Align, Reason and Learn: Enhancing Medical Vision-and-Language Pre-training with Knowledge

Zhihong Chen
Shenzhen Research Institute of Big Data,
The Chinese University of Hong Kong,
Shenzhen, China
zhihongchen@link.cuhk.edu.cn

Guanbin Li*
Sun Yat-sen University,
Guangzhou, China
liguanbin@mail.sysu.edu.cn

Xiang Wan*
Shenzhen Research Institute of Big Data,
The Chinese University of Hong Kong,
Shenzhen, China
Pazhou Lab,
Guangzhou, China
wanxiang@sribd.cn

ABSTRACT

Medical vision-and-language pre-training (Med-VLP) has received considerable attention owing to its applicability to extracting generic vision-and-language representations from medical images and texts. Most existing methods mainly contain three elements: uni-modal encoders (i.e., a vision encoder and a language encoder), a multi-modal fusion module, and pretext tasks, with few studies considering the importance of medical domain expert knowledge and explicitly exploiting such knowledge to facilitate Med-VLP. Although there exist knowledge-enhanced vision-and-language pre-training (VLP) methods in the general domain, most require off-the-shelf toolkits (e.g., object detectors and scene graph parsers), which are unavailable in the medical domain. In this paper, we propose a systematic and effective approach to enhance Med-VLP by structured medical knowledge from three perspectives. First, considering knowledge can be regarded as the intermediate medium between vision and language, we align the representations of the vision encoder and the language encoder through knowledge. Second, we inject knowledge into the multi-modal fusion model to enable the model to perform reasoning using knowledge as the supplementation of the input image and text. Third, we guide the model to put emphasis on the most critical information in images and texts by designing knowledge-induced pretext tasks. To perform a comprehensive evaluation and facilitate further research, we construct a medical vision-and-language benchmark including three tasks. Experimental results illustrate the effectiveness of our approach, where state-of-the-art performance is achieved on all downstream tasks. Further analyses explore the effects of different components of our approach and various settings of pre-training.[1]

CCS CONCEPTS

• **Computing methodologies → Artificial intelligence**; **Machine learning**; **Multi-task learning**.

---

*Corresponding authors.
[1]The source code is available at https://github.com/zhjohnchan/ARL.

KEYWORDS

1 INTRODUCTION

Medical data streams from various sources, among which vision and language are two critical ones. It includes image data (e.g., radiography, magnetic resonance imaging, and computed tomography) and text data (e.g., radiology reports and medical texts). Medical vision-and-language pre-training (Med-VLP) aims to jointly process data from these two modalities to learn generalizable multi-modal representations from large-scale medical image-text data. It enables a vision-and-language model to address a wide range of medical vision-and-language tasks (e.g., medical visual question answering (Med-VQA), medical image-text classification (Med-ITC), and medical image-text retrieval (Med-ITR)), which can be crucial for alleviating the data scarcity problem in the medical field.

In the past few years, vision-and-language pre-training (VLP) has drawn sustaining attention [8, 9, 24, 28, 35, 43, 47] and achieved state-of-the-art performance on many vision-and-language tasks in the general domain. In general, a VLP system consists of three elements: (i) uni-modal encoders (i.e., a vision encoder and a language encoder) that encode images and texts into image and text features, respectively; (ii) a multi-modal fusion module that performs the fusion of the encoded image and text features; (iii) pretext tasks (e.g., masked image modeling (MIM), masked language modeling (MLM), and image-text matching (ITM)) that assist the learning of VLP models. More recently, some studies [15, 22, 29, 37] applied VLP to the medical domain and significantly improved the performance for medical vision-and-language tasks (especially for Med-VQA). These methods are superior in capturing the mappings between images and texts and thus enable the pre-trained models to understand the complicated cross-modal information. For example, [15, 22] proposed to perform the pre-training on medical image-text pairs to capture medical knowledge, and the evaluation on Med-VQA has demonstrated the validity of their proposed methods.

Although these methods have motivated the learning of image-text correspondences through well-designed model architectures

and pretext tasks, most of them disregard the complementary information (i.e., knowledge) shared by different modalities and still lack the explicit knowledge modeling for Med-VLP. Even in the general domain, there are only a few VLP studies [7, 9, 28, 55] on incorporating external knowledge into the pre-training process. For instance, ERNIE-ViL [55] constructed a scene graph from the input text to build the semantic connections between vision and language and emphasized the importance of keywords (e.g., objects, attributes, and relationships between objects) through the designs of pretext tasks. ROSITA [9] used a unified scene graph shared by the input image and text to enhance the semantic alignments between vision and language. Similarly, KB-VLP [7] used object tags detected from images and knowledge graph embeddings extracted from texts to enhance the learning of knowledge-aware representations. However, the aforementioned studies require off-the-shelf toolkits (e.g., object detectors and scene graph parsers), which are generally unavailable in the medical domain. Furthermore, they might be limited in scalability as their performance depends heavily on the reliability of the object detectors or scene graph parsers. Therefore, it is expected to have a better solution to exploit external knowledge more appropriately and systematically and further improve the generalization ability of Med-VLP methods.

In this paper, we propose a systematic approach to Med-VLP enhanced by structured expert domain knowledge from the Unified Medical Language System [4] (UMLS), a large medical knowledge base containing many biomedical terminologies with the associated information, such as synonyms and categorical groupings. To ensure the effectiveness and efficiency of our approach, structured knowledge is injected into the Med-VLP system from three perspectives: (i) Aligning Through Knowledge: It uses knowledge as the intermediate medium between vision and language to align the image and text features encoded by the uni-modal encoders; (ii) Reasoning Using Knowledge: It develops a knowledge-enhanced multi-modal fusion module to integrate knowledge into the interaction process of the image and text features; (iii) Learning From Knowledge: It constructs knowledge-induced pretext tasks to assist the model in capturing underlying critical medical information of the images and texts to promote the medical vision-and-language understanding. As a result, the proposed method is able to learn cross-modal domain-specific knowledge from large-scale medical image-text datasets and medical knowledge bases to promote the learning of semantically aligned and knowledge-aware image and text representations. We perform the pre-training on three large-scale medical image-text datasets, i.e., ROCO [40], MedICaT [44], and MIMIC-CXR [21]. To verify the effectiveness of our approach and facilitate further research, we construct a medical vision-and-language understanding benchmark including three tasks (i.e., Med-VQA, Med-ITC, and Med-ITR). Experimental results demonstrate the effectiveness of our approach, where state-of-the-art performance is achieved on all datasets.

2 RELATED WORK

Vision-and-Language Pre-training (VLP) Motivated by the success of the self-supervised pre-training recipe of BERT in NLP, there has been an increasing interest in developing VLP methods to address a wide range of vision-and-language tasks. In general, VLP

methods can be categorized with respect to three perspectives. For the designs of the uni-modal encoders, different methods adopt different image features (e.g., region features [27, 35], patch embeddings [24, 26, 51], and grid features [20]) and distinct text features (e.g., statistic embeddings [24] and dynamic embeddings [13]). For multi-modal fusion modules, existing methods can be classified into two categories (i.e., single-stream and dual-stream). In specific, for the single-stream fusion, the models [8, 27, 28, 43] use a single Transformer for early and unconstrained fusion between modalities; for the dual-stream fusion, the models [35, 47, 55] adopt the co-attention mechanism to interact different modalities. For pretext tasks, inspired by uni-modal pre-training schemes such as MLM [10, 33] and causal language modeling [6], existing studies explore a variety of pre-training tasks, including MLM [27, 35, 47], MIM [8, 35], ITM [27, 58], image-text contrastive [26] and prefix language modeling [51]. This paper adopts a purely Transformer-based backbone architecture using the dual-stream fusion with ViT-based grid features and BERT-based dynamic text features and three common pretext tasks (i.e., MLM, MIM, and ITM).

Medical Vision-and-Language Pre-training (Med-VLP) Being one of the applications and extensions of VLP to the medical domain, Med-VLP aims to understand the content and relations between medical images and their corresponding texts. It can be traced back to [29], which explored the performance of four vision-and-language models pre-trained in the general domain on a disease classification task. Then MMBERT [22], PubMedCLIP [14], and MedViLL [37] performed pre-training on medical image-text data before fine-tuning on the downstream tasks. Compared with these studies, we design a more appropriate and systematic scheme for Med-VLP from four aspects (i.e., pre-training datasets, model designs, pre-training tasks, and evaluation benchmarks).

Knowledge-Enhanced Pre-training For uni-modal pre-training in CV and NLP, many works have investigated how to incorporate knowledge into the pre-trained models. According to the knowledge injection schemes, existing studies can be classified into four categories: embeddings combination [41, 59], data structure compatibility [16, 32, 45], knowledge supervision [46, 49], and neural-symbolic methods [2]. For VLP, knowledge can be acquired from both the image and text modalities, and there are several works [9, 28, 55] studying to integrate knowledge into their methods. ERNIE-ViL [55] built detailed semantic alignments between vision and language based on the scene graph parsed from the text. ROSITA [9] proposed integrating extra cross-modal knowledge mappings to enhance the learning of semantic alignments between vision and language. Different from them, we revisit existing knowledge-enhanced methods and propose to inject knowledge from three VLP-specific perspectives without requiring object detectors or scene graph parsers, which are unavailable in the medical domain.

3 THE PROPOSED APPROACH

We follow the standard pre-train-and-fine-tune paradigm for medical vision-and-language understanding. In the pre-training stage, the framework develops a variety of pretext tasks to train the Med-VLP model using medical image-text pairs. In the fine-tuning stage, the pre-trained Med-VLP model is transferred to various medical
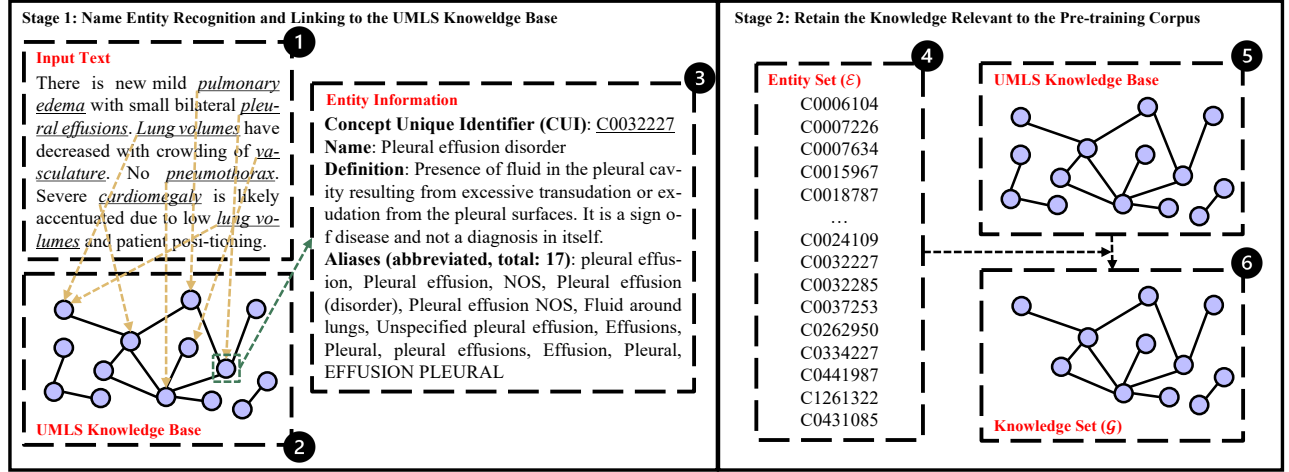
Figure 1: The flowchart of knowledge extraction from UMLS, a medical knowledge base. It contains two main stages, where the first stage is to link entities of each input text to the knowledge base and the second stage is to retain the knowledge relevant to the pre-training corpus to form our knowledge set. Numbers are marked for ease of reading.

vision-and-language downstream tasks. An overview of the proposed approach is demonstrated in Figure 2, and the details of the general Med-VLP framework, the knowledge extraction process (shown in Figure 1), and the injection of knowledge into the general Med-VLP framework are introduced in the following subsections.

3.1 The General Med-VLP Framework

The general Med-VLP framework can be partitioned into three major components, i.e., the uni-modal encoders, the multi-modal fusion module, and the pretext tasks. The overall description of the three components is detailed below.

Uni-modal Encoders In the Med-VLP framework, there is a vision encoder and a language encoder, which encode the input image and text into image and text features, respectively.

For the vision encoder, we study the use of vision Transformer [12] (ViT). In ViT, an input image $I \in \mathbb{R}^{H \times W \times C}$ is first segmented into patches $\{x_1^v, x_2^v, \ldots, x_{N_v}^v\}$, where $H \times W$ is the image resolution, C is the number of channels, N_v is the number of patches, $x_i^v \in \mathbb{R}^{P^2 \times C}$ and $P \times P$ is the patch resolution. Then the patches $\{x_1^v, x_2^v, \ldots, x_{N_v}^v\}$ are flattened and linearly projected into patch embeddings through a linear transformation $E^v \in \mathbb{R}^{P^2 C \times D}$ and a special learnable token embedding $x_I^v \in \mathbb{R}^D$ is prepended for the aggregation of visual information. Therefore, the input image representations are obtained via summing up the patch embeddings and learnable 1D position embeddings $E_{pos}^v \in \mathbb{R}^{(N_v+1) \times D}$:

$$X^v = [x_I^v; x_1^v E^v; x_2^v E^v; \ldots; x_{N_v}^v E^v] + E_{pos}^v. \tag{1}$$

Then X^v is fed into a Transformer model with L_v Transformer layers. Finally, we obtain the contextualized image representations $H^v = [h_I^v; h_1^v; h_2^v; \ldots; h_{N_v}^v]$.

For the language encoder, we follow BERT [10] to tokenize the input text to subword tokens $\{x_1^l, x_2^l, \ldots, x_{N_l}^l\}$ by WordPiece [53] and then represent subword tokens as $\{x_1^l, x_2^l, \ldots, x_{N_l}^l\}$, where

$x_i^l \in \mathbb{R}^V$ are the one-hot form of x_i^l, V is the vocabulary size, and N_l is the number of tokens. Subsequently, 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 $x_T^l \in \mathbb{R}^D$ and a special boundary token embedding $x_{SEP}^l \in \mathbb{R}^D$ are added to the text sequence. Therefore, the input text representations are computed via summing up the token embeddings and text position embeddings $E_{pos}^l \in \mathbb{R}^{(N_l+2) \times D}$:

$$X^l = [x_T^l; x_1^l E^l; \ldots; x_{N_l}^l E^l; x_{SEP}^l] + E_{pos}^l. \tag{2}$$

Then X^l is fed into a Transformer model with L_l Transformer layers. Finally, we obtain the contextualized text representations $H^l = [h_T^l; h_1^l; h_2^l; \ldots; h_{N_l}^l; h_{SEP}^l]$.

Multi-model 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 for vision and language, respectively, each of which is a stack of L_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 feed-forward sub-layer. The attention mechanism is applied in the self-attention and cross-attention sub-layers and is defined as

$$\text{ATTN}(Q, K, V) = \text{softmax}\left(QK^\top / \sqrt{D_k}\right) \cdot V, \tag{3}$$

where Q, K, and V are the query, key, value matrices linearly transformed from the corresponding input sequences, respectively, and D_k is the dimension of K. In the self-attention sub-layer, the representations interact within modalities:

$$H^{vs} = \text{ATTN}(H^v, H^v, H^v), \tag{4}$$

$$H^{ls} = \text{ATTN}(H^l, H^l, H^l), \tag{5}$$

where H^{vs} and H^{ls} are the self-attention outputs for vision and language, respectively. Then residual connections followed by layer
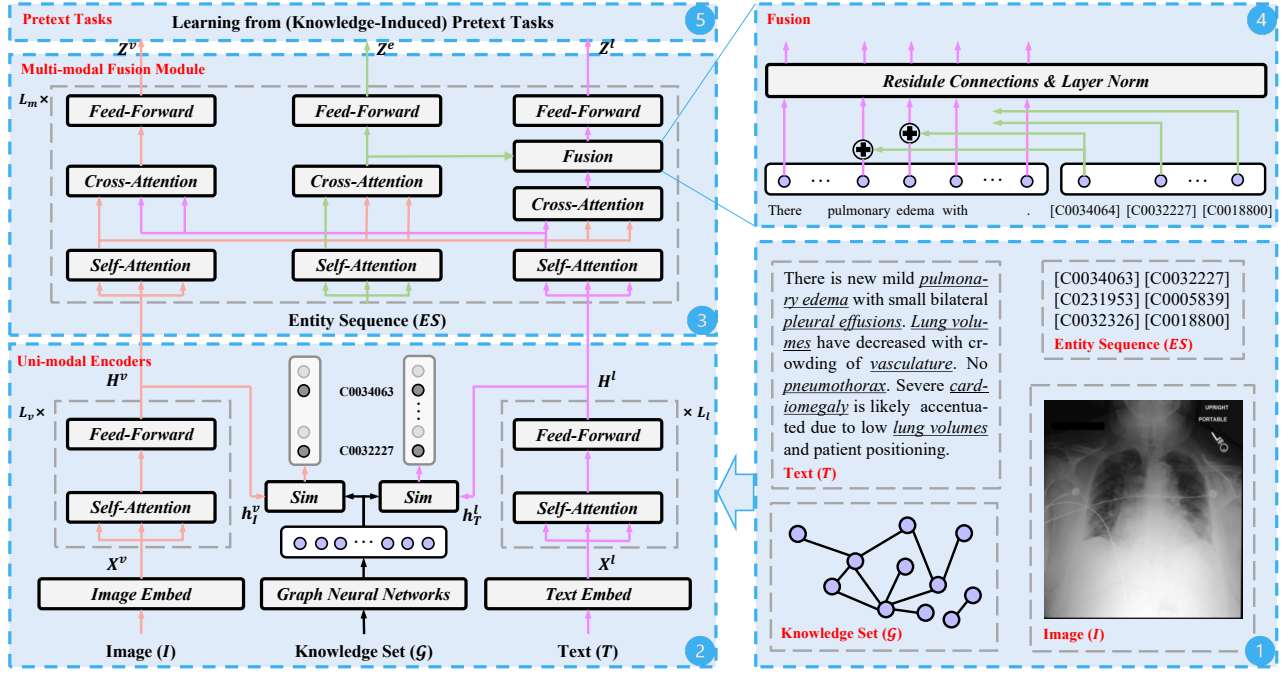
Figure 2: The overall architecture of our proposed approach, where the inputs, uni-modal encoders (with the "aligning through knowledge" process), multi-modal fusion module (with the "reasoning using knowledge" process), pretext tasks (with the "learning from knowledge" process) are shown in blue dash boxes. Numbers are marked for ease of reading.

normalization are employed to H^{vs} and H^{ls} and we denote the results as H^{vs} and H^{ls}, respectively, for simplicity. In the cross-attention sub-layer, the representations interact across modalities to integrate cross-modal information into their representations:

$$H^{vc} = \text{ATTN}(H^{vs}, H^{ls}, H^{ls}), \qquad (6)$$

$$H^{lc} = \text{ATTN}(H^{ls}, H^{vs}, H^{vs}), \qquad (7)$$

where H^{vc} and H^{lc} are the cross-attention outputs for vision and language, respectively. Similarly, residual connections followed by layer normalization are employed to H^{vc} and H^{lc}, and we denote the results as H^{vc} and H^{lc}, respectively, for simplicity. Finally, H^{vc} and H^{lc} are input to the feed-forward sub-layer (i.e., a multi-layer perceptron (MLP)) to obtain the multi-modal representations $Z^v = [z^v_I; z^v_1; z^v_2; \ldots; z^v_{N_v}]$ for vision and $Z^l = [z^l_T; z^l_1; z^l_2; \ldots; z^l_{N_l}; z^l_{SEP}]$ for language.

Pretext Tasks Given the aforementioned structure (denoted as $\mathcal{M}_\theta$) with its parameters θ, the Med-VLP framework develops various pretext tasks (e.g., masked language modeling (MLM), masked image modeling (MIM), and image-text matching (ITM)) to guide the learning of θ. Assuming there are S pretext tasks, the learning process can be formalized as

$$\theta^*, \theta^*_1, \ldots, \theta^*_S = \underset{\theta, \theta_1, \ldots, \theta_S}{\arg\min} \sum_{i=1}^{S} L_i(Y_i, \mathcal{D}_{\theta_i}(\mathcal{M}_\theta(I, T))), \qquad (8)$$

where L_i are the loss functions of pretext tasks, Y_i are the corresponding ground-truth labels, and $\mathcal{D}_{\theta_i}$ are the prediction heads with their parameters θ_i.

3.2 Knowledge Extraction

Although knowledge graphs (KGs) have shown their effectiveness in many natural language processing (NLP) tasks [32, 41, 59] and computer vision (CV) tasks [36, 50], the existing Med-VLP methods rarely consider incorporating KGs to provide rich structured knowledge for better vision-and-language understanding.

Therefore, we propose to enhance Med-VLP by leveraging external domain expert knowledge from UMLS. In doing so, we extract knowledge through two stages, as illustrated in Figure 1. The first stage is to apply a named entity recognition, and linking tool ScispaCy [38] to pre-process the texts in the pre-training corpus to link entities in the texts to the UMLS knowledge base for entity disambiguation. Therefore, for each image-text pair, there is an entity sequence $ES = \{x^e_1, x^e_2, \ldots, x^e_{N_{es}}\}$ aligning to the token sequence $T = \{x^l_1, x^l_2, \ldots, x^l_{N_l}\}$, where x^e_i are the extracted entities and N_{es} is the length of the entity sequence. Following [11], to record the position of the extracted entities, we adopt an entity matching matrix $P \in \mathbb{R}^{N_l \times N_{es}}$, where each element is represented by

$$P_{ij} = \begin{cases} 1 & x^l_i \in x^e_j \\ 0 & x^l_i \notin x^e_j \end{cases}, \qquad (9)$$

where P is employed to assist the interaction between the text and the entity sequence in the knowledge injection process (as described in the next subsection). After pre-processing all the texts, we can obtain an entity set $\mathcal{E} = \{e_i\}_{i=1}^{N_e}$ containing all the N_e entities related to the pre-training corpus. The second stage is to extract relevant knowledge graph triples from the UMLS knowledge base once both

the head and tail entities of the triple are in the entity set $\mathcal{E}$. We denote the extracted knowledge graph (i.e., a sub-graph of the UMLS knowledge base) as the knowledge set $\mathcal{G} = \{k_i = (h_i, r_i, t_i)\}_{i=1}^{N_g}$, where N_g is the number of knowledge graph triples, k_i are the knowledge graph triples, and h_i, r_i and t_i represent the head entity, relation, and tail entity, respectively.

3.3 Knowledge Injection

To integrate knowledge into the general Med-VLP framework, first, we perform knowledge representation following two steps: (i) We apply knowledge representation learning algorithm (e.g., TransE [5]) to the knowledge graph $\mathcal{G} = \{k_i = (h_i, r_i, t_i)\}_{i=1}^{N_g}$ to obtain the entity embeddings $\{e_i\}_{i=1}^{N_e}$, where $e_i \in \mathbb{R}^{D_e}$ and D_e is the dimension of the entity embeddings; (ii) We adopt Graph Neural Networks (e.g., Graph Attention Networks [48]) to take account of the whole structure of the graph to aggregate local information in the graph neighborhood for each node, and obtain the entity representations (denoted as $\{e_i\}_{i=1}^{N_e}$ for simplicity, where $e_i \in \mathbb{R}^{D_e}$).

Afterwards, given the input image I and text $T = \{x_1^l, x_2^l, \ldots, x_{N_l}^l\}$ with its corresponding entity sequence $ES = \{x_1^e, x_2^e, \ldots, x_{N_{es}}^e\}$, we develop three essential and systematic designs to inject knowledge from the following perspectives:

(i) Aligning Through Knowledge Knowledge can be regarded as the intermediate medium between vision and language, where knowledge can be used as an explanation of the meaning behind both images and texts. In most cases, entities in knowledge graph triples can be treated as anchor points that appear in the image and are mentioned in the accompanying text. Motivated by this fact, we propose to align the image representations and the text representations from uni-modal encoders through knowledge. Similar to [26], it serves two purposes: It improves the unimodal encoders to better understand the semantic meaning of images and texts; It eases the learning of semantic alignments between images and texts.

Formally, given the aggregated image representation h_I^v and text representation h_T^l, we calculate the image-knowledge and text-knowledge similarity followed by a sigmoid function:

$$p_i^v = \text{sigmoid}(e_i^\top W_{vk} h_I^v), i = 1, ..., N_e, \tag{10}$$

$$p_i^l = \text{sigmoid}(e_i^\top W_{lk} h_T^l), i = 1, ..., N_e, \tag{11}$$

where $W_{vk} \in \mathbb{R}^{D_e \times D}$ and $W_{lk} \in \mathbb{R}^{D_e \times D}$ are trainable weights for the linear transformation. Therefore, the alignments for image-knowledge and text-knowledge are learned explicitly through minimizing the following functions:

$$L_{vk} = -\sum_{i=1}^{N_e} (y_i \log p_i^v + (1 - y_i) \log (1 - p_i^v)), \tag{12}$$

$$L_{lk} = -\sum_{i=1}^{N_e} (y_i \log p_i^l + (1 - y_i) \log (1 - p_i^l)), \tag{13}$$

where y_i can be defined as

$$y_i = \begin{cases} 1 & e_i \in ES \\ 0 & e_i \notin ES \end{cases}. \tag{14}$$

Therefore, knowledge is employed as an intermediate medium to enhance and smooth image-text mappings by doing so.

(ii) Reasoning Using Knowledge As the supplementation of the input image and text, knowledge can also be utilized to assist the reasoning of the Med-VLP model. In doing so, we enhance the multi-modal fusion module by knowledge.

Formally, given the entity sequence $ES = \{x_1^e, x_2^e, ..., x_{N_{es}}^e\}$, first, we extract its entity representations $H^e = [h_1^e; h_2^e; \ldots; h_{N_{es}}^e]$. Second, we apply self-attention to the entity representations to encode the contextualized information:

$$H^{es} = \text{ATTN}(H^e, H^e, H^e), \tag{15}$$

where H^{es} is the self-attention outputs of entity representations. Residual connections followed by layer normalization are employed to H^{es}, and we denote the outputs as H^{es} for simplicity. Third, to interact the entities with the image, since there is no available toolkits to structuralize the input image to construct mappings between image patches and entities, we directly perform cross-attention on H^{es} and H^{vs}:

$$H^{ec} = \text{ATTN}(H^{es}, H^{vs}, H^{vs}), \tag{16}$$

where H^{ec} is the cross-attention outputs of entity representations. Residual connections followed by layer normalization are employed to H^{ec}, and we denote the outputs as H^{ec} for simplicity. Fourth, since the mappings between the entities and the text are recorded by P, we can use it to fuse H^{ec} and H^{ls} through:

$$\tilde{H}^{lc} = P H^{ec} + H^{lc}, \tag{17}$$

where $\tilde{H}^{lc}$ is the text representations encoded with the image and entity information. Finally, we apply residual connections followed by layer normalization to $\tilde{H}^{lc}$ and input it to the feed-forward sub-layer to complete the knowledge-enhanced multi-modal fusion. In the meantime, H^{ec} is input to another feed-forward sub-layer to produce the representations Z^e for the next layer.

(iii) Learning From Knowledge Knowledge can help us to induce more sophisticated pretext tasks to guide the model to learn more informative representations. In our paper, we follow [46] to design a knowledge-induced mask generation strategy. Specifically, when performing the MLM task given the input text $T = \{x_1^l, x_2^l, \ldots, x_{N_l}^l\}$ with its entity sequence $ES = \{x_1^e, x_2^e, \ldots, x_{N_{es}}^e\}$, we do not mask subword tokens in T randomly. Instead, we randomly sample entities from ES and then mask consecutive spans of subword tokens belonging to the sampled entities. Since entities can be abstract or have a physical existence, it can force the model to focus on critical medical information in both images and texts.

Therefore, knowledge can be injected into the Med-VLP framework in a systematic way through the above three designs.

4 EXPERIMENTAL SETTINGS

4.1 Pre-training Setup

Datasets In our experiments, we perform the pre-training on three datasets, which are described as follows:

- **ROCO** [40]: a dataset of radiology figure-caption pairs from PubMed Central, an open-access biomedical literature database. It has over 81,000 radiology images (from various imaging modalities) and their corresponding captions.

Table 1: Results on the Med-VQA task (including three datasets, i.e., VQA-RAD, SLACK, and VQA-2019) to compare with the state-of-the-art methods. Dark and light grey colors highlight the top and second best results on each evaluation metric.

Dataset		MFB [56]	SAN [54]	BAN [23]	MEVF-SAN [39]	MEVF-BAN [39]	CPRD-BAN [30]	COND-REA [57]	MTPT [15]	MMBERT [22]	Ours
VQA-RAD	Open	14.50	31.30	37.40	49.20	49.20	52.50	60.00	61.50	63.10	67.60
	Closed	74.30	69.50	72.10	73.90	77.20	77.90	79.30	80.90	77.90	86.76
	Overall	50.60	54.30	58.30	64.10	66.10	67.80	71.60	73.20	72.00	79.16
SLACK	Open	72.20	74.00	74.60	75.30	77.80	79.50	-	-	-	81.89
	Closed	75.00	79.10	79.10	78.40	79.80	83.40	-	-	-	91.35
	Overall	73.30	76.00	76.30	76.50	78.60	81.10	-	-	-	85.59
VQA-2019	Overall	-	-	-	68.90	77.86	-	-	-	77.90	80.32

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

Dataset	Modalities	Methods	Accuracy
	Image-Only	ResNet-101 [18]	63.84
	Text-Only	LSTM [19]	59.20
		RoBERTa [33]	75.40
MELINDA		SciBERT [3]	77.70
	Multi-Modal	NLF [52]	76.60
		SAN [54]	72.30
		ViL-BERT [35]	78.60
		Ours	80.51

Table 3: Results on the Med-ITR task (i.e., the ROCO dataset) to compare with the state-of-the-art methods, where the zero-shot (Ours (ZS)) and fine-tuned results (Ours (FT)) are shown.

Methods	T2I			I2T		
	R@1	R@5	R@10	R@1	R@5	R@10
ViT+BERT [13]	5.25	15.85	25.85	6.85	21.25	31.60
ViLT [24]	9.75	28.95	41.40	11.90	31.90	43.20
METER [13]	11.30	27.25	39.60	14.45	33.30	45.10
Ours (ZS)	23.50	49.05	63.00	23.45	50.60	62.05
Ours (FT)	29.65	56.95	69.30	29.35	57.50	70.40

- **MedICaT** [44]: a dataset of medical figure-caption pairs also extracted from PubMed Central. Different from ROCO, 75% of its figures are compound figures, including several sub-figures. It contains over 217,000 images with their captions and inline textual references.
- **MIMIC-CXR** [21]: the largest radiology dataset to date that consists of 473,057 chest X-ray images (in frontal or lateral views) and 206,563 reports from 63,478 patients from the Beth Israel Deaconess Medical Center.

For all the datasets, we exclude those samples with the length of their texts less than 3. For ROCO and MedICaT, we filter non-radiology samples, and for MIMIC-CXR, we only keep images in the frontal view. As for the dataset split, we adopt the official splits of ROCO and MIMIC-CXR. For MedICaT, we randomly sample 1,000 image-text pairs for validation and 1,000 for test, and the remaining image-text pairs are used for training.

Implementation Details For the uni-modal encoders, we use the vision encoder with CLIP-ViT-B [42] (L_v = 12) and the language encoder with RoBERTa-base [33] (L_l = 12). For the multi-modal fusion module, we set the number of Transformer layers L_m = 6, and the dimension of the hidden states D = 768 with the number of heads set to 12. For knowledge representation and injection, we set the dimension of the hidden states D_e = 256. For the pretext tasks, we adopt (knowledge-enhanced) MLM, MIM [17], and ITM, where the masking ratios of MLM and MIM are set to 15% and 75%, respectively. For the optimization, the models are trained with

AdamW optimizer [34] for 100,000 steps with the learning rates for the uni-modal encoders and the remaining parameters set to 1e-5 and 5e-5, respectively. The warm-up ratio is set to 10%, and the learning rate is linearly decayed to 0 after warm-up. Besides, we use center-crop to resize each image to the size of 288×288.

4.2 Vision-and-Language Transfer Tasks

To evaluate the performance, we construct a medical vision-and-language understanding benchmark including three tasks. The details of the tasks and fine-tuning strategies are described below.

Medical Visual Question Answering (Med-VQA) This task requires the model to answer natural language questions about a medical image. We adopt three publicly available Med-VQA datasets (i.e., VQA-RAD [25], SLACK [31] and VQA-2019 [1]), where VQA-RAD consists of 315 images and 3515 questions, SLACK contains 642 images and 14,028 questions, and VQA-2019 contains 4,200 images and 15,292 questions. To fine-tune 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. During training, the models are trained with a binary cross-entropy loss with a batch size of 64.

Medical Image-Text Classification (Med-ITC) This task aims to produce the classification label given an image-text pair. We evaluate the performance on MELINDA [52], a Biomedical Experiment Method Classification dataset that contains 5,371 image-text pairs. To fine-tune 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 (Med-ITR) The target of this task is to calculate a similarity score between an image and a text and then perform cross-modal retrieval. There are two subtasks for this task, where image-to-text (I2T) retrieval requires retrieving the most relevant texts from a large pool of texts given an image, and vice versa for text-to-image (T2I) retrieval. We conduct experiments on the ROCO dataset and measure both zero-shot and fine-tuned performance. To fine-tune 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 all tasks, we use the AdamW optimizer with the learning rate set to 5e-6 and 2.5e-5 for the model backbone and prediction heads, respectively, and the warm-up ratio set to 10%. For the evaluation metrics, we adopt accuracy for Med-VQA and Med-ITC, and Recall@K[3] (K=1, 5, 10) for Med-ITR.

5 EXPERIMENTAL RESULTS

5.1 Main Results

We compare the proposed approach with existing methods on the same datasets, with their results reported in Table 1, 2, and 3. There are several observations drawn from different aspects. First, our approach achieves the best performance on all tasks, which confirms the validity of the proposed pre-training approach. Second, for Med-VQA, our approach achieves significant improvements even compared with those pre-training methods (e.g., MTPT and MM-BERT), which indicates the usefulness of incorporating knowledge into the pre-training process. Third, for Med-ITC, the results of those strong baselines (i.e., RoBERTa, SciBERT, and ViL-BERT) are achieved by continued pre-training on the MELINDA dataset. Our approach achieves better performance without such requirements, which indicates that an appropriate design can alleviate the need for continued pre-training on downstream datasets. Fourth, for Med-ITR, the proposed approach achieves a substantial improvement compared with the state-of-the-art methods, where ViLT and METER are two strong baselines in the general domain. This shows that it is necessary to design an appropriate approach (including pre-training data and methods) for the medical domain.

5.2 Ablation Studies

To illustrate the effectiveness of our proposed approach, we perform an ablation study on the three proposed knowledge injection designs. The experiments are conducted on the VQA-RAD dataset, and the results are reported in Table 4.

We have the following observations. First, for the model parameters, only the RK design brings extra 12M parameters (∼3.4% of the whole model) while other designs do not add additional parameters,

---

[2]The time complexity of the ranking process is $O(N^2)$, where N is the sample number.
[3]Recall@K corresponds to whether the ground truth is included among top K results.

Table 4: Ablation study on the three knowledge injection designs (i.e., aligning through knowledge (AK), reasoning using knowledge (RK), and learning from knowledge (LK)) on the VQA-RAD dataset, with the model parameters (Para.).

ID	AK	RK	LK	Para.	Open	Closed	Overall
1				350M	65.36	84.19	76.72
2	✓			350M	67.04	84.98	77.88
3		✓		362M	65.56	86.40	78.10
4			✓	350M	65.92	85.29	77.61
5	✓	✓		362M	67.60	86.08	78.76
6	✓		✓	350M	66.11	86.40	78.32
7		✓	✓	362M	65.36	86.40	78.05
8	✓	✓	✓	362M	67.60	86.76	79.16

which justifies introducing knowledge to Med-VLP through our approach can be done with a small price. Second, the results of pre-training with two designs (ID 5, 6, and 7) and one design (ID 2, 3, and 4) are consistently better than those of pre-training with one design (ID 2, 3, and 4) and without any knowledge injection design (ID 1), respectively. This demonstrates the excellent compatibility and complementarity of our design perspectives, which is critical in a multi-component approach and allows us to develop more designs under such a framework. Third, injecting the knowledge into the multi-modal fusion module (ID 3) achieves a significant improvement. The reason behind this might be that knowledge (i.e., entities here) serves three functions: (i) It smooths the interaction process of the image and text representations; (ii) It provides information at a greater granularity than words; (iii) It removes ambiguity between diverse words by linking to the knowledge base. Fourth, performing the aligning process can further improve the performance of the RK design (ID 5), which can be explained by the fact that the aligning processing can produce better knowledge representations for the RK process. Fifth, our full approach (ID 8) achieves the best performance, which confirms the effectiveness of the proposed framework for medical knowledge injection.

5.3 Qualitative Analysis

To further investigate the effectiveness of our approach, we perform qualitative analysis on some cases with their results shown in Figure 3. For "aligning through knowledge", in the input text of this case, there is an entity "*Brain magnetic resonance imaging*" which links to the entity "*C4028269: Nuclear magnetic resonance imaging brain*" in the UMLS knowledge base. The sub-figures (3(C) and 3(D)) show that the image and text representations produced by the uni-modal encoders have high similarity with the entity representation of "*C4028269*", which implicitly pulls the image and text representations close (as shown in the sub-figure 3(E)). For "reasoning using knowledge", the sub-figures 3(H), 3(I), and 3(J) illustrate that using entities is beneficial for aligning the text with the image, where the learned entity-image attention mappings are better than the subword-image mappings. The reason behind this is that entities are more complete semantic units. In contrast, words (or subwords) have a smaller granularity than entities, making the
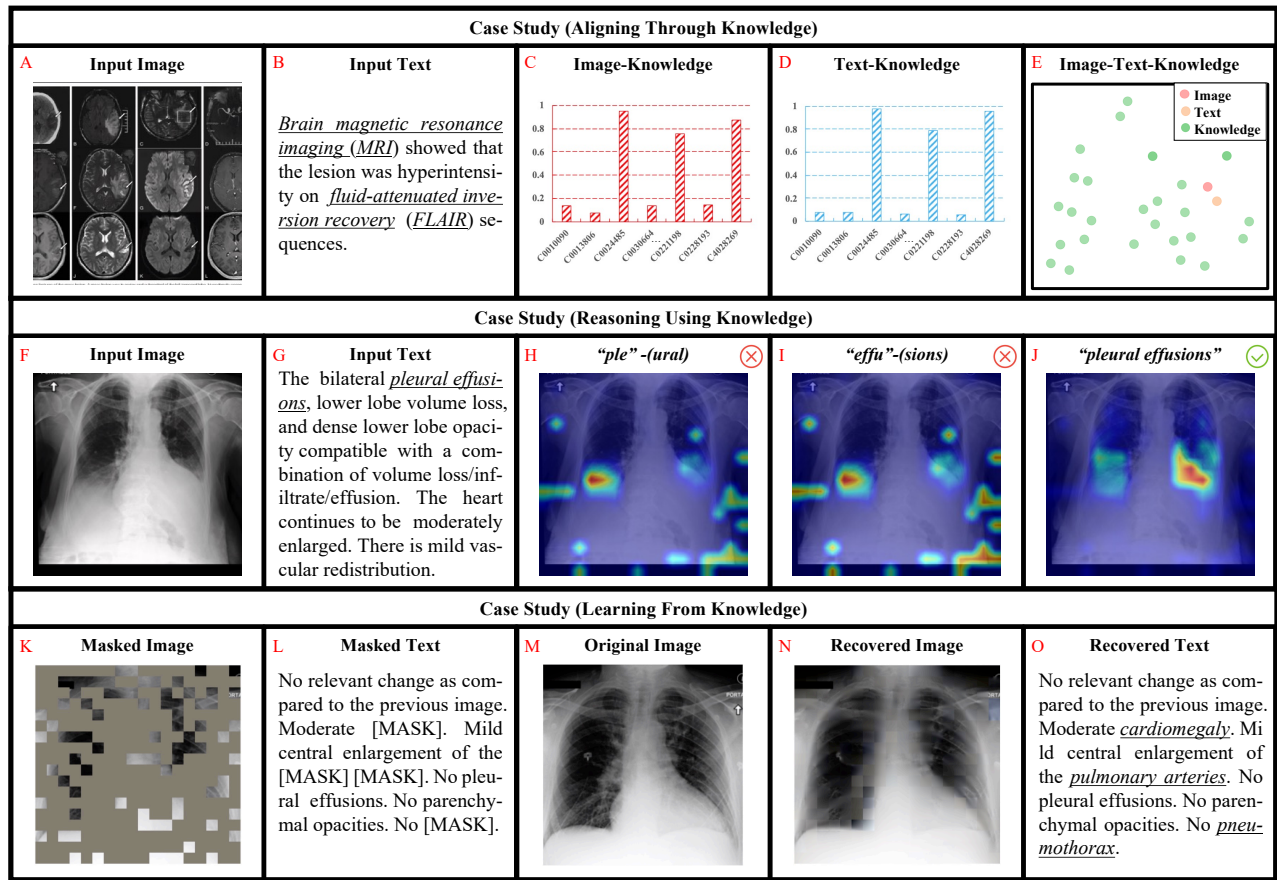
Figure 3: Visualizations of the three proposed designs from the pre-trained model, where image-knowledge and text-knowledge similarity and the image-text-knowledge t-SNE visualization are shown for "aligning through knowledge"; subword-image and entity-image attention mappings (colors from blue to red representing the weights from low to high) are shown for "reasoning using knowledge"; the recovered text (with the addition reconstructed image) is shown for "learning from knowledge".

correspondences between images and texts more challenging to learn. For "learning from knowledge", the masked medical entities are correctly recovered by the pre-trained model (as shown in the sub-figure 3(O)) since the knowledge-induced pretext task guides the model to put more emphasis on the medical knowledge-related information. In addition, the masked and recovered images are also shown in the sub-figures 3(K) and 3(M), respectively, which shows the high quality of the image reconstruction. In summary, these cases reveal that injecting knowledge through the three proposed designs is essential in modeling the hidden structures among the images and texts better to promote Med-VLP.

6 CONCLUSION

In this paper, we propose to pre-train the medical vision-and-language model with medical domain knowledge, where the knowledge is injected into the Med-VLP framework from three aspects: (i) aligning the image and text representations through knowledge before their interaction; (ii) treating knowledge as the supplementation of the input image and text to assist the reasoning during the multi-modal fusion process; (iii) utilizing knowledge to induce more sophisticated pretext tasks to guide the model put more emphasis on the critical medical information. To perform a comprehensive evaluation and facilitate further research, we construct a medical vision-and-language understanding benchmark, including three tasks (i.e., Med-VQA, Med-ITC, and Med-ITR). Experimental results on the downstream datasets demonstrate the effectiveness of our approach, which achieves state-of-the-art performance. Further analyses investigate the effects of different components in our approach and show that our approach is able to better learn the correspondences between vision and language so as to produce more generic and effective vision-and-language representations.

REFERENCES

[1] Asma Ben Abacha, Sadid A. Hasan, Vivek Datla, Joey Liu, Dina Demner-Fushman, and Henning Müller. 2019. VQA-Med: Overview of the Medical Visual Question Answering Task at ImageCLEF 2019. In *CLEF*.

[2] Saeed Amizadeh, Hamid Palangi, Alex Polozov, Yichen Huang, and Kazuhito Koishida. 2020. Neuro-Symbolic Visual Reasoning: Disentangling. In *International Conference on Machine Learning*. PMLR, 279–290.

[3] Iz Beltagy, Kyle Lo, and Arman Cohan. 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)*. 3615–3620.

[4] Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic acids research* 32, Database issue (2004), D267–D270.

[5] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems* 26 (2013).

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[7] Kezhen Chen, Qiuyuan Huang, Yonatan Bisk, Daniel McDuff, and Jianfeng Gao. 2021. KB-VLP: Knowledge Based Vision and Language Pretraining. In *Proceedings of the 38 th International Conference on Machine Learning, PMLR 139, 2021. ICML, workshop, 2021*.

[8] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*. Springer, 104–120.

[9] Yuhao Cui, Zhou Yu, Chunqi Wang, Zhongzhou Zhao, Ji Zhang, Meng Wang, and Jun Yu. 2021. ROSITA: Enhancing Vision-and-Language Semantic Alignments via Cross-and Intra-modal Knowledge Integration. In *Proceedings of the 29th ACM International Conference on Multimedia*. 797–806.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*. 4171–4186.

[11] Shizhe Diao, Jiaxin Bai, Yan Song, Tong Zhang, and Yonggang Wang. 2020. ZEN: Pre-training Chinese Text Encoder Enhanced by N-gram Representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. 4729–4740.

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

[13] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Zicheng Liu, Michael Zeng, et al. 2021. An Empirical Study of Training End-to-End Vision-and-Language Transformers. *arXiv preprint arXiv:2111.02387* (2021).

[14] Sedigheh Eslami, Gerard de Melo, and Christoph Meinel. 2021. Does CLIP Benefit Visual Question Answering in the Medical Domain as Much as it Does in the General Domain? *arXiv preprint arXiv:2112.13906* (2021).

[15] Haifan Gong, Guanqi Chen, Sishuo Liu, Yizhou Yu, and Guanbin Li. 2021. Cross-Modal Self-Attention with Multi-Task Pre-Training for Medical Visual Question Answering. In *Proceedings of the 2021 International Conference on Multimedia Retrieval*. 456–460.

[16] Bin He, Di Zhou, Jinghui Xiao, Xin Jiang, Qun Liu, Nicholas Jing Yuan, and Tong Xu. 2020. BERT-MK: Integrating Graph Contextualized Knowledge into Pre-trained Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. 2281–2290.

[17] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2021. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377* (2021).

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[19] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[20] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. 2020. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849* (2020).

[21] Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. 2019. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042* (2019).

[22] Yash Khare, Viraj Bagal, Minesh Mathew, Adithi Devi, U Deva Priyakumar, and CV Jawahar. 2021. Mmbert: Multimodal bert pretraining for improved medical vqa. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 1033–1036.

[23] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear attention networks. *Advances in Neural Information Processing Systems* 31 (2018).

[24] Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*. PMLR, 5583–5594.

[25] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data* 5, 1 (2018), 1–10.

[26] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems* 34 (2021).

[27] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557* (2019).

[28] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*. Springer, 121–137.

[29] Yikuan Li, Hanyin Wang, and Yuan Luo. 2020. 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, 1999–2004.

[30] Bo Liu, Li-Ming Zhan, and Xiao-Ming Wu. 2021. 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, 210–220.

[31] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021. SLAKE: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 1650–1654.

[32] Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-bert: Enabling language representation with knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 2901–2908.

[33] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[34] Ilya Loshchilov and Frank Hutter. 2018. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

[35] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems* 32 (2019).

[36] Kenneth Marino, Ruslan Salakhutdinov, and Abhinav Gupta. 2017. The More You Know: Using Knowledge Graphs for Image Classification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 20–28.

[37] Jong Hak Moon, Hyungyung Lee, Woncheol Shin, and Edward Choi. 2021. Multimodal Understanding and Generation for Medical Images and Text via Vision-Language Pre-Training. *arXiv preprint arXiv:2105.11333* (2021).

[38] Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*. 319–327.

[39] Binh D Nguyen, Thanh-Toan Do, Binh X Nguyen, Tuong Do, Erman Tjiputra, and Quang D Tran. 2019. Overcoming data limitation in medical visual question answering. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 522–530.

[40] Obioma Pelka, Sven Koitka, Johannes Rückert, Felix Nensa, and Christoph M Friedrich. 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, 180–189.

[41] Matthew E Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A Smith. 2019. Knowledge Enhanced Contextual Word Representations. 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)*. 43–54.

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.

[43] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. In *International Conference on Learning Representations*.

[44] Sanjay Subramanian, Lucy Lu Wang, Ben Bogin, Sachin Mehta, Madeleine van Zuylen, Sravanthi Parasa, Sameer Singh, Matt Gardner, and Hannaneh Hajishirzi. 2020. MedICaT: A Dataset of Medical Images, Captions, and Textual References. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. 2112–2120.

[45] Tianxiang Sun, Yunfan Shao, Xipeng Qiu, Qipeng Guo, Yaru Hu, Xuan-Jing Huang, and Zheng Zhang. 2020. CoLAKE: Contextualized Language and Knowledge Embedding. In *Proceedings of the 28th International Conference on Computational Linguistics*. 3660–3670.

[46] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223* (2019).

[47] Hao Tan and Mohit Bansal. 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)*. 5100–5111.

[48] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. In *International Conference on Learning Representations*.

[49] Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021. KEPLER: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics* 9 (2021), 176–194.

[50] Xiaolong Wang, Yufei Ye, and Abhinav Gupta. 2018. Zero-shot recognition via semantic embeddings and knowledge graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6857–6866.

[51] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2021. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904* (2021).

[52] Te-Lin Wu, Shikhar Singh, Sayan Paul, Gully Burns, and Nanyun Peng. 2021. MELINDA: A Multimodal Dataset for Biomedical Experiment Method Classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 14076–14084.

[53] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144* (2016).

[54] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. 2016. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 21–29.

[55] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. ERNIE-ViL: Knowledge Enhanced Vision-Language Representations through Scene Graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 3208–3216.

[56] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. 2017. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *Proceedings of the IEEE international conference on computer vision*. 1821–1830.

[57] Li-Ming Zhan, Bo Liu, Lu Fan, Jiaxin Chen, and Xiao-Ming Wu. 2020. Medical visual question answering via conditional reasoning. In *Proceedings of the 28th ACM International Conference on Multimedia*. 2345–2354.

[58] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5579–5588.

[59] Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced Language Representation with Informative Entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 1441–1451.