Clinical-BERT: Vision-Language Pre-Training for Radiograph Diagnosis and Reports Generation

Bin Yan, Mingtao Pei *

Beijing Laboratory of Intelligent Information Technology
School of Computer Science, Beijing Institute of Technology
{bean.yan, peimt}@bit.edu.cn

Abstract

In this paper, we propose a vision-language pre-training model, Clinical-BERT, for the medical domain, and devise three domain-specific tasks: Clinical Diagnosis (CD), Masked MeSH Modeling (MMM), Image-MeSH Modeling (IMM), together with one general pre-training task: Masked Language Modeling (MLM), to pre-train the model. The CD task helps the model to learn medical domain knowledge by predicting disease from radiographs. Medical Subject Headings (MeSH) words are important semantic components in radiograph reports, and the MMM task helps the model focus on the prediction of MeSH words. The IMM task helps the model learn the alignment of MeSH words with radiographs by matching scores obtained by a two-level sparse attention: region sparse attention and word sparse attention. Region sparse attention generates corresponding visual features for each word, and word sparse attention enhances the contribution of images-MeSH matching to the matching scores. To the best of our knowledge, this is the first attempt to learn domain knowledge during pre-training for the medical domain. We evaluate the pre-training model on Radiograph Diagnosis and Reports Generation tasks across four challenging datasets: MIMIC-CXR, IU X-Ray, COV-CTR, and NIH, and achieve state-of-the-art results for all the tasks, which demonstrates the effectiveness of our pre-training model.

Introduction

Vision-Language (VL) pre-training aims to learn general representations between images and text, that can help to improve the performance of downstream tasks such as Visual Question Answering, Visual Grounding, and Image Captioning. Many BERT (Devlin et al. 2018) based pre-training methods (Zhou et al. 2020; Huang et al. 2021; Zhuge et al. 2021) are proposed recently, and achieve promising performance under the context of general domain.

However, there are only few researches about VL pre-training for medical domain, in which there are also many vision-language tasks such as radiograph diagnosis (Rajpurkar et al. 2017) and reports generation (Jing, Xie, and Xing 2018; Chen et al. 2020b). Directly applying the model pre-trained in general domain to medical domain will cause

---

*Corresponding author.

Figure 1: Comparison of sentences in different domains. Radiograph reports are less diverse compared to natural language captions. In the upper part, words highlighted in same color indicate same semantic. In the bottom part, words highlighted are full matches (green) or sub-matches (yellow) for MeSH.

performance degradation since the two domains have different characteristics.

For example, radiograph report generation is similar to image captioning, but the radiograph report needs to be more exact than natural captions. As shown in Figure 1, in the five captions of the first image, the people are stated as "men", "teams" or "guys"; their actions are expressed as "play", "keep" or "kick"; the soccer is expressed as "soccer" or "ball". Such expressions with synonyms do not change the semantic information of sentences. However, radiograph reports requires the use of Medical Subject Headings (MeSH) words, such as "Lungs" and "Pleural Effusion", which are unique in expression and are important semantic components in the reports. It is logical that the MeSH words require more focus than other words.

As mentioned above, currently, there are only few researches about VL pre-training for medical domain. In current researches, models are usually pre-trained by general pre-training tasks such as Masked Language Modeling (MLM) and Image Report Matching (IRM). In these tasks, MeSH words and other words are treated equally, which ignores the important domain knowledge. The MeSH words should receive more attention in pre-training tasks, which can enable the pre-trained model to learn domain knowledge and obtain better performance in the downstream tasks.

Therefore, we propose a VL pre-training model Clinical-BERT, which can learn the knowledge of medical domain through three domain-specific pre-training tasks. The three tasks are Clinical Diagnosis (CD), Masked MeSH Modeling (MMM), and Image-MeSH Matching (IMM). In the CD task, we treat it as a multi-label classification problem, that is, to predict diseases from radiographs, that learns medical domain knowledge. The MMM task and IMM task focus on the MeSH words. In the MMM task, we randomly mask MeSH words instead of arbitrary words to help the model focus on MESH words. In the IMM task, we design a two-level sparse attention: region sparse attention and word sparse attention, to help the model learn the alignment of MeSH words with radiographs. The region sparse attention obtains region features that match each word, and word sparse attention assigns higher weights to the MeSH words. We also pre-train our model by Masked Language Modeling (MLM) task, in which both sequence to sequence and bidirectional prediction are employed as objectives.

Our model is pre-trained on MIMIC-CXR (Johnson et al. 2019), which contains large mounts of image-report pairs. We evaluate the pre-trained model on two downstream tasks: radiograph report generation and radiograph diagnosis. The radiograph report generation task is conducted on MIMIC-CXR, COV-CTR (Li et al. 2020b) and IU X-Ray (Demner-Fushman et al. 2016), and the radiograph diagnosis task is conducted on NIH ChestX-ray14 (Wang et al. 2017) dataset. The proposed pre-training model achieves promising performance on both tasks, which demonstrates that the pre-trained model can benefit greatly from the learning of medical domain knowledge. The contributions of this paper can be summarized as follows:

- We propose a VL pre-training model Clinical-BERT, which can learn medical domain knowledge to improve the performance of downstream tasks in medical domain. To the best of our knowledge, this is the first attempt to learn domain knowledge during pre-training for the medical domain.

- We design three domain-specific pre-training tasks to learn domain knowledge: the CD task learns a multi-label classification for diagnosis, the MMM task focuses on the predication of MeSH words and the IMM task aligns images and reports by region and word sparse attention.

- We conduct extensive experiments on downstream tasks of radiograph diagnosis and report generation, and achieve state-of-the-art performance on both tasks.

Related Works

Vision-Language Pre-Training

Pre-training models, such as BERT (Devlin et al. 2018), have recently achieved revolutionary progress in language tasks, and many BERT-based cross-modal pre-training models are proposed for VL understanding or generation tasks. These models can be categorized into two types: single-stream models (Su et al. 2020; Zhou et al. 2020; Li et al. 2020c; Hu et al. 2021) and two-stream models (Lu et al. 2019; Murahari et al. 2020; Li et al. 2020a). Generally,

single-stream models feed different modalities in a unified Transformer (Vaswani et al. 2017) encoder, while two-stream models adopt different Transformer encoders to process different modalities. Li et al. (Li et al. 2019b) proposed a single-stream model that reuses the self-attention to implicitly align the elements of input text and regions. Zhou et al. (Zhou et al. 2020) proposed a unified pre-training model with sequence to sequence and bidirectional prediction as objectives, and adapted the model to both understanding and generation tasks. Zhuge et al. (Zhuge et al. 2021) proposed to pay more attention to image-text coherence for the fashion domain.

For medical domain, some pre-training models (Lee et al. 2020; Zhang et al. 2020a) were proposed for single modality, and achieved promising results. Recently, Moon et al. (Moon et al. 2021) employed multi-modal pre-training for the medical domain, and proposed the Medical Vision Language Learner (MedViLL). However, they not fully utilized the domain knowledge. Different from current researches, we focus on pre-training the model with the medical domain knowledge.

Radiograph Reports Generation

Radiograph report generation receives more and more attention recently. The existing methods are divided into hierarchical models (Jing, Xie, and Xing 2018; Li et al. 2019a; Zhang et al. 2020b; Jing, Wang, and Xing 2019; Liu et al. 2019; Ni et al. 2020; Yang et al. 2021; Liu, Ge, and Wu 2021) and Transformer-based models (Chen et al. 2020b; Liu et al. 2021) .

In hierarchical models, reports are learned and generated hierarchically. Jing et al. (Jing, Xie, and Xing 2018) proposed the CoAtt, a co-attention model to generate radiograph reports hierarchically. Zhang et al. (Zhang et al. 2020b) extracted disease relationships from reports as prior knowledge and represented them as knowledge graphs, which improved the accuracy of generated reports.

Different from hierarchical models, the Transformer-based models learn the reports in one forward process. Chen et al. (Chen et al. 2020b) proposed a memory-driven Transformer, which can record key information of the generation process to assist the reports generation.

Radiograph Diagnosis

Radiograph diagnosis is a hot research topic, and lots of researches (Wang et al. 2018; Rajpurkar et al. 2017; Li et al. 2018; Luo et al. 2020; Yan et al. 2018; Chen et al. 2020a) have been proposed and achieved promising results. CheXNet (Rajpurkar et al. 2017) is a benchmark model pre-trained on ChestX-ray dataset. Luo et al. (Luo et al. 2020) introduced external data to assist the diagnosis and achieved excellent results.

Clinical-BERT

VL pre-training has achieved impressive results on cross-modal tasks under the context of general domain. However, applying these models directly to the medical domain would achieve poor results because of the domain gap. Therefore,
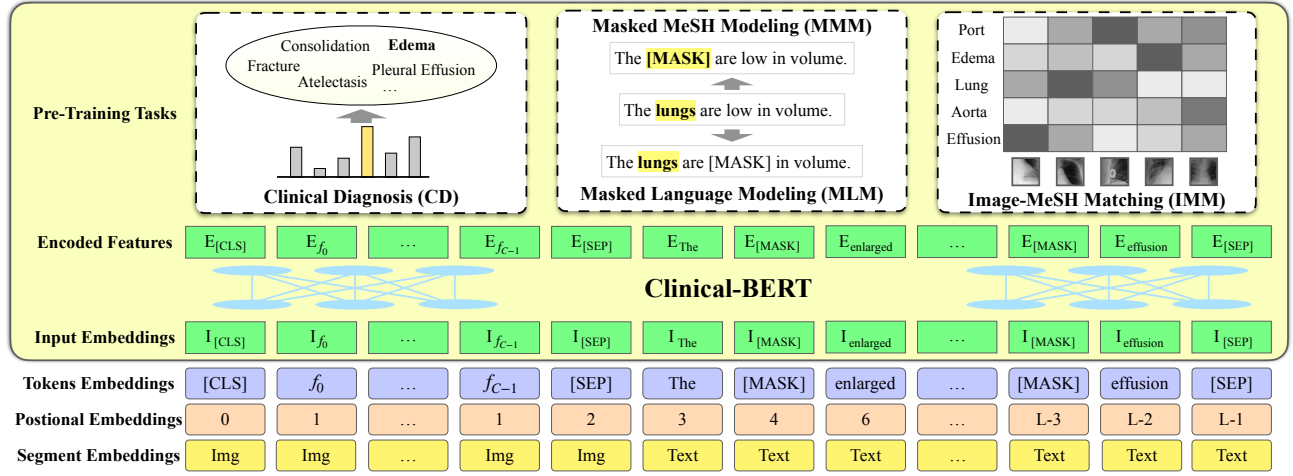
Figure 2: The architecture of Clinical-BERT. The BERT encoder is adopted to learn vision and language representation jointly. The CD, MMM, and IMM are devised domain-specific tasks. Words highlighted are MeSH words.

we propose the Clinical-BERT to learn the knowledge of the medical domain for medical tasks.

The overall architecture of the proposed Clinical-BERT is shown in Figure 2. There are three domain-specific pre-training tasks: Clinical Diagnosis (CD), Masked MeSH Modeling (MMM), and Image-MeSH Matching (IMM), and one general pre-training task: Masked Language Modeling (MLM). The single-stream BERT-base (Devlin et al. 2018) is adopted as the backbone to model both vision and language representations in a unified semantic space.

Input Representations

Given a radiograph and the corresponding report, we first extract visual features $f \in R^{C \times H \times W}$ from the radiograph by employing a convolutional neural network, where C is the number of feature channels, H and W are the height and width of features, respectively. We collapse the spatial dimensions of f and obtain $H \times W$ feature vectors $f_i \in R^C$ where f_i denotes the i-th feature vector. Then, the words in the report are embedded as word embeddings, and visual features and word embeddings are concatenated to obtain token embeddings, including the embeddings of special tokens [CLS], [SEP], [PAD], [UNK], and [MASK]. [CLS] indicates the start of the visual input. [SEP] is used to split the visual and linguistic input, and is also used as the end of the input. [PAD] is a language token used to pad the word sequence to a specified length. [UNK] denotes the filtered out words which occur less than three times. [MASK] is used to replace language tokens, and the replaced tokens will be predicted during the pre-training.

Then, input embeddings I are obtained by adding token embeddings with positional embeddings and segment embeddings. In particular, the position sequence of visual input is marked as 1 for positional embeddings, and segment embeddings contain vision (Img) and language (Text) tags. The input embeddings are fed into the BERT encoder to obtain encoded features E, and the encoded features are fed into the followed task-specific networks.

Pre-Training Tasks

Clinical Diagnosis (CD) We use the CheXpert (Irvin et al. 2019) labels as the disease annotation, which contains 14 categories of diseases. For each disease, there are four tags: positive, negative, uncertain, and absent. Specifically, we adopt the "U-Zeros" strategy to handle the source annotations, that is, all positive tags are marked as 1, and the others are marked as 0. Then, the CD task turns to be a multi-label classification problem.

In this task, diseases are predicted according to both vision and language modalities. We take the encoder output $E_{[CLS]}$ on the special token [CLS] as holistic visual representations and the $E_{[SEP]}$ on the first [SEP] as holistic linguistic features. These features are learned with bidirectional objective, that is, tokens are encoded based on their surrounding tokens. Then, the Hadamard product of $E_{[CLS]}$ and $E_{[SEP]}$ is obtained as the joint representation, which is fed into a neural network for disease prediction. The neural network consists of two fully connected (FC) layers and one ReLU activation layer, and the Sigmoid function is applied to predict scores between 0 and 1. The binary cross-entropy loss is employed for optimization:

$$\mathcal{L}_{\text{CD}} = -\frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{\mathcal{D}} \left[d_i \log p_i + (1 - d_i) \log (1 - p_i) \right], \quad (1)$$

where $N = 14$ is the number of diseases, $d_i \in \{0, 1\}$ indicates the presence or absence of the i-th disease, p_i is the predicted scores, and $\mathcal{D}$ denotes the dataset.

Masked MeSH Modeling (MMM) MeSH words are important for describing radiographs, hence the MMM task is proposed to focus on the prediction of MeSH words, which can help the model to learn the domain knowledge. Denote radiograph and the corresponding report as X and Y, respectively, where $Y = \{y_0, ..., y_l, ..., y_{L-1}\}$, y_l is the l-th word in the report, and L is the length of the report. We
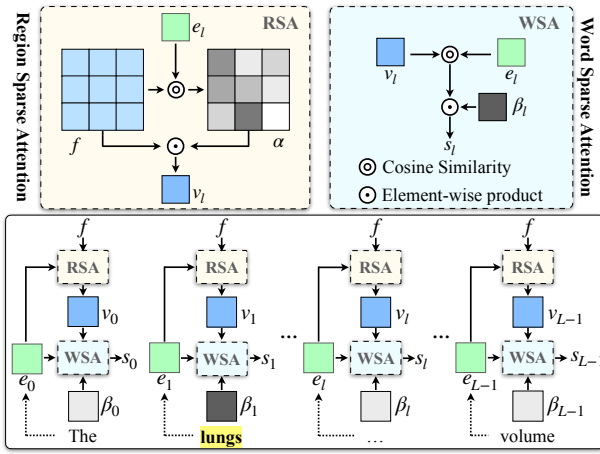
Figure 3: The details of the two-level sparse attention. Firstly, region features v_l are generated by the RSA. Then, matching scores for image-words pairs are generated by the WSA. The dotted line indicates the word encoding process.

label MeSH words in the report according to the MeSH table (Lipscomb 2000), and get tags $T = \{t_0, ..., t_l, ..., t_{L-1}\}$, where $t_l \in \{0, 1\}$. $t_l = 1$ denotes that the l-th word is a full match or sub-match of the MeSH words, and $t_l = 0$ denotes that the l-th word does not match the MeSH words. All the MeSH words in reports have an 80% chance of being replaced with [MASK] tokens, 10% chance of being replaced with random MeSH words, and 10% change of remaining unchanged.

The MMM task is to predict masked MeSH words based on the tokens on their left and the visual features f. The encoded features $E_{[MASK]}$ on the l-th [MASK] token, which is learned based on $\{y_0, ..., y_{l-1}\}$ and f with sequence to sequence objective, is projected to words likelihood to predict words. The prediction network consists of two FC layers, one normalization layer, and one ReLU layer. The cross-entropy loss is used for optimization:

$$\mathcal{L}_{\mathrm{MMM}} = -\mathbb{E}_{(f,y)\sim\mathcal{D}} \left[t_l \log p\left(y_l \mid y_0, ..., y_{l-1}, f \right) \right], \quad (2)$$

Masked Language Modeling (MLM) In the MLM task, 15% of the language tokens are replaced by [MASK], random tokens, and original tokens with 80%, 10%, and 10% chance, respectively. Specifically, MeSH words are not concerned in this task. The tokens behind the current token are invisible in the generation task, so for better adaptation to the generation task, we follow VLP (Zhou et al. 2020) to set two objectives: bidirectional (bi) prediction and sequence to sequence (s2s) prediction.

The goal is to predict the masked words, and the prediction process is similar to the MMM task. The MLM task shares the prediction network with MMM task, and adopts cross-entropy loss $\mathcal{L}_{MLM} = \mathcal{L}_{\mathrm{s2s}} + \mathcal{L}_{\mathrm{bi}}$ for optimization:

$$\mathcal{L}_{\mathrm{s2s}} = -\mathbb{E}_{(f,y)\sim\mathcal{D}} \left[\log p\left(y_l \mid y_0, ..., y_{l-1}, f \right) \right], \quad (3)$$

$$\mathcal{L}_{\mathrm{bi}} = -\mathbb{E}_{(f,y)\sim\mathcal{D}} \left[\log p\left(y_l \mid y_{\backslash l}, f \right) \right], \quad (4)$$

where $y_{\backslash l}$ is the surrounding tokens of the l-th word.

Image-MeSH Matching (IMM) In the IMM task, we align the images and MeSH words in a certain latent space by learning a cross-modal matching score. Inspired by Zhang et al. (Zhang et al. 2021), we propose a two-level sparse attention to learn the alignment: region sparse attention (RSA) and word sparse attention (WSA), as illustrated in Figure 3.

The RSA generates aligned region features for each word. This process mimics the focus of radiologists' interest when writing reports according to different observations. In particular, we regard each visual feature vector f_i as a region feature according to their perception field. Then, the region feature aligned to the l-th word can be formulated as $v_l = \sum_{i=1}^{C} \alpha_{l,i} f_i$, where C is the number of regions, $\alpha_{l,i}$ is the corresponding weight. The $\alpha_{l,i}$ is formulated as:

$$\alpha_{l,i} = \frac{\exp\left(\rho_1 \cos\left(e_l, f_i\right)\right)}{\sum_{h=1}^{C} \exp\left(\rho_1 \cos\left(e_l, f_h\right)\right)}, \quad (5)$$

where e_l is the encoded feature of the l-th word, $\cos(\cdot)$ is cosine similarity function. ρ_1 is a sharpening hyper-parameter, that approximates formula (5) to the argmax function when $\rho_1 \to \infty$.

Like the radiologists usually focusing on multiple while a small percent of regions in the radiograph, we employ a sparse attention mechanism to force the model to focus on a small set of critical regions. The top K weights are maintained, with the rest of the weights are set to negative infinity. Then, the weight $\alpha_{l,j}$ is overwritten by employing Softmax normalization to the revised weights $\alpha_{l,j}^K$:

$$\alpha_{l,j} = \frac{\exp\left(\alpha_{l,j}^K\right)}{\sum_{j=1}^{R} \exp\left(\alpha_{l,j}^K\right)}. \quad (6)$$

The WSA forces the model to focus on semantic components in the report to increase the contribution of MeSH words to the matching score. Firstly, the matching score of the encoded features e_l of the l-th word and aligned region features v_l is obtained by cosine similarity: $s_l = \beta_l \cos(e_l, v_l)$, where β is obtained by employing Softmax normalization to MeSH tags T: $\beta = \mathrm{Softmax}(T)$. Then, the matching score between radiograph and corresponding report is calculated as formula (7), which is also the loss function of the IMM:

$$\mathcal{L}_{IMM} = -\mathbb{E}_{\mathcal{D}} \left[\log \left(\sum_{l=1}^{L} \exp\left(\rho_2 s_l\right) \right)^{\frac{1}{\rho_2}} / \tau \right], \quad (7)$$

where ρ_2 and τ are sharpening hyper-parameters. In practice, we set $\rho_1 = \rho_2 = 4$ and $\tau = 0.01$. It should be noted that the calculation of formula (7) excludes the samples involved in the MMM. The reason is that replacing MeSH words in such samples by other words will result in image mismatches.

The overall pre-training loss is:

$$\mathcal{L} = \mathcal{L}_{CD} + \mathcal{L}_{MMM} + \mathcal{L}_{MLM} + \mathcal{L}_{IMM}. \quad (8)$$

Dataset	Model	B@1↑	B@2↑	B@3↑	B@4↑	R↑	C↑
IU X-Ray	R2Gen (Chen et al. 2020b)	0.470	0.304	0.219	0.165	0.371	-
	CMN (Chen et al. 2021)	0.475	0.309	0.222	*0.170*	0.375	-
	KERP (Li et al. 2019a)	0.482	*0.325*	*0.226*	0.162	0.339	0.280
	PPKED (Liu et al. 2021)	*0.483*	0.315	0.224	0.168	*0.376*	*0.351*
	Ours	**0.495**	**0.330**	**0.231**	**0.170**	**0.376**	**0.432**
COV-CTR	ST (Vinyals et al. 2015)	0.697	0.621	0.568	0.515	0.723	0.659
	COATT (Jing, Xie, and Xing 2018)	0.709	0.645	0.603	0.552	**0.748**	0.672
	ASGK (Li et al. 2020b)	*0.712*	*0.659*	*0.611*	*0.570*	0.746	*0.684*
	Ours	**0.759**	**0.713**	**0.675**	**0.641**	*0.737*	**1.218**

Table 1: Results on IU X-Ray and COV-CTR. B@n for BLEU-n, R for ROUGE-L, C for CIDEr, and M for METEOR. The results are quoted from published literature. Numbers in bold are the best result, and numbers in italic are the second best.

Model	B@1↑	B@2↑	B@3↑	B@4↑	M↑	R↑	Precision↑	Recall↑	F1↑
TOPDOWN (Anderson et al. 2018)	0.317	0.195	0.130	0.092	0.128	0.267	0.322	0.239	0.249
R2Gen (Chen et al. 2020b)	0.353	0.218	0.145	0.103	0.142	0.277	0.333	0.273	0.276
CMN (Chen et al. 2021)	0.353	0.218	0.148	0.106	0.142	*0.278*	*0.334*	*0.275*	*0.278*
PPKED (Liu et al. 2021)	*0.360*	*0.224*	*0.149*	*0.106*	**0.149**	**0.284**	-	-	-
Ours	**0.383**	**0.230**	**0.151**	**0.106**	*0.144*	0.275	**0.397**	**0.435**	**0.415**

Table 2: Comparison results on the pre-training dataset MIMIC-CXR. NLG and clinical accuracy metrics are reported.

Experiments

Datasets

We pre-train the Clinical-BERT on MIMIC-CXR (Johnson et al. 2019) dataset, which contains 377,110 chest X-ray images and 227,835 reports. The CheXpert labels are adopted as annotations for the CD task, which are extracted from the reports by the rule-based CheXpert labeler (Irvin et al. 2019). For a fair comparison, we use the official splitting for training, validation, and testing.

The radiograph reports generation task is conducted on IU X-Ray (Demner-Fushman et al. 2016) and COV-CTR (Li et al. 2020b). IU X-Ray consists of 7,470 frontal and lateral-view chest X-ray images and 3,955 corresponding reports. The findings and impression sections are concatenated as the ground-truth reports. COV-CTR is a Chinese medical reports dataset, which consists of 728 images and the corresponding reports. We randomly split both datasets into training, validation, and testing in the ratio of 7:1:2.

The radiograph diagnosis task is conducted on NIH (Wang et al. 2017). The NIH contains 112,120 images from 32,717 patients, and each image is labeled with 14 disease labels. The official splitting set is adopted in the experiment.

Downstream Tasks

We evaluate our model on the radiograph reports generation task and the radiograph diagnosis task by end-to-end fine-tuning the pre-trained model. The flowcharts of these two downstream tasks are illustrated in Figure 4.

Radiograph Reports Generation (RRG) Given a radiograph, the RRG task is to output coherent reports for the radiograph. We fine-tune the pre-trained Clinical-BERT without the bidirectional objective. During the inference stage, firstly, visual features and special tokens [CLS] and [SEP] are embedded as input sequences, the [MASK] token is embedded and concatenated for the prediction of the first word.
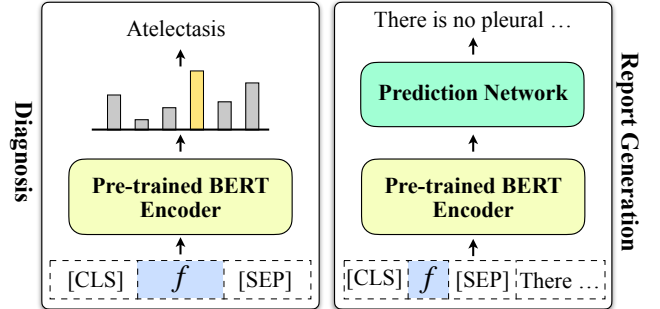

Figure 4: The flowcharts of downstream tasks. f is the visual features

Then, the [MASK] token is replaced by the predicted word, and a new [MASK] token is appended to the input sequence. This step repeats until the [SEP] token is generated.

The performance of the RRG is evaluated by Natural Language Generation (NLG) metrics that include BLEU (Papineni et al. 2002), METEOR (Denkowski and Lavie 2011), CIDEr (Vedantam, Zitnick, and Parikh 2015) and ROUGE-L (Lin 2004) scores. In addition, clinical accuracy is adopted to evaluate whether the generated reports give an accurate diagnosis. We label the generated reports with the CheXpert labeler, and the clinical accuracy is reported as the precision, recall, and F1 scores for the generated reports.

Radiograph Diagnosis (RD) Given a radiograph, the RD task is to output the disease labels for the radiograph. Similarly, visual features and special tokens $[CLS]$ and $[SEP]$ are embedded as the input sequences. Then, we employ one FC layer to the encoded output $E_{[CLS]}$ for the prediction of labels. The Sigmoid function is used to predict scores between 0 and 1. When the predicted score is above the threshold (set as 0.5 in our experiment), the corresponding disease

Image	Ground Truth	Baseline	Ours
	Single portable view of the ==chest==. ==Right chest wall port== is again seen. Streaky left basilar and right upper ==lung opacities== are seen suggestive of ==atelectasis== or scarring. ==Calcified mediastinal nodes== are again seen. ==Cardiomediastinal== silhouette is within normal limits. No acute ==osseous== abnormality detected.	Lung volumes are low. Heart size is normal. Mediastinal and hilar contours are unremarkable. Streaky ==opacities== in the ==lung== bases likely reflect areas of ==atelectasis==. Pulmonary vasculature is not engorged. No focal consolidation pleural effusion or pneumothorax is seen. There are no acute ==osseous== abnormalities.	Ap portable upright view of the ==chest==. ==Port-a-cath== resides over the right ==chest wall== with catheter tip in the region of the mid svc. Streaky ==opacities== in the ==lung== bases likely reflect areas of ==atelectasis==. The ==cardiomediastinal== silhouette is normal. Imaged ==osseous== structures are intact.

Figure 5: Examples from MIMIC-CXR. Words highlighted in green are terms that occur in both generated and ground-truth reports. Terms not generated are highlighted in yellow.
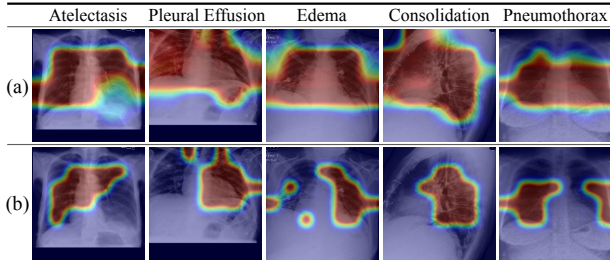

Figure 6: Attention maps of MeSH words. Row (a) is the attention map without sparsity and row (b) is the attention map with sparsity.

is regarded as positive. The Area Under Curve (AUC) for pathology is adopted as the evaluation metric.

Disease	CheXNet	WS	DME	Ours
Atelectasis	0.7795	*0.7924*	0.7891	**0.8293**
Cardiomegaly	0.8816	0.8814	*0.9069*	**0.9174**
Consolidation	0.7542	0.7598	*0.7681*	**0.8061**
Edema	0.8496	0.8478	*0.8610*	**0.9007**
Effusion	0.8268	0.8415	*0.8418*	**0.8851**
Emphysema	0.9249	**0.9422**	*0.9396*	0.9249
Fibrosis	0.8219	0.8326	*0.8381*	**0.8385**
Hernia	0.9323	0.9341	*0.9371*	**0.9379**
Infiltration	0.6894	0.7095	**0.7184**	*0.7156*
Mass	0.8307	*0.8470*	0.8376	**0.8524**
Nodule	0.7814	**0.8105**	*0.7985*	0.7807
Pleural T.	0.7925	**0.8083**	*0.8036*	0.7857
Pneumonia	0.7354	0.7397	*0.7419*	**0.7703**
Pneumothorax	0.8513	0.8795	**0.9063**	*0.8853*
Mean	0.8180	0.8302	*0.8349*	**0.8450**

Table 3: Results of radiograph diagnosis on NIH dataset. "Pleural T." denotes "Pleural Thickening".

Implementation Details

We employ the pre-trained uncased BERT-base (Devlin et al. 2018) as backbone, and DenseNet121 (Huang, Liu, and Weinberger 2017) pre-trained on ImageNet (Russakovsky et al. 2015) as visual feature extractor. The task-specific modules are initialized randomly.

For the pre-training, we set the data ratio for s2s:bi:MMM tasks to 0.5:0.25:0.25. Words in the MIMIC-CXR that occur more than 3 times are tokenized and 7861 tokens are obtained. The maximum length of the report is set to 100. All images are resized and cropped into 224×224. The AdamW (Loshchilov and Hutter 2019) optimizer is adopted with a weight decay of 0.01. Batch size is set as 256 with gradient accumulation (every 4 steps). The learning rate for the backbone and the visual extractor are $1e-4$ and $5e-5$, respectively. We pre-train the model for 50 epochs. All experiments are run on two Nvidia 3090 GPUs. The number of model parameters is 102M, the memory cost is 36G for each gradient accumulation step, the training time is 48 hours for 50 epochs, and the inference time is 0.1 seconds per image.

For the fine-tuning of the RRG, the data ratio of s2s:MMM is set to 0.75:0.25. Words that occur more than 3 times are tokenized into 764 tokens for IU X-Ray and the maximum prediction length is set to 60. Jieba (Sun 2012) is employed to tokenize words in COV-CTR and yields 333 tokens, and the maximum prediction length is set to 80. Cross-entropy is used as a loss function. We adopt the AdamW optimizer with a batch size of 16 for both datasets and a learning rate of $1e-5$. During the inference stage, we adopt the beam search strategy with a beam size of 5.

For the fine-tuning of the RD, the binary cross-entropy is used as a loss function. We adopt the AdamW optimizer with a batch size of 32 and a learning rate of $1e-5$.

Results on Downstream tasks

The comparison results of the RRG task are reported in Table 1. For IU X-Ray, our model achieves the best results under all NLG metrics. For the clinical accuracy, we achieve the precision, recall, and F1 scores of 48.52%, 42.79%, and 45.47%, respectively, with an average 8.6% gain compared with R2Gen (Chen et al. 2020b). For COV-CTR, we achieve the best results under five of the six NLG metrics. Furthermore, the AUC for COVID-19 prediction is up to 94.72%, with a large margin with 17.8% gain compared to ASGK (Li et al. 2020b). The experimental results show that the designed pre-training tasks can effectively learn domain knowledge and improve performance.

We also report the results on the test set of pre-training dataset MIMIC-CXR for the RRG task in Table 2. It can be seen that our model achieves the best results for BLEU-n and clinical accuracy metrics, and achieves comparable results for METEOR and ROUGE-L. The results demonstrate that reports generated by our model are not only fluent but also more accurate in clinical diagnosis.

We show some generated reports in Figure 5 for qualitative analysis. The Baseline is the model pre-trained with the MLM task. It can be seen that reports generated by our model pre-trained with medical domain tasks are accurate

Dataset	Fine-tune From	B@1↑	B@2↑	B@3↑	B@4↑	R↑	C↑	M↑	△(%)
	Baseline	0.465	0.299	0.213	0.158	0.369	0.328	0.196	-
	+CD	0.479	0.312	0.221	0.165	0.377	0.361	0.197	4.7
IU X-Ray	+CD+MMM	0.489	0.317	0.229	**0.173**	**0.385**	0.378	0.203	1.4
	+CD+MMM+IMM (w/o sparsity)	0.488	0.321	0.230	0.171	0.371	0.402	**0.211**	1.5
	+CD+MMM+IMM (w/ sparsity)	**0.495**	**0.330**	**0.231**	0.170	0.376	**0.432**	0.209	1.8
	Baseline	0.345	0.215	0.145	0.104	0.274	0.144	0.135	-
	+CD	0.359	0.221	0.148	0.105	0.271	0.146	0.139	2.0
MIMIC-CXR	+CD+MMM	0.363	0.224	0.149	0.105	0.271	0.149	0.136	0.3
	+CD+MMM+IMM (w/o sparsity)	0.372	0.228	0.151	0.105	0.274	0.144	0.142	1.2
	+CD+MMM+IMM (w/ sparsity)	**0.383**	**0.230**	**0.151**	**0.106**	**0.275**	**0.151**	**0.144**	1.6

Table 4: Ablation study results on IU X-Ray and MIMIC-CXR.
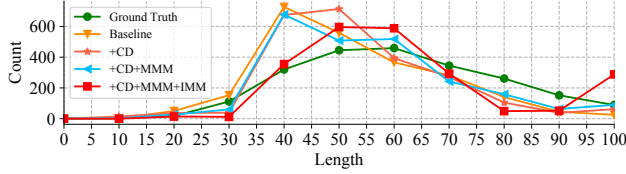
Figure 7: The length distribution of generated reports from different pre-training models on MIMIC-CXR.

K	6	8	12	16	20	24
B@1	0.357	0.381	**0.383**	0.378	0.380	0.382
B@2	0.216	0.227	**0.230**	0.225	0.227	0.228
M	0.137	0.144	**0.144**	0.143	0.144	0.144
R	0.271	0.272	**0.275**	0.270	0.274	0.270

Table 5: Effect of K for the sparsity on the MIMI-CXR.

in semantic, and most MeSH words are predicted accurately (highlighted in green), such as "Atelectasis", "Lung", "Osseous", and "chest". However, some MeSH words are not predicted (highlighted in yellow). The reason may be that there is no ground truth for the IMM task, making the joint representation of modalities less accurate, which limits the prediction of MeSH words. We also show the attention maps of MeSH words in Figure 6. It can be seen that our model can focus on regions related to MeSH, and the area of attention maps are significantly reduced by the introduced sparsity.

The results of the RD task on the NIH dataset are reported in Table 3. We compare our model with state-of-the-art models CheXNet (Rajpurkar et al. 2017), WS (Yan et al. 2018) and DME (Luo et al. 2020), and achieve the best results on nine of the fourteen pathologies and on the average AUC. The results demonstrate that the domain knowledge learned by the domain-specific pre-training tasks can improve the performance of radiograph diagnosis.

Ablation Study

We conduct ablation studies to evaluate the effectiveness of the domain-specific pre-training tasks. The Baseline denotes the baseline model pre-trained on MIMIC-CXR without the domain-specific pre-training tasks, +CD, +CD+MMM, and +CD+MMM+IMM denote the models pre-trained with different domain-specific pre-training tasks, and the effectiveness of the sparsity in IMM is also analyzed. The experimental results are shown in Table 4.

The CD task optimizes the joint representation of modalities for a better understanding of radiographs. The MLM task focuses on predicting randomly chosen words and ensures the consistency of sentences. While the MMM task focuses on the prediction of MeSH words and guarantees the accuracy of sentences. The IMM task aligns vision and language in the latent space. Region sparse attention focuses on the regions related to words. While word sparse attention focuses on the matching of MeSH words to images. The introduced sparsity reduces the area of interest, and enables the generation of more accurate reports. The average gains after adding tasks are listed in Table 4. Figure 7 shows the length distribution of the reports generated by different models. The length distribution of the reports generated by the final model is closer to the real distribution, which further demonstrates the effectiveness of devised pre-training tasks.

We also analyze the effect of K for the sparsity in IMM, which reduces the area of interest, and in turn, affects the report generation. We pre-train our model with different values and fine-tune it on MIMIC-CXR. The results are shown in Table 5. We can see that our model obtains best performance when $K = 12$. It should be noticed that the performance drops when K increased, which suggests that only a small percent of regions are contributive to the reports generation.

Conclusion and Future Work

In this paper, we propose a pre-training model Clinical-BERT for the medical domain. The model is pre-trained by three domain-specific tasks: Clinical Diagnosis (CD), Masked MeSH Modeling (MMM), Image-MeSH Matching (IMM), and one general pre-training task: Masked Language Modeling (MLM). The domain-specific tasks focus on the learning of medical knowledge and medical subject headings, which can help the understanding of radiographs and greatly boost the performance of downstream tasks. The results on radiograph diagnosis and report generation prove the effectiveness of the domain-specific pre-training.

In the future, we would employ localization information of organs for the IMM task to further improve the prediction accuracy of MeSH words. In addition, more downstream tasks will be evaluated, such as image-report retrieval and medical visual question answering.

References

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. *CVPR*, 6077–6086.

Chen, B.; Li, J.; Lu, G.; and Zhang, D. 2020a. Lesion Location Attention Guided Network for Multi-Label Thoracic Disease Classification in Chest X-Rays. *IEEE Journal of Biomedical and Health Informatics*, 24: 2016–2027.

Chen, Z.; Shen, Y.; Song, Y.; and Wan, X. 2021. Cross-modal Memory Networks for Radiology Report Generation. In *ACL/IJCNLP*.

Chen, Z.; Song, Y.; Chang, T.-H.; and Wan, X. 2020b. Generating Radiology Reports via Memory-driven Transformer. In *EMNLP*.

Demner-Fushman, D.; Kohli, M.; Rosenman, M.; Shooshan, S. E.; Rodriguez, L. M.; Antani, S.; Thoma, G.; and McDonald, C. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association : JAMIA*, 23 2: 304–10.

Denkowski, M. J.; and Lavie, A. 2011. Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. In *WMT@EMNLP*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Hu, X.; Yin, X.; Lin, K.; Zhang, L.; Gao, J.; Wang, L.; and Liu, Z. 2021. VIVO: Visual Vocabulary Pre-Training for Novel Object Captioning. In *AAAI*.

Huang, G.; Liu, Z.; and Weinberger, K. Q. 2017. Densely Connected Convolutional Networks. *CVPR*, 2261–2269.

Huang, Z.; Zeng, Z.; Huang, Y.; Liu, B.; Fu, D.; and Fu, J. 2021. Seeing Out of tHe bOx: End-to-End Pre-training for Vision-Language Representation Learning. *arXiv preprint*, arXiv:2104.03135.

Irvin, J. A.; Rajpurkar, P.; Ko, M.; Yu, Y.; Ciurea-Ilcus, S.; Chute, C.; Marklund, H.; Haghgoo, B.; Ball, R. L.; Shpanskaya, K.; Seekins, J.; Mong, D.; Halabi, S.; Sandberg, J.; Jones, R. H.; Larson, D.; Langlotz, C.; Patel, B.; Lungren, M.; and Ng, A. 2019. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. In *AAAI*.

Jing, B.; Wang, Z.; and Xing, E. 2019. Show, Describe and Conclude: On Exploiting the Structure Information of Chest X-ray Reports. In *ACL*.

Jing, B.; Xie, P.; and Xing, E. 2018. On the Automatic Generation of Medical Imaging Reports. In *ACL*.

Johnson, A. E. W.; Pollard, T.; Berkowitz, S. J.; Greenbaum, N. R.; Lungren, M.; ying Deng, C.; Mark, R.; and Horng, S. 2019. MIMIC-CXR: A large publicly available database of labeled chest radiographs. *arXiv preprint*, arXiv:1901.07042.

Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; and Kang, J. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36: 1234 – 1240.

Li, C. Y.; Liang, X.; Hu, Z.; and Xing, E. 2019a. Knowledge-driven Encode, Retrieve, Paraphrase for Medical Image Report Generation. *arXiv preprint*, arXiv:1903.10122.

Li, G.; Duan, N.; Fang, Y.; Jiang, D.; and Zhou, M. 2020a. Unicoder-VL: A Universal Encoder for Vision and Language by Cross-modal Pre-training. In *AAAI*.

Li, L. H.; Yatskar, M.; Yin, D.; Hsieh, C.-J.; and Chang, K.-W. 2019b. VisualBERT: A Simple and Performant Baseline for Vision and Language. *arXiv preprint*, arXiv:1908.03557.

Li, M.; Wang, F.; Chang, X.; and Liang, X. 2020b. Auxiliary Signal-Guided Knowledge Encoder-Decoder for Medical Report Generation. *arXiv preprint*, arXiv:2006.03744.

Li, X.; Yin, X.; Li, C.; Hu, X.; Zhang, P.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; Choi, Y.; and Gao, J. 2020c. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In *ECCV*.

Li, Z.; Wang, C.; Han, M.; Xue, Y.; Wei, W.; Li, L.-J.; and Li, F. 2018. Thoracic Disease Identification and Localization with Limited Supervision. *CVPR*, 8290–8299.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *ACL*.

Lipscomb, C. E. 2000. Medical subject headings (MeSH). *Bulletin of the Medical Library Association*, 88(3): 265.

Liu, F.; Ge, S.; and Wu, X. 2021. Competence-based Multimodal Curriculum Learning for Medical Report Generation. In *ACL/IJCNLP*.

Liu, F.; Wu, X.; Ge, S.; Fan, W.; and Zou, Y. 2021. Exploring and Distilling Posterior and Prior Knowledge for Radiology Report Generation. In *CVPR*, 13753–13762.

Liu, G.; Hsu, T.; McDermott, M. B. A.; Boag, W.; Weng, W.; Szolovits, P.; and Ghassemi, M. 2019. Clinically Accurate Chest X-Ray Report Generation. *arXiv preprint*, arXiv:1904.02633.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *ICLR*.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *NeurIPS*.

Luo, L.; Yu, L.; Chen, H.; Liu, Q.; Wang, X.; Xu, J.; and Heng, P. 2020. Deep Mining External Imperfect Data for Chest X-Ray Disease Screening. *IEEE Transactions on Medical Imaging*, 39: 3583–3594.

Moon, J. H.; Lee, H.; Shin, W.; and Choi, E. 2021. Multimodal Understanding and Generation for Medical Images and Text via Vision-Language Pre-Training. *arXiv preprint*, arXiv:2105.11333.

Murahari, V. S.; Batra, D.; Parikh, D.; and Das, A. 2020. Large-scale Pretraining for Visual Dialog: A Simple State-of-the-Art Baseline. In *ECCV*.

Ni, J.; Hsu, C.-N.; Gentili, A.; and McAuley, J. 2020. Learning Visual-Semantic Embeddings for Reporting Abnormal Findings on Chest X-rays. *arXiv preprint*, arXiv:2010.02467.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *ACL*.

Rajpurkar, P.; Irvin, J. A.; Zhu, K.; Yang, B.; Mehta, H.; Duan, T.; Ding, D.; Bagul, A.; Langlotz, C.; Shpanskaya, K.; Lungren, M.; and Ng, A. 2017. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. *arXiv preprint*, arXiv:1711.05225.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M. S.; Berg, A.; and Fei-Fei, L. 2015. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 115: 211–252.

Su, W.; Zhu, X.; Cao, Y.; Li, B.; Lu, L.; Wei, F.; and Dai, J. 2020. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. *arXiv preprint*, arXiv:1908.08530.

Sun, J. 2012. Jieba chinese word segmentation tool.

Vaswani, A.; Shazeer, N. M.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. *arXiv preprint*, arXiv:1706.03762.

Vedantam, R.; Zitnick, C. L.; and Parikh, D. 2015. CIDEr: Consensus-based image description evaluation. *CVPR*, 4566–4575.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. *CVPR*, 3156–3164.

Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Bagheri, M.; and Summers, R. 2017. ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. *CVPR*, 3462–3471.

Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; and Summers, R. 2018. TieNet: Text-Image Embedding Network for Common Thorax Disease Classification and Reporting in Chest X-Rays. *CVPR*, 9049–9058.

Yan, C.; Yao, J.; Li, R.; Xu, Z.; and Huang, J. 2018. Weakly Supervised Deep Learning for Thoracic Disease Classification and Localization on Chest X-rays. *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*.

Yang, X.; Ye, M.; You, Q.; and Ma, F. 2021. Writing by Memorizing: Hierarchical Retrieval-based Medical Report Generation. *arXiv preprint*, arXiv:2106.06471.

Zhang, H.; Koh, J. Y.; Baldridge, J.; Lee, H.; and Yang, Y. 2021. Cross-Modal Contrastive Learning for Text-to-Image Generation. *arXiv preprint*, arXiv:2101.04702.

Zhang, Y.; Jiang, H.; Miura, Y.; Manning, C. D.; and Langlotz, C. 2020a. Contrastive Learning of Medical Visual Representations from Paired Images and Text. *arXiv preprint*, arXiv:2010.00747.

Zhang, Y.; Wang, X.; Xu, Z.; Yu, Q.; Yuille, A.; and Xu, D. 2020b. When Radiology Report Generation Meets Knowledge Graph. *arXiv preprint*, arXiv:2002.08277.

Zhou, L.; Palangi, H.; Zhang, L.; Hu, H.; Corso, J. J.; and Gao, J. 2020. Unified Vision-Language Pre-Training for Image Captioning and VQA. In *AAAI*.

Zhuge, M.; Gao, D.; Fan, D.-P.; Jin, L.; Chen, B.; Zhou, H.; Qiu, M.; and Shao, L. 2021. Kaleido-BERT: Vision-Language Pre-training on Fashion Domain. *arXiv preprint*, arXiv:2103.16110.