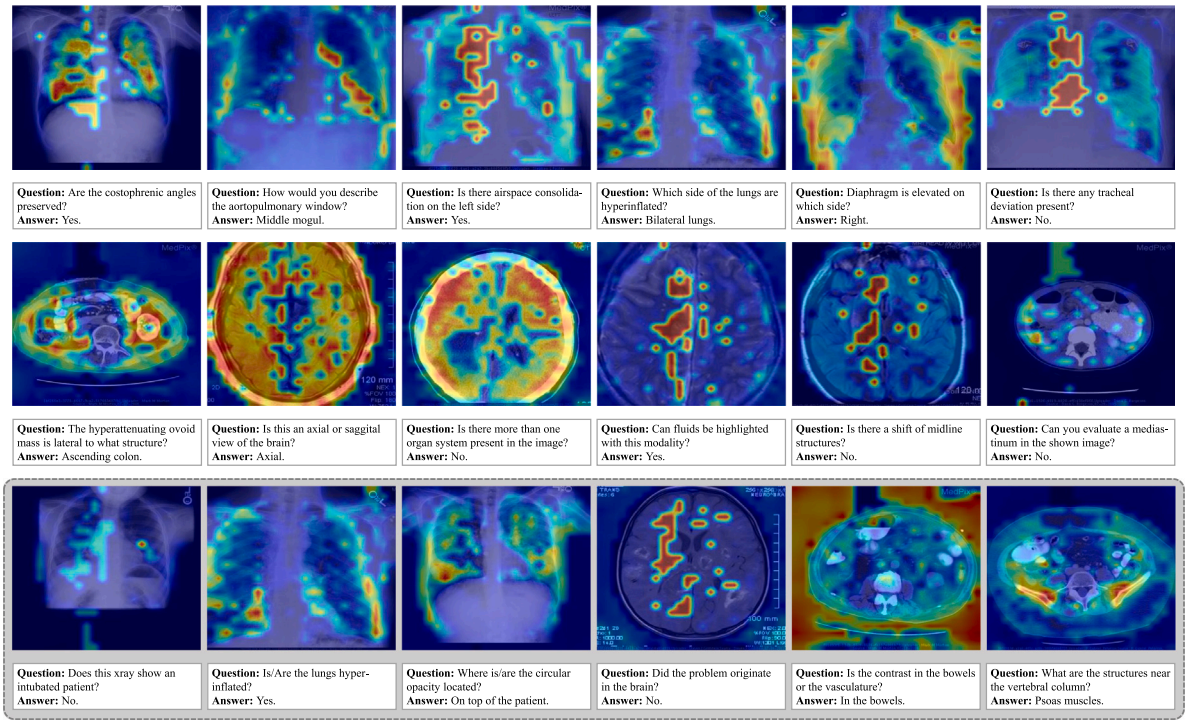

Fig. 4. Illustrations of attention maps of chest X-ray and CT cases generated by our best-performing model on the VQA-RAD test set, paired with the corresponding questions and answers, where the first two rows contain the well-learned attention mappings and the last row shows the failure cases.

Table 5
Limitation investigation, where we show the gap between two types of pre-trained vision-and-language models: (i) Dual-Encoder: better for uni-modal tasks in general; (ii) Fusion-Encoder: better for multi-modal tasks in general. Besides, those models pre-trained without using texts (denoted as "w/o Text Information") are also listed for comparison. 1%, 10%, 100% refer to the different portions of training data.

Type	Model	CheXpert (AUC)			RSNA (AUC)		
		1%	10%	100%	1%	10%	100%
w/o Text Information	Random Init.	56.10	62.60	65.70	58.90	69.40	74.10
	ImageNet Init.	74.40	79.70	81.40	74.90	74.50	76.30
Dual-Encoder	ConVIRT (Zhang et al., 2022a)	85.90	86.80	87.30	77.40	80.10	81.30
	GLoRIA (Huang et al., 2021)	86.60	87.80	88.10	86.10	88.00	88.60
	MGCA (Wang et al., 2022b)	88.80	89.10	89.70	89.10	89.90	90.80
Fusion-Encoder	M^3AE (Ours)	84.00	86.42	88.87	86.66	88.04	89.52

"aortopulmonary window", and "the left side" as relevant areas to answer the questions. In the new three examples, the questions are query-related, causing the model to focus on both sides of the chest X-ray images to capture more global information. Similarly, for the following three examples of MRI images, the model focuses on almost all parts of the images. In the last three cases, the questions are more common and can be answered with common features of the imaging modality. Consequently, the model pays less attention to the images. However, there are still some failure cases where the model focuses on the discrete parts, as illustrated in the bottom row of Fig. 4.

5.5. Limitation analysis

To provide a further reference for future research, we investigate the limitations of the proposed approach. Practically, there exist two typical types, *i.e.*, the fusion-encoder type and the dual-encoder type, depending on whether a heavy fusion module is used. It has been observed from previous studies (Bao et al., 2022; Singh et al., 2022) that the former is superior at multi-modal tasks owing to the sufficient interaction between modalities; the latter is good at uni-modal and cross-modal tasks due to the single-modality encoding ability. Our proposed M^3AE belongs to the fusion-encoder type with a multi-modal fusion module.

To investigate the limitations of M^3AE, we transfer the pre-trained models to medical image classification tasks on two popular benchmark datasets (e.g., CheXpert (Irvin et al., 2019) and RSNA Pneumonia (Shih et al., 2019)). We follow Zhang et al. (2022a), Huang et al. (2021) used different portions of training data to testify the transferring ability, with the results reported in Table 5. There are several observations. First, domain-specific and text-assisted pre-training can boost the performance of downstream tasks, where dual-encoder and fusion-encoder models outperform the random-initialization and ImageNet-initialization models to a large extent. This owes to the domain knowledge from medical images and texts can be modeled and learned during the pre-training process. Second, when the training data is limited (e.g., 1%), it could be observed that dual-encoder models achieve better results due to the smaller gap between pre-training and transferring. The reason behind this is that the visual and textual information is not fused in the pre-training procedure for dual-encoder models. Thus, the vision and language encoders could be used in a separate way when transferring to downstream uni-modal tasks. Third, the performance gap of dual-encoder and fusion-encoder models narrows as the number of training data increases when observing the performance of different models on both datasets. In the meantime, we conducted an additional experiment to validate the performance of MGCA in the VQA-RAD dataset. Specifically, we replace the vision

encoder and language encoder in our framework with the pre-trained MGCA ones and then directly fine-tune the model. The result in the VQA-RAD test set is 73.17. This observation might be explained by the fact that fusion-encoder models (like our model) are equipped with a fusion layer to fuse the information from different modalities in the pre-training stage, which makes it better at the tasks requiring the joint understanding of vision and language.

Therefore, in the future, the unification of dual-encoder and fusion-encoder models could be further studied, where the characteristics of both types of models could be integrated, and the two types of models could facilitate each other.

6. Conclusion

This paper explores the potential of utilizing massive medical image-text data to improve medical vision-and-language understanding. Such data provides valuable context and structural information, which can significantly enhance the ability of models to understand medical images and text. To this end, we propose a simple yet effective approach called multi-modal masked autoencoders (M^3AE), designed to pre-train models on large-scale medical image-text pairs. Our approach incorporates three critical design choices, including masking ratios, representation selection for reconstruction, and decoder designs, which are carefully crafted to optimize the performance of the pre-training process. To comprehensively evaluate the effectiveness of our approach, we introduce a medical vision-and-language understanding benchmark consisting of three tasks: medical visual question answering (Med-VQA), medical image-text classification, and medical image-text retrieval. Experimental results on various datasets demonstrate that our approach achieves state-of-the-art performance across all tasks, surpassing existing approaches by a significant margin. Overall, our work provides a contribution to the field of medical vision-and-language understanding, demonstrating the potential of utilizing multi-modal data for enhancing the performance of models in this domain.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

I have shared the link to my code in the attached file.

Acknowledgments

This work is supported in part by the Shenzhen Science and Technology Program (JCYJ20220818103001002), in part by the Guangdong Provincial Key Laboratory of Big Data Computing, The Chinese University of Hong Kong, Shenzhen, in part by the National Natural Science Foundation of China (NO. 62322608, 61976250), in part by the Open Project Program of the Key Laboratory of Artificial Intelligence for Perception and Understanding, Liaoning Province (AIPU, No. 20230003).

References

Abacha, A.B., Hasan, S.A., Datla, V.V., Liu, J., Demner-Fushman, D., Müller, H., 2019. VQA-med: Overview of the medical visual question answering task at imageclef 2019. In: CLEF (Working Notes), Vol. 2.

Acosta, J.N., Falcone, G.J., Rajpurkar, P., Topol, E.J., 2022. Multimodal biomedical AI. Nat. Med. 28 (9), 1773–1784.

Bao, H., Dong, L., Piao, S., Wei, F., 2021. BEiT: BERT pre-training of image transformers. In: International Conference on Learning Representations.

Bao, H., Wang, W., Dong, L., Liu, Q., Mohammed, O.K., Aggarwal, K., Som, S., Piao, S., Wei, F., 2022. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. Adv. Neural Inf. Process. Syst. 35, 32897–32912.

Beltagy, I., Lo, K., Cohan, A., 2019. SciBERT: A pretrained language model for scientific text. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. EMNLP-IJCNLP, pp. 3615–3620.

Chambon, P., Bluethgen, C., Delbrouck, J.-B., Van der Sluijs, R., Połacin, M., Chaves, J.M.Z., Abraham, T.M., Purohit, S., Langlotz, C.P., Chaudhari, A., 2022. RoentGen: Vision-language foundation model for chest X-ray generation. arXiv preprint arXiv:2211.12737.

Chen, X., Ding, M., Wang, X., Xin, Y., Mo, S., Wang, Y., Han, S., Luo, P., Zeng, G., Wang, J., 2023. Context autoencoder for self-supervised representation learning. Int. J. Comput. Vis. 1–16.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020a. A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning. PMLR, pp. 1597–1607.

Chen, Z., Li, G., Wan, X., 2022. Align, reason and learn: Enhancing medical vision-and-language pre-training with knowledge. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 5152–5161.

Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J., 2020b. Uniter: Universal image-text representation learning. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX. Springer, pp. 104–120.

Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., Sutskever, I., 2020c. Generative pretraining from pixels. In: International Conference on Machine Learning. PMLR, pp. 1691–1703.

Chen, Z., Song, Y., Chang, T.H., Wan, X., 2020d. Generating radiology reports via memory-driven transformer. arXiv preprint arXiv:2010.16056.

Clark, K., Luong, M.T., Le, Q.V., Manning, C.D., 2019. ELECTRA: Pre-training text encoders as discriminators rather than generators. In: International Conference on Learning Representations.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2019. 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, Volume 1 (Long and Short Papers). pp. 4171–4186.

Do, T., Nguyen, B.X., Tjiputra, E., Tran, M., Tran, Q.D., Nguyen, A., 2021. Multiple meta-model quantifying for medical visual question answering. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24. Springer, pp. 64–74.

Dong, X., Bao, J., Zhang, T., Chen, D., Zhang, W., Yuan, L., Chen, D., Wen, F., Yu, N., Guo, B., 2023. Peco: Perceptual codebook for bert pre-training of vision transformers. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 37, No. 1. pp. 552–560.

Dong, L., Yang, N., Wang, W., Wei, F., Liu, X., Wang, Y., Gao, J., Zhou, M., Hon, H.W., 2019. Unified language model pre-training for natural language understanding and generation. Adv. Neural Inf. Process. Syst. 32.

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

Dou, Z.Y., Xu, Y., Gan, Z., Wang, J., Wang, S., Wang, L., Zhu, C., Peng, N., Liu, Z., Zeng, M., 2021. An empirical study of training end-to-end vision-and-language transformers. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 18145–18155.

Fang, Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., Huang, T., Wang, X., Cao, Y., 2023. Eva: Exploring the limits of masked visual representation learning at scale. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19358–19369.

Gao, P., Ma, T., Li, H., Lin, Z., Dai, J., Qiao, Y., 2022. MCMAE: Masked convolution meets masked autoencoders. Adv. Neural Inf. Process. Syst. 35, 35632–35644.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R., 2021. Masked autoencoders are scalable vision learners. arXiv preprint arXiv:2111.06377.

He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020. Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9729–9738.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.

Huang, S.C., Shen, L., Lungren, M.P., Yeung, S., 2021. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3942–3951.

Huang, L., You, S., Zheng, M., Wang, F., Qian, C., Yamasaki, T., 2022. Green hierarchical vision transformer for masked image modeling. Adv. Neural Inf. Process. Syst. 35, 19997–20010.

Huang, Z., Zeng, Z., Liu, B., Fu, D., Fu, J., 2020. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. arXiv preprint arXiv:2004.00849.

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al., 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, No. 01. pp. 590–597.

Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T., 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In: International Conference on Machine Learning. PMLR, pp. 4904–4916.

Jing, B., Xie, P., Xing, E., 2018. On the automatic generation of medical imaging reports. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2577–2586.

Khare, Y., Bagal, V., Mathew, M., Devi, A., Priyakumar, U.D., Jawahar, C., 2021. Mmbert: Multimodal bert pretraining for improved medical vqa. In: 2021 IEEE 18th International Symposium on Biomedical Imaging. ISBI, IEEE, pp. 1033–1036.

Kim, J.H., Jun, J., Zhang, B.T., 2018. Bilinear attention networks. Adv. Neural Inf. Process. Syst. 31.

Kim, W., Son, B., Kim, I., 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In: International Conference on Machine Learning. PMLR, pp. 5583–5594.

Lau, J.J., Gayen, S., Abacha, A.B., Demner-Fushman, D., 2018. A dataset of clinically generated visual questions and answers about radiology images. Sci. Data 5.

Li, Y., Wang, H., Luo, Y., 2020a. A comparison of pre-trained vision-and-language models for multimodal representation learning across medical images and reports. In: 2020 IEEE International Conference on Bioinformatics and Biomedicine. BIBM, IEEE, pp. 1999–2004.

Li, L.H., Yatskar, M., Yin, D., Hsieh, C.-J., Chang, K.-W., 2019. Visualbert: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557.

Li, X., Yin, X., Li, C., Hu, X., Zhang, P., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., Choi, Y., Gao, J., 2020b. Oscar: Object-semantics aligned pre-training for vision-language tasks. In: European Conference on Computer Vision. pp. 121–137.

Li, G., Zheng, H., Liu, D., Wang, C., Su, B., Zheng, C., 2022. Semmae: Semantic-guided masking for learning masked autoencoders. Adv. Neural Inf. Process. Syst. 35, 14290–14302.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Liu, F., Wu, X., Ge, S., Fan, W., Zou, Y., 2021a. Exploring and distilling posterior and prior knowledge for radiology report generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13753–13762.

Liu, B., Zhan, L.M., Wu, X.M., 2021b. Contrastive pre-training and representation distillation for medical visual question answering based on radiology images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 210–220.

Liu, B., Zhan, L.M., Xu, L., Ma, L., Yang, Y., Wu, X.M., 2021c. SLAKE: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In: 2021 IEEE 18th International Symposium on Biomedical Imaging. ISBI, IEEE, pp. 1650–1654.

Loshchilov, I., Hutter, F., 2018. Decoupled weight decay regularization. In: International Conference on Learning Representations.

Lu, J., Batra, D., Parikh, D., Lee, S., 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Adv. Neural Inf. Process. Syst. 32.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J., 2013. Distributed representations of words and phrases and their compositionality. Adv. Neural Inf. Process. Syst. 26.

Monajatipoor, M., Rouhsedaghat, M., Li, L.H., Jay Kuo, C.-C., Chien, A., Chang, K.-W., 2022. Berthop: An effective vision-and-language model for chest x-ray disease diagnosis. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part V. Springer, pp. 725–734.

Moon, J.H., Lee, H., Shin, W., Choi, E., 2021. Multi-modal understanding and generation for medical images and text via vision-language pre-training. arXiv preprint arXiv:2105.11333.

Moor, M., Banerjee, O., Abad, Z.S.H., Krumholz, H.M., Leskovec, J., Topol, E.J., Rajpurkar, P., 2023. Foundation models for generalist medical artificial intelligence. Nature 616 (7956), 259–265.

Mu, N., Kirillov, A., Wagner, D., Xie, S., 2022. Slip: Self-supervision meets language-image pre-training. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI. Springer, pp. 529–544.

Müller, P., Kaissis, G., Zou, C., Rückert, D., 2021. Joint learning of localized representations from medical images and reports. arXiv preprint, arXiv:2112.02889.

Najdenkoska, I., Zhen, X., Worring, M., Shao, L., 2022. Uncertainty-aware report generation for chest X-rays by variational topic inference. Med. Image Anal. 82, 102603.

Nguyen, B.D., Do, T.T., Nguyen, B.X., Do, T., Tjiputra, E., Tran, Q.D., 2019. Overcoming data limitation in medical visual question answering. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 522–530.

Pelka, O., Koitka, S., Rückert, J., Nensa, F., Friedrich, C.M., 2018. Radiology objects in context (ROCO): a multimodal image dataset. In: Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis. Springer, pp. 180–189.

Peng, Z., Dong, L., Bao, H., Ye, Q., Wei, F., 2022. Beit v2: Masked image modeling with vector-quantized visual tokenizers. arXiv preprint arXiv:2208.06366.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I., 2021. Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning.

Seenivasan, L., Islam, M., Krishna, A.K., Ren, H., 2022. Surgical-VQA: Visual question answering in surgical scenes using transformer. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VII. Springer, pp. 33–43.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 618–626.

Shih, G., Wu, C.C., Halabi, S.S., Kohli, M.D., Prevedello, L.M., Cook, T.S., Sharma, A., Amorosa, J., Arteaga, V.A., Galperin-Aizenberg, M., Gill, R.R., Godoy, M.C.B., Hobbs, S., Jeudy, J., Laroia, A., Shah, P.N., Vummidi, D.R., Yaddanapudi, K., Stein, A., 2019. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. Radiology. Artif. Intell. 1 1, e180041.

Shin, H.C., Roberts, K., Lu, L., Demner-Fushman, D., Yao, J., Summers, R.M., 2016. Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2497–2506.

Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., Kiela, D., 2022. Flava: A foundational language and vision alignment model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15638–15650.

Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., Dai, J., 2019. VL-BERT: Pre-training of generic visual-linguistic representations. In: International Conference on Learning Representations.

Subramanian, S., Wang, L.L., Bogin, B., Mehta, S., van Zuylen, M., Parasa, S., Singh, S., Gardner, M., Hajishirzi, H., 2020. MedICaT: A dataset of medical images, captions, and textual references. In: Findings of the Association for Computational Linguistics: EMNLP 2020. pp. 2112–2120.

Tan, H., Bansal, M., 2019. LXMERT: Learning cross-modality encoder representations from transformers. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. EMNLP-IJCNLP, pp. 5100–5111.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.

Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O.K., Singhal, S., Som, S., et al., 2022a. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. arXiv preprint arXiv:2208.10442.

Wang, H., Tang, Y., Wang, Y., Guo, J., Deng, Z.-H., Han, K., 2023. Masked image modeling with local multi-scale reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2122–2131.

Wang, Z., Zhou, L., Wang, L., Li, X., 2021. A self-boosting framework for automated radiographic report generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2433–2442.

Wang, F., Zhou, Y., Wang, S., Vardhanabhuti, V., Yu, L., 2022b. Multi-granularity cross-modal alignment for generalized medical visual representation learning. In: Oh, A.H., Agarwal, A., Belgrave, D., Cho, K. (Eds.), Advances in Neural Information Processing Systems.

Wei, C., Fan, H., Xie, S., Wu, C.Y., Yuille, A., Feichtenhofer, C., 2022a. Masked feature prediction for self-supervised visual pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14668–14678.

Wei, Y., Hu, H., Xie, Z., Zhang, Z., Cao, Y., Bao, J., Chen, D., Guo, B., 2022b. Contrastive learning rivals masked image modeling in fine-tuning via feature distillation. arXiv preprint arXiv:2205.14141.

Wei, L., Xie, L., Zhou, W., Li, H., Tian, Q., 2022c. Mvp: Multimodality-guided visual pre-training. In: European Conference on Computer Vision. Springer, pp. 337–353.

Wettig, A., Gao, T., Zhong, Z., Chen, D., 2023. Should you mask 15% in masked language modeling? In: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. pp. 2977–2992.

Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J.R., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G.S., Hughes, M., Dean, J., 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv, arXiv:1609.08144.

Wu, T.L., Singh, S., Paul, S., Burns, G., Peng, N., 2021. MELINDA: A multimodal dataset for biomedical experiment method classification. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35. pp. 14076–14084.

Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H., 2022. Simmim: A simple framework for masked image modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9653–9663.

Yan, A., He, Z., Lu, X., Du, J., Chang, E., Gentili, A., McAuley, J., Hsu, C.n., 2021. Weakly supervised contrastive learning for chest X-Ray report generation. In: Findings of the Association for Computational Linguistics: EMNLP 2021. pp. 4009–4015.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V., 2019. Xlnet: Generalized autoregressive pretraining for language understanding. Adv. Neural Inf. Process. Syst. 32.

Yang, Z., He, X., Gao, J., Deng, L., Smola, A., 2016. Stacked attention networks for image question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 21–29.

Yao, L., Huang, R., Hou, L., Lu, G., Niu, M., Xu, H., Liang, X., Li, Z., Jiang, X., Xu, C., 2022. FILIP: Fine-grained interactive language-image pre-training. In: International Conference on Learning Representations.

Yu, F., Tang, J., Yin, W., Sun, Y., Tian, H., Wu, H., Wang, H., 2021. ERNIE-ViL: Knowledge enhanced vision-language representations through scene graphs. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35. pp. 3208–3216.

Yu, Z., Yu, J., Fan, J., Tao, D., 2017. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1821–1830.

Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P., 2022a. Contrastive learning of medical visual representations from paired images and text. In: Machine Learning for Healthcare Conference. PMLR, pp. 2–25.

Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., Gao, J., 2021. Vinvl: Revisiting visual representations in vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5579–5588.

Zhang, C., Zhang, C., Song, J., Yi, J.S.K., Zhang, K., Kweon, I.S., 2022b. A survey on masked autoencoder for self-supervised learning in vision and beyond. arXiv preprint arXiv:2208.00173.

Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T., 2021. Image BERT pre-training with online tokenizer. In: International Conference on Learning Representations.