Multi-modal Understanding and Generation for Medical Images and Text via Vision-Language Pre-Training

Jong Hak Moon*, Hyungyung Lee*, Woncheol Shin, Young-Hak Kim, and Edward Choi

***Abstract*— Recently a number of studies demonstrated impressive performance on diverse vision-language multi-modal tasks such as image captioning and visual question answering by extending the BERT architecture with multi-modal pre-training objectives. In this work we explore a broad set of multi-modal representation learning tasks in the medical domain, specifically using radiology images and the unstructured report. We propose Medical Vision Language Learner (MedViLL), which adopts a BERT-based architecture combined with a novel multi-modal attention masking scheme to maximize generalization performance for both vision-language understanding tasks (diagnosis classification, medical image-report retrieval, medical visual question answering) and vision-language generation task (radiology report generation). By statistically and rigorously evaluating the proposed model on four downstream tasks with three radiographic image-report datasets (MIMIC-CXR, Open-I, and VQA-RAD), we empirically demonstrate the superior downstream task performance of MedViLL against various baselines, including task-specific architectures. The source code is publicly available at: https://github.com/SuperSupermoon/MedViLL**

***Index Terms*— Healthcare, Medical, Multimodal Learning, Representation Learning, Self-Supervised Learning, Vision-and-Language**

I. Introduction

Vision-Language (VL) multi-modal research using radiographic images and associated free-text description (e.g., Chest X-rays and radiology report) is one of the most important and interesting works in the medical informatics [4], [6], [18], [22], [24], [31], [42]. Although each VL modality provides different representations to the researcher, images and reports contain mutually helpful semantic information. Consequently, advances in VL multi-modal research can be beneficial in improving the quality of clinical care by providing automated support for a variety of tasks such as diagnosis classification [6], [18], [42], report generation [24], [42]. Owing to the high dimensionality, heterogeneity, and systemic biases, however, handling both image and clinical report to learn joint representation poses significant technical challenges. The development of VL multi-modal learning has produced tremendous progress recently by extending the BERT-based architecture [10] in deep learning area. BERT-based VL model is the typical pretrain-then-transfer approach that makes the model learn a representation of each modality by performing multiple pre-training tasks. After pre-training the model, it is transferred to various vision language understanding (VLU) (e.g., visual question answering, text-conditioned image retrieval and vice versa) and vision language generation (VLG) (e.g., image captioning) downstream tasks by making only minor additions to the base architecture.

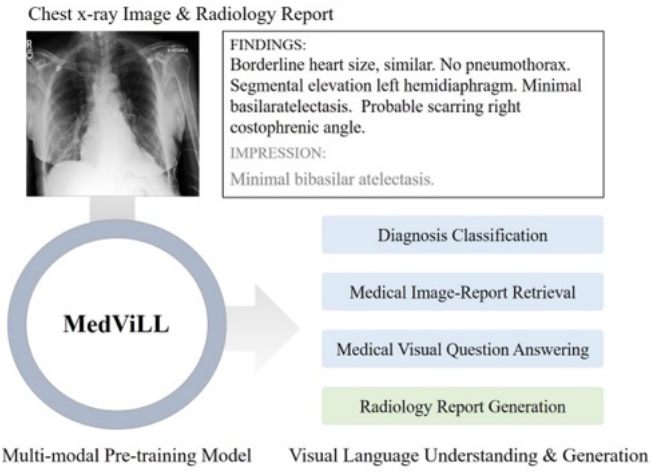


Fig. 1: **Overview of MedViLL.** During the pre-training, MedViLL learns joint representation, then fine-tuned for VLU and VLG tasks.

However, despite significant improvements reported for a wide range of downstream tasks utilizing pre-trained models, most previous studies focused on either the VLU tasks or the VLG tasks [14], [20], [26], [36], [39], suggesting the challenging nature of learning meaningful representations for both VLU and VLG at the same time. Some recent studies tried to tackle both tasks at the same time by proposing a hybrid model using the encoder and decoder of the Transformer [37], [38], [45], or with a unified BERT model by sharing knowledge using different types of the self-attention mask [47]. These works demonstrated promising results even when a single BERT-based architecture was trained to aim at both tasks.

While VL multi-modal pre-training has no doubt seen significant progress in recent years, it was mainly developed under the context of general domain (e.g., using MS-COCO). Vision and language, however, is one of the most frequently used information in the medical domain as well, often produced in the form of radiology images and corresponding free-text report. VL multi-modal pre-training therefore has great potential to be

First authors*: Jong Hak Moon, and Hyungyung Lee are equally contributed. Corresponding author: Edward Choi.

Jong Hak Moon, Hyungyung Lee, Woncheol Shin, and Edward Choi are with EdLab, Graduate School of AI, KAIST, Daejeon, South Korea (e-mail: jhak.moon@kaist.ac.kr; ttumyche@kaist.ac.kr; swc1905@kaist.ac.kr; edwardchoi@kaist.ac.kr).

Young-Hak Kim is with Department of Cardiology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, South Korea(e-mail: mdyhkim@amc.seoul.kr).

widely used in healthcare such improving diagnosis accuracy, automatically generating reports, or answering questions from physicians. Despite its huge potential, VL multi-modal pre-training in the medical domain has only recently received attention, where Li et al. [21] only demonstrated improved diagnosis accuracy of VL pre-trained models. In order to truly understand, however, whether a model has effectively learned both vision and text representation, it must be evaluated on diverse VLU and VLG tasks beyond simple diagnosis classification. This motivates us to investigate whether an integrated model is possible for a wide range of VLU and VLG tasks.

In this paper, we aim to develop a model that can learn multipurpose joint representations of vision and language in the medical domain (Fig.1). The more immediate focus of our approach to enhance downstream tasks such as diagnosis and treatment delivery is case-based reasoning, discovering underlying patterns in data, and generating semantically accurate disease profiles. The main contributions of this paper can be summarized as follows:

1) We propose Medical Vision Language Learner (MedViLL), a multi-modal pre-training model for medical images and reports with a novel self-attention scheme.
2) We demonstrate the effectiveness of our approach with detailed ablation study on extensive vision-language understanding and generation-based downstream tasks, including diagnosis classification, medical image-report retrieval, medical visual question answering, and radiology report generation.
3) We demonstrate the generalization ability of our approach under the transfer learning setting using two separate Chest X-ray datasets, where we pre-train a model on one dataset and perform diverse downstream tasks on another.

To the best of our knowledge, this is the first study that conducts both VLU and VLG tasks with a unified VL pre-training model in the medical domain. We expect that our pretrained VL model will enable more effective cross-task knowledge sharing, and reduce the development costs by eliminating the need for separate models for different tasks.

II. Related Work

A. Radiology Practices

In radiology practice, physicians identify various clinical findings based on radiographic images and the patient's clinical history, then summarize these findings and overall impressions in a clinical report [17], [32]. To accelerate the diagnostic process, [33], [34] enhance perceptual quality of noisy radiographic images. Diagnostic observations are described as positive, negative, or uncertain about the clinical findings, including the detailed location and severity of the findings. Such clinical reports are currently being used as a standard method to communicate in the clinical setting. A combination of vision and language data helps further improve the model performance in both image annotation and automatic report generation [22].

B. VL Multimodal Researches in the Medical Domain

Although various models have been gradually developed for language modeling [2], [19], [25], [35], CNN-RNN based models still dominate in VL multi-modal learning in the medical domain, and these models were mainly designed for a task-specific method of either VLU or VLG tasks. TieNet [42] is a pioneering CNN-RNN model with image-report attention mechanism for VLU (e.g., diagnosis classification) and VLG (e.g., report generation) tasks by using ChestX-ray14 [41] dataset. Liu et al. [24] only focus on the VLG task to generate the radiology report utilizing a CNN-RNN-RNN architecture with a hierarchical generation strategy from the MIMIC-CXR [16] and Open-I dataset [9]. Hsu et al. [12] focuses on a VLU task, specifically image-report retrieval in the MIMIC-CXR dataset, based on supervised and unsupervised methods. The most recent studies [4], [23], [43], [46] focus on either VLU or VLG task. Li et al. [21] compares 4 different BERT-based pre-training models on a VLU task, specifically classifying thoracic findings in the MIMIC-CXR and Open-I dataset. With a focus on VLG task [4], [23], [43], [46], EMIXER [4] is a GAN-based approach that simultaneously generates a pair of X-ray images and corresponding reports based on diagnosis labels. Other recent VLG studies [4], [23], [43], [46] focus on report generation that is as close to the ground truth as possible utilizing both the frontal and lateral view images to generate a single corresponding report. Liu et al. [23] proposed a Transformer encoder-decoder based prior and posterior knowledge distilling approaches using 3 different modalities (Vision, Language, and Knowledge Graph). Wang et al. [43] proposed a self-boosting framework with 3 different modules (CNN-based object detector as image encoder, Sentence-BERT as report encoder, and additional LSTM layers as decoder) using image and report based on the cooperation of the main tasks of generation and an auxiliary task of image-report matching. Yang et al. [46] proposed MedWriter that incorporates a hierarchical retrieval mechanism to automatically extract both report and sentence-level templates. In this paper, we focus on learning a joint representation of a single image (frontal view) and its corresponding report to perform both VLU and VLG tasks with fine-tuning.

C. VL Multimodal Researches in the General Domain

For better understanding of VL multimodality, many works have been proposed recently [7], [14], [26], [38], [39], [47] in the general domain. Among numerous variants of VL pre-training setup, we focus on three components that are most relevant to our approach: input embedding stream, visual feature embedding, and downstream tasks.

1) Input Embedding Stream: Existing models can be divided into two groups based on their architecture as a single- [7], [14], [20], [38] or two-stream [26], [39] with the marginal difference in downstream task performance [5], [28]. However, the two-stream architecture has a greater number of parameters, whereas the single stream architecture allows early interaction between two modalities by sharing parameters and processing stacks [7], [14], [36], [47]. For architectural simplicity and

time/space efficiency, we design our model with a single-stream architecture.

2) Visual Feature Embedding: For visual feature embedding, most of the recent works [7], [20], [26] are inspired by [3] utilizing pre-trained object detectors [30] to extract the region-based visual inputs. However, the representation capability of this approach is limited by the given categories of the object detection task, leading to information gaps for language understanding [14]. In contrast to the region-based visual embedding, PixelBERT [14] suggests CNN-based visual encoder with random pixel sampling to improve the robustness of visual feature learning and avoid over-fitting [11]. Since there is no applicable off-the-shelf object detector model to extract region-based feature in the medical domain [8], [22], [29], we adopt the CNN-based visual feature embedding.

3) Downstream tasks: VLU and VLG tasks are typical downstream tasks of the VL pre-trained model for tackling more complex tasks that combine vision with language. In this regard, a number of previous works [7], [14], [26], [39] use BERT-based vision-language joint encoder to perform VLU tasks. On the other hand, VLG tasks typically require an encoder for embedding the vision features and a decoder that generates text [38]. Unified VLP [47] conducts these two disparate tasks (VLU and VLG) with a single BERT-based architecture by repeatedly alternating the mask type with a fixed ratio between bidirectional and sequence-to-sequence mask during pre-training. Inspired by this unified pre-training approach, we explore different types of masks and their effects on diverse VLU and VLG downstream tasks.

III. Materials and Methods

A. Dataset

We used publicly available MIMIC-CXR [16] and Open-I [9] datasets. MIMIC-CXR [16] contains 377,110 Chest X-ray images and corresponding free-text reports. Also, Open-I dataset contains 3,851 reports and 7,466 Chest X-ray images. Since the dataset contains frontal and lateral view images, it is required to distinguish between view positions [6], [24] to avoid miss-match findings between an image and a report pair. Therefore, given the dominance of the anteroposterior (AP) frontal view in ICU (Intensive Care Units) settings (e.g., 38.89% of all studies containing at least one AP view image), we perform all experiments on unique 91,685 AP view image and associated report pairs following the official split of MIMIC-CXR (train 89,395, valid 759, test 1,531) and 3,547 image-report pairs from the official Open-I dataset. We use Open-I to test the generalization ability of the models, where all models are pre-trained on MIMIC-CXR, then fine-tuned for downstream tasks on a completely unseen Open-I dataset. Our pre-processing procedure is illustrated in Fig. 2.

B. VL Pre-training Model

Our proposed architecture MedViLL is a single BERT-based model that learn unified contextualized vision-language representation. The overall architecture of MedViLL is illustrated in Fig.3.

TABLE I: Notation explanation appearing in this section

Notation	Description
v	Chest x-ray image
$\mathbf{v}$	Visual feature
$\mathbf{l}$	Visual location feature
$\mathbf{s}_V$	Visual semantic embedding
w	Clinical report
$\mathbf{w}$	Language feature
$\mathbf{p}$	Language position embedding
$\mathbf{s}_L$	Language semantic embedding
$\tilde{\mathbf{v}}$	Final visual feature embedding
$\tilde{\mathbf{w}}$	Final language feature embedding
$\tilde{\mathbf{H}}$	Joint embedding
$\bar{\mathbf{H}}$	Contextualized embedding

1) Visual Feature Embedding: We use a CNN to extract visual features from the medical image. The visual features are obtained from the last convolution layer, then flattened along the spatial dimension. Further, we encode the absolute positions of visual input as additional information for explicitly injecting the same body position information in the x-ray images. Given a Chest x-ray image v, we denote the flattened visual feature obtained from the last CNN layer $\mathbf{v}$, and the location feature $\mathbf{l}$ as follows:

$$\mathbf{v} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_K\}, \quad \mathbf{v}_i \in \mathbb{R}^c \tag{1}$$

$$\mathbf{l} = \{\mathbf{l}_1, \mathbf{l}_2, \ldots, \mathbf{l}_K\}, \quad \mathbf{l}_i \in \mathbb{R}^c \tag{2}$$

where K indicates the number of visual features (*i.e.*, height $\times$ width) and c the hidden dimension size (*i.e.*, channel size). The final visual feature embeddings $\tilde{\mathbf{v}} = \{\tilde{\mathbf{v}}_1, \tilde{\mathbf{v}}_2, \ldots, \tilde{\mathbf{v}}_K\}$ are computed as follows:

$$\tilde{\mathbf{v}}_i = \mathbf{v}_i + \mathbf{l}_i + \mathbf{s}_V \tag{3}$$

where $\mathbf{s}_V$ is a semantic embedding vector shared by all visual feature to differentiate themselves from language embeddings. The final visual features $\tilde{\mathbf{v}}$ are fed into a fully-connected layer, to be projected into the same embedding space $\mathbb{R}^d$ has the language embeddings. During pre-training, we randomly sample a subset of the final visual features to avoid overfitting and enhance the semantic knowledge learning of visual input [15]. We use k to denote the number of sampled visual features, whereas K denotes the number of all visual features.

2) Language Feature Embedding: For language feature embedding, we follow BERT [10] to encode the textual information. A given clinical report w is first split into a sequence of N tokens (i.e. subwords) $\{w_1, ..., w_N\}$ using the WordPiece tokenizer [44]. The tokens are then converted to vector representations $\mathbf{w} = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_N\}, \quad \mathbf{w}_i \in \mathbb{R}^d$ via a lookup table, where d is the embedding dimension size. We denote position embeddings as $\mathbf{p} = \{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_N\}, \quad \mathbf{p}_i \in \mathbb{R}^d$. The final language feature embeddings $\tilde{\mathbf{w}} = \{\tilde{\mathbf{w}}_1, \tilde{\mathbf{w}}_2, \ldots, \tilde{\mathbf{w}}_N\}$ are obtained as follows:

$$\tilde{\mathbf{w}}_i = \mathbf{w}_i + \mathbf{p}_i + \mathbf{s}_L \tag{4}$$

where $\mathbf{s}_L$ is a semantic embedding vector shared by all language feature to differentiate themselves from visual embeddings.

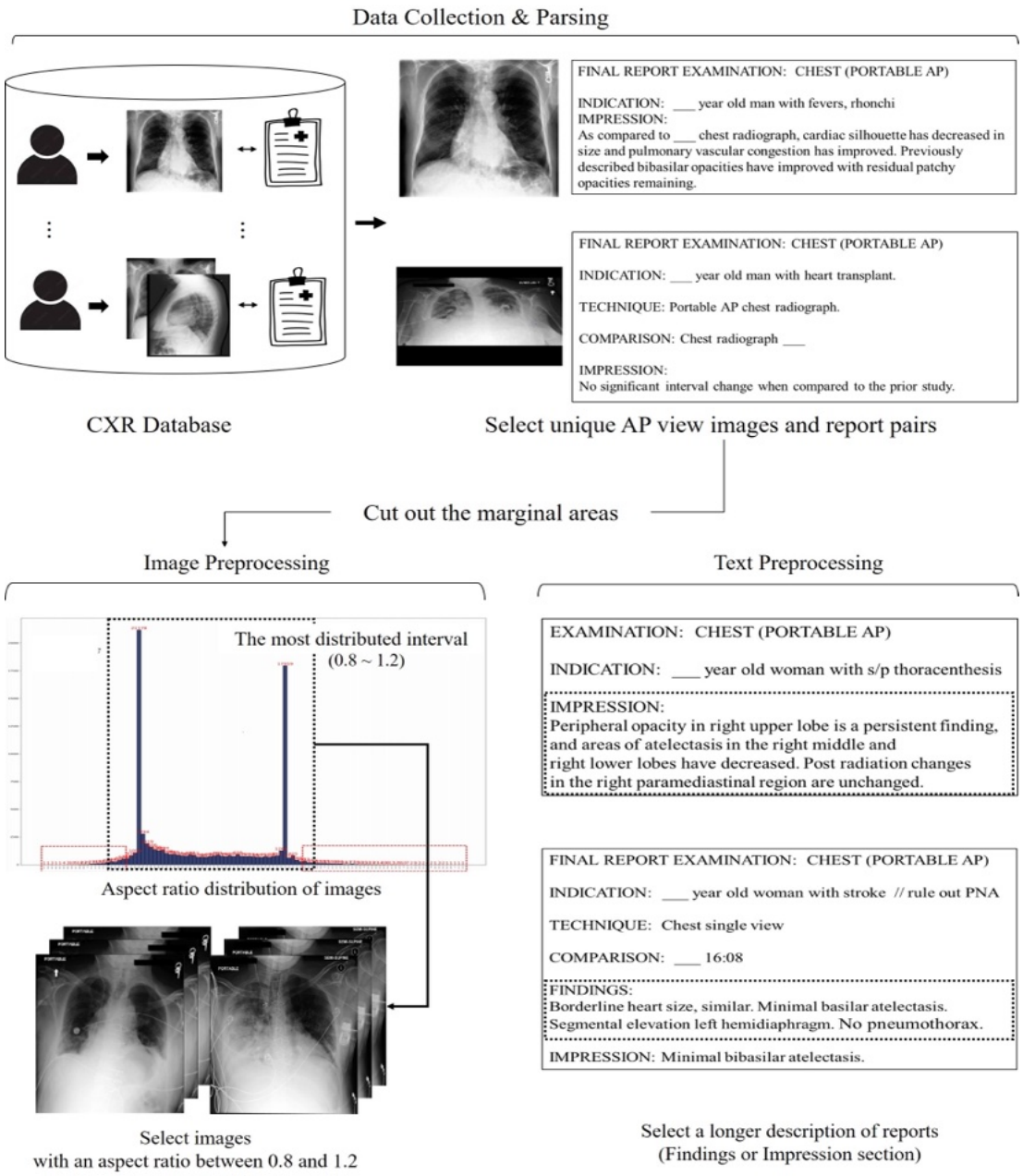


Fig. 2: **Data collection and parsing.** We pre-process the X-ray image and report data are as follows. First, for the X-ray image, we cut out the marginal space of the original image and resize all the images to 512 x 512, keeping the aspect ratio. Then for the report, we select a longer description (Findings or Impression section) which may contain detailed information associated with the X-ray imaging.

3) Joint Embedding: After obtaining visual embedding $\tilde{\mathbf{v}} \in \mathbb{R}^d$ and language embeddings $\tilde{\mathbf{w}} \in \mathbb{R}^d$, we concatenate them to construct the input sequence to the joint embedding component (Fig.3 (C)). Using additional special tokens CLS and SEP, we define the input to the joint embedding block as $\tilde{\mathbf{H}} = \{\overline{CLS}, \tilde{\mathbf{v}}_1, \ldots, \tilde{\mathbf{v}}_K, \overline{SEP}_V, \tilde{\mathbf{w}}_1, \ldots, \tilde{\mathbf{w}}_N, \overline{SEP}_L\} \in \mathbb{R}^{d \times S}$ where $S = N + K + 3$. Note that $\overline{CLS}, \overline{SEP}_V$ and $\overline{SEP}_L$ are obtained by summing the special tokens with corresponding position and semantic embeddings as in Fig.3. The contextualized embedding produced by the joint embedding block are denoted as $\bar{\mathbf{H}} = \{\overline{CLS}, \bar{\mathbf{v}}_1, \ldots, \bar{\mathbf{v}}_K, \overline{SEP}_V, \bar{\mathbf{w}}_1, \ldots, \bar{\mathbf{w}}_N, \overline{SEP}_L\}$.

4) Pre-training Objectives: To pre-train MedViLL and align visual features with language features, we take the Masked Language Modeling (MLM) and Image Report Matching (IRM) tasks, which were used in various forms in previous work [7], [14], [20], [36]. For the MLM task, we follow BERT to replace 15% of the input text tokens $\{w_1, \ldots, w_N\}$ with the special MASK token, a random token, or the original token with a probability of 80%, 10% and 10% respectively. The model is trained to recover these masked tokens based on the contextual observation of their surrounding language tokens and the visual tokens, by minimizing the following negative log-likelihood.

$$L_{MLM}(\theta) = -\mathbb{E}_{(v,w)\sim D}\Big[\log P_\theta(w_m|v, w_{\setminus m})\Big] \quad (5)$$

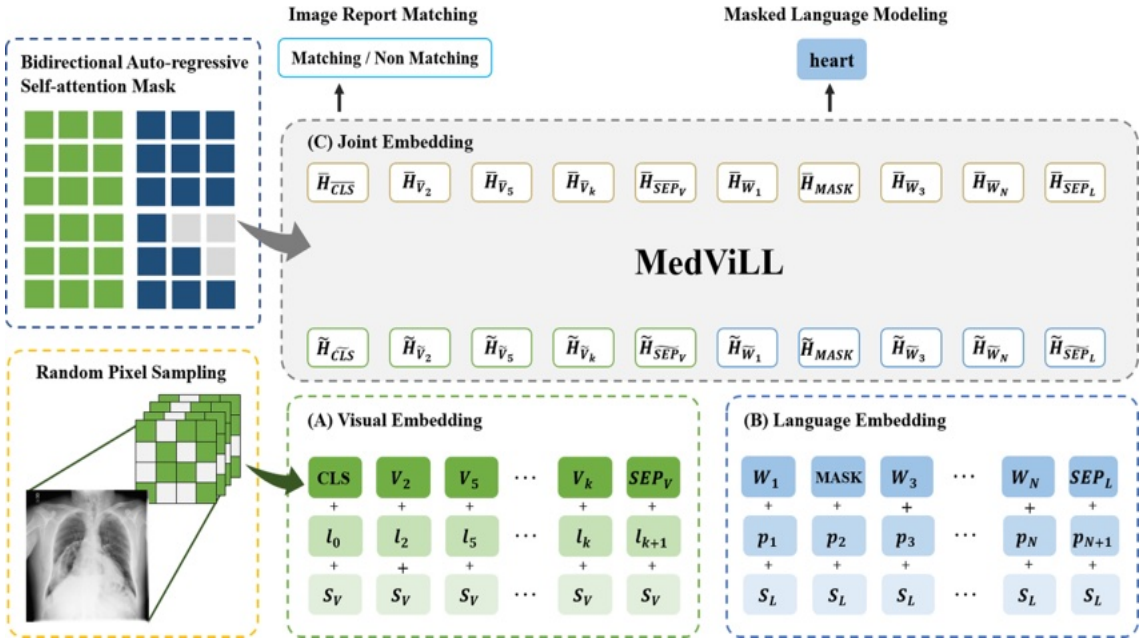


Fig. 3: **Architecture of the MedViLL.** MedViLL is a single stream BERT model for the cross-modal embedding. Chest X-ray images are randomly sampled from the last feature map of the CNN model as visual inputs. Also, each report is parsed with the BERT tokenizer to get language input. MedViLL is pre-trained with masked language modeling and image report matching tasks, and flexibly applied to VLU and VLG downstream tasks.

where θ is the trainable parameters of MedViLL. A pair of images and its corresponding report (v, w) is sampled from the training set D, where w can be divided into the masked tokens w_m and their complements $w_{\backslash m}$. $\mathbb{E}_{(v,w)\sim D}$ is the average for the training set D, and $P_\theta(w_m|v, w_{\backslash m})$ is the probability of w_m given v and $w_{\backslash m}$. IRM task encourages the model to learn both visual and textual features by training the model to predict whether a given pair of image and report (v, w) is a matching pair or not. During pre-training, we randomly sample both matching image-report pairs and non-matching image-report pairs with 1:1 ratio from the dataset. Note that, however, while selecting a matching pair is straightforward (X-ray images come with a corresponding report), sampling a non-matching pair is not, because two different report can be semantically the same (*e.g.*, "No findings.", "Nothing noticeable."). Therefore, when sampling for non-matching image-report pairs, we use diagnosis labels. Specifically, in our IRM task, a non-matching report is defined as the ones that are extracted different positive diagnosis labels than the matching report. The joint contextualized embedding $\overline{CLS}$ is used to classify whether the input image and report are a matching pair or not, with the following loss function,

$$L_{IRM}(\theta) = -\mathbb{E}_{(v,w)\sim D}\Big[y \log P_\theta(v, w)\Big] - \mathbb{E}_{(v,w')\sim D}\Big[(1-y)\log(1-P_\theta(v, w'))\Big] \quad (6)$$

where (v, w) denotes a matching image-report pair, (v, w') a non-matching pair, y is the label (1 for matching, and 0 for unmatching), $\mathbb{E}_{(v,w)\sim D}$ is the average for the training set D, and $P_\theta(v, w)$ is the probability of the (v, w) being paired.

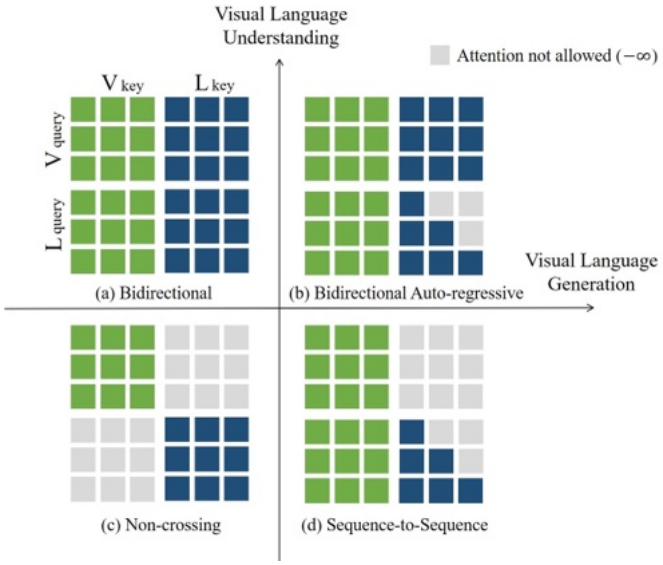


Fig. 4: **Self-attention mask schemes.** Four types of self-attention masks and the quadrant for the difference in performance in the downstream task of each attention mask. (a) Bidirectional, (b) Bidirectional Auto-regressive, (c) Non-crossing, (d) Sequence-to-Sequence self-attention masks.

C. Self-Attention Mask Schemes

We explore several types of self-attention masks to encourage the model to learn universal multi-modal representations. Bi (Bidirectional) attention mask (Fig. 4 (a)) that allows all inputs to interact freely for unconstrained context learning between the visual-language modalities. S2S (Sequence-to-Sequence) causal attention mask (Fig. 4 (d)), on the other hand, allows restricted context learning; language features are only allowed to attend to previous words, while visual features are not allowed to attend to any language features, in order to prevent leaking

information from the future. Bi & S2S uses both Bi and S2S masks alternately during pre-training (in every mini-batch, use S2S with 75% chance and Bi with 25% chance) to perform both VLU and VLG downstream tasks. In this work, we propose a new self-attention mask, Bidirectional Auto-Regressive (BAR) (Fig. 4 (b)), to closes the gap between Bi and S2S while taking advantage of both. BAR allows image features to be mixed with language features during pre-training (as opposed to S2S mask), while preserving the causal nature of auto-regressive language generation. The self-attention mask $M \in R^{S\times S}$, $S = N+K+3$ consists of 0s and negative infinities as below.

$$M_{jk} = \begin{cases} 0, & \text{(attention allowed)} \\ -\infty, & \text{(attention not allowed)} \end{cases} \qquad j,k = 1,...,S. \tag{7}$$

And a single attention head in the self-attention module can be formulated as follows:

$$Attention = \text{softmax}\left(SA + M\right)V, \quad SA = \frac{QK^T}{\sqrt{d_k}} \tag{8}$$

where Q, K, V, and d_k indicate queries, keys, values, and dimension of queries and keys respectively [40].

$$SA = \begin{bmatrix} CLS_q{\cdot}CLS_k & \cdots & CLS_q{\cdot}W_{1k} & \cdots & CLS_q{\cdot}SEP_{Lk} \\ V_{1q}{\cdot}CLS_k & \cdots & V_{1q}{\cdot}W_{1k} & \cdots & V_{1q}{\cdot}SEP_{Lk} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ SEP_{Vq}{\cdot}CLS_k & \cdots & SEP_{Vq}{\cdot}W_{1k} & \cdots & SEP_{Vq}{\cdot}SEP_{Lk} \\ W_{1q}{\cdot}CLS_k & \cdots & W_{1q}{\cdot}W_{1k} & \cdots & W_{1q}{\cdot}SEP_{Lk} \\ W_{2q}{\cdot}CLS_k & \cdots & W_{2q}{\cdot}W_{1k} & \cdots & W_{2q}{\cdot}SEP_{Lk} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ W_{Kq}{\cdot}CLS_k & \cdots & W_{Kq}{\cdot}W_{Lk} & \cdots & W_{Kq}{\cdot}SEP_{Lk} \\ SEP_{Lq}{\cdot}CLS_k & \cdots & SEP_{Lq}{\cdot}W_{Lk} & \cdots & SEP_{Lq}{\cdot}SEP_{Lk} \end{bmatrix} \tag{9}$$

where q and k indicate query and key vectors respectively. Since the self-attention matrix is computed from the query and key vectors of vision-language modalities according to the Eq. (9), the computed self-attention matrix can be divided into 4 subparts of queries and key combinations by modality type.

$$SA_{q,k} = SA_{CLS_q:SEP_{Vq},CLS_k:SEP_{Vk}} \tag{10}$$
$$+ SA_{CLS_q:SEP_{Vq},W_{1k}:SEP_{Lk}} \tag{11}$$
$$+ SA_{W_{1q}:SEP_{Lq},CLS_k:SEP_{Vk}} \tag{12}$$
$$+ SA_{W_{1q}:SEP_{Lq},W_{1k}:SEP_{Lk}} \tag{13}$$

where Eq. (10) is the attention of query and key from vision, Eq. (11) is an attention mask of query from the vision and key from language, Eq. (12) is an attention mask of query from the language and key from vision, and Eq. (13) is an attention mask of query and key from language features. We combine the attention mask matrix M for the subparts of SA because adding negative infinity to the calculated attention value will result in zero in the softmax operation.

$$BAR_M = \begin{bmatrix} 0 & \cdots & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 \\ 0 & \cdots & -\infty & \cdots & -\infty \\ 0 & \cdots & 0 & \cdots & -\infty \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & -\infty \\ 0 & \cdots & 0 & \cdots & 0 \end{bmatrix} \in \mathbb{R}^{S\times S} \tag{14}$$

Therefore, BAR attention mask (Eq. (14)) allows the attention calculations of all possible combinations except for the Eq. (13). Intuitively, this self-attention mask scheme applies auto-regressive attention masks to language modality to enhance joint embedding between vision and language modalities and perform well in both generation and understanding tasks. We implemented four different models each using different types of self-attention masks during pre-training; Bi, S2S, BAR and Bi & S2S. In addition, we also experiment with Non-crossing attention mask (Fig. 4 (c)) as a baseline to investigate the impact of multi-modal representation learning. As non-crossing attention mask restricts the interaction between two modalities, we add one additional CLS token at the beginning of the language features, so that both $\overline{CLS}_V$ and $\overline{CLS}_L$ can be used for the IRM pre-training task. Three types of self-attention mask matrices (Bi, S2S and Non-crossing) are described in the appendix.

IV. RESULTS AND DISCUSSION

A. Dataset Analysis

Although both MIMIC-CXR and Open-I consist of chest X-ray images and report pairs, the two datasets could have different characteristics since they were collected from separate institutions. Specifically, the diagnostic information represented by the two X-ray image sets could be differently distributed. Therefore, to analyzes the difference in the distribution of diagnostic labels between two datasets, we compared positive labels acquired from the Chexpert labeler results. As seen in Fig. 5, a mild imbalance was observed in MIMIC-CXR where the class ratios ranged from 13.39% (support devices) to 1.2% (pneumonia, and pleural other). On the other hand, a severe imbalance was observed in Open-I compared to MIMIC-CXR with the maximum class ratios of 28.8% (Others, and cardiomegaly) and the minimum of 1.07% (support devices). This shows that Open-I not only differs from MIMIC-CXR in terms of data volume, but also in terms of clinical properties. Therefore, we believe evaluating the MIMIC-CXR-pre-trained models on Open-I is an appropriate setup to test the generalization capability of the models. The distribution of diagnosis label is illustrated in Fig. 5.

B. Implementation details

We use ResNet-50 pre-trained on ImageNet as a visual feature extractor. The input image size is (512x512x3), and the last feature map (16x16x2048) of ResNet-50 is flattened by spatial dimensions and we randomly sample 180 visual features (180x2048) during pre-training, while we use all features (256x2048) for every downstream task. To embed text token, each sequence from reports is truncated or padded to 253 tokens in length by considering maximum embedding size. For the joint embedding, we adopt BERT-base architecture which comprised of 12 Transformer layers. Each layer contains 12 attention heads, 768 embedded hidden size and 0.1 drop-out probability. We adopt AdamW optimizer with learning rate $1e^{-5}$ settings for visual backbone and Transformer. All models were trained on 8 RTX-3090 GPU with the batch size of 128 and 50 epochs for the pre-training model.

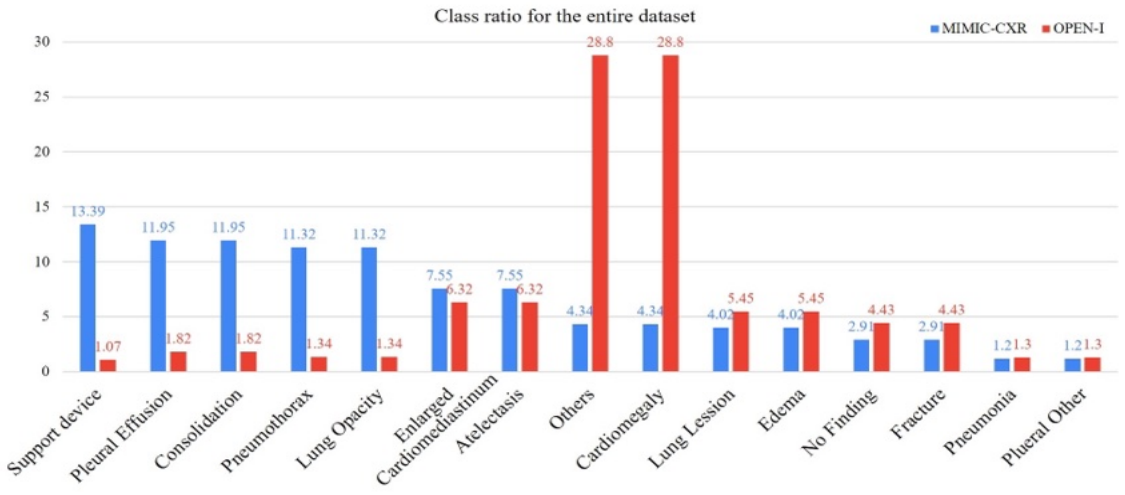


Fig. 5: **Dataset Analysis.** We compare the distribution of diagnosis labels over the entire dataset. Due to the different scales of the two datasets, each label was represented as a percentage over the entire dataset.

TABLE II: Model AUROC and F1 scores for the diagnosis classification task on MIMIC-CXR and Open-I. Inference time(ms) on MIMIC-CXR: MedViLL(12.5), Bi&S2S (13), Bi (13), S2S (13), Non-crossing (12.5), Fine-tuning Only (15.5), CNN & Transformer (10.5).

Dataset	Metrics	MedViLL	Bi&S2S	Bi	S2S	Non-crossing	Fine-tuning Only	CNN & Transformer
MIMIC-CXR	avg AUROC	0.980 (0.00)	0.979 (0.00)	**0.984 (0.00)**	0.982 (0.00)	0.980 (0.00)	0.969 (0.00)	0.831 (0.00)
	avg F1	0.839 (0.00)	0.846 (0.00)	**0.852 (0.00)**	0.846 (0.00)	0.824 (0.00)	0.807 (0.00)	0.491 (0.00)
	p-value (avg AUROC)	-	0.005	1.97E-15	0.003	0.254	1.70E-36	3.41E-102
	p-value (avg F1)	-	9.59E-28	7.85E-42	1.62E-26	2.02E-43	4.90E-63	2.70E-122
Open-I	avg AUROC	**0.892 (0.00)**	0.827 (0.00)	0.758 (0.00)	0.720 (0.00)	0.589 (0.00)	0.723 (0.00)	0.709 (0.00)
	avg F1	**0.407 (0.01)**	0.301 (0.01)	0.295 (0.01)	0.256 (0.01)	0.185 (0.00)	0.300(0.00)	0.245 (0.01)
	p-value (avg AUROC)	-	6.94E-83	4.00E-98	1.23E-101	4.66E-122	1.41E-103	1.35E-109
	p-value (avg F1)	-	1.05E-93	7.04E-95	7.17E-101	2.08E-110	6.49E-94	2.69E-104

TABLE III: Medical Image-Report Retrieval performance on MIMIC-CXR and Open-I. Inference time(ms) of Report-to-Image and Image-to-Report on MIMIC-CXR: MedViLL(7.6, 7.8), Bi&S2S (8.2, 7.8), Bi (7.6, 7.6), S2S (7.6, 7.7), Non-crossing (7.7, 7.8), Fine-tuning Only (7.8, 7.6), CNN & Transformer (5.3, 5.3).

Task	Models	MIMIC-CXR					OpenI				
		MRR	H@5	R@5	P@5	p-value	MRR	H@5	R@5	P@5	p-value
Report-to-Image	MedViLL	56.5(0.01)	77.0(0.01)	47.4(0.01)	19.9(0.00)	-	51.3(0.01)	73.0(0.01)	12.9(0.00)	31.7(0.00)	-
	Bi&S2S	55.5(0.01)	76.7(0.01)	46.7(0.01)	19.7(0.00)	1.20E-05	46.4(0.01)	68.1(0.01)	10.5(0.00)	28.8(0.01)	3.71E-27
	Bi	58.0(0.01)	78.2(0.01)	48.2(0.01)	20.2(0.00)	1.60E-10	**51.4(0.01)**	**74.8(0.01)**	**13.3(0.00)**	32.0(0.01)	0.843
	S2S	**58.8(0.01)**	**79.1(0.01)**	**48.9(0.01)**	**20.3(0.00)**	1.89E-18	48.6(0.01)	67.2(0.01)	10.3(0.01)	**32.9(0.01)**	2.28E-14
	Non-crossing	54.7(0.01)	77.0(0.01)	47.2(0.01)	19.5(0.00)	4.07E-12	48.6(0.01)	68.4(0.01)	11.2(0.00)	31.1(0.01)	3.88E-18
	Fine-tuning Only	41.8(0.01)	61.6(0.01)	35.8(0.01)	15.8(0.00)	3.14E-53	36.9(0.01)	54.4(0.01)	5.4(0.00)	20.7(0.01)	1.72E-53
	CNN & Transformer	11.4(0.01)	15.2(0.02)	5.1(0.00)	3.6(0.01)	9.43E-71	36.2(0.04)	56.6(0.04)	5.0(0.00)	21.4(0.04)	4.94E-19
Image-to-Report	MedViLL	55.8 (0.01)	75.5(0.01)	47.1(0.01)	19.7(0.00)	-	**50.4(0.01)**	63.8(0.01)	12.9(0.00)	35.5(0.01)	-
	Bi&S2S	54.5(0.01)	75.5(0.01)	47.8(0.01)	19.9(0.00)	6.32E-08	45.8(0.01)	54.0(0.01)	10.1(0.00)	35.8(0.00)	8.55E-29
	Bi	56.7(0.01)	76.3(0.01)	47.6(0.01)	20.2(0.00)	0.0002	48.5(0.01)	**65.8(0.01)**	**13.7(0.00)**	32.3(0.01)	3.17E-12
	S2S	**57.9(0.01)**	**78.5(0.01)**	**49.7(0.01)**	**20.7(0.00)**	2.72E-13	45.4(0.01)	53.6(0.01)	8.9(0.00)	**36.9(0.00)**	6.84E-31
	Non-crossing	54.6(0.01)	75.7(0.01)	47.6(0.01)	20.0(0.00)	3.84E-07	42.6(0.01)	61.2(0.01)	11.0(0.00)	28.0(0.01)	1.15E-40
	Fine-tuning Only	41.4(0.01)	60.8(0.01)	36.3(0.01)	15.7(0.00)	5.56E-56	45.2(0.00)	49.7(0.01)	5.1(0.00)	35.0(0.00)	2.64E-29
	CNN & Transformer	12.0(0.02)	15.3(0.02)	5.1(0.00)	4.0(0.01)	1.09E-52	37.9(0.06)	54.0(0.06)	5.0(0.00)	23.0(0.06)	1.16E-12

C. Task-specific Downstream Model Strategy

1) Diagnosis Classification: For a given image-report pair, we use the positive labels extracted from the report by the Chexpert labeler as the diagnosis labels. As a single pair could have multiple diagnosis labels up to the maximum of 14 (*i.e.* multi-label classification), we use 14 linear heads on top of $\overline{CLS}$ and fine-tune the model using the binary cross-entropy loss. All models are evaluated with the micro average AUROC, and micro average F1 score.

2) Medical Image-Report Retrieval: There are two subtasks for medical image-report retrieval, where image-to-report (I2R) retrieval requires the model to retrieve the most relevant report from a large pool of reports given an image, and vice versa for report-to-image (R2I) retrieval. Given an image, any report that contains the same Chexpert diagnosis labels as the original matching report is considered a positive image-report pair, and a negative pair otherwise. The final multi-modal representation $\overline{CLS}$ is used as the input to a binary classifier to classify the

TABLE IV: Model accuracy on the VQA-RAD dataset. O.E. stands for Open-ended question and C.E. stands for close-ended question. For MEVF [27], we used the reported results from the original paper. Inference time(ms) on MIMIC-CXR: MedViLL(19.46), Bi&S2S(19.52), Bi(19.43), S2S(19.51), Non-crossing(19.58), Fine-tuning Only(19.61), CNN & Transformer(17.42).

Models	ALL				CHEST			
	O.E.	C.E.	p-value of O.E.	p-value of C.E.	O.E.	C.E.	p-value of O.E.	p-value of C.E.
MedViLL	**0.595(0.032)**	0.777(0.071)	-	-	**0.587(0.033)**	**0.782(0.123)**	-	-
Bi&S2S	0.541(0.038)	0.76(0.027)	2.93E-07	0.224	0.566(0.074)	0.766(0.035)	0.164	0.519
Bi	0.58(0.038)	**0.784(0.03)**	0.124	0.643	0.562(0.04)	0.767(0.035)	0.013	0.549
S2S	0.505(0.042)	0.73(0.025)	1.81E-12	0.002	0.517(0.07)	0.723(0.048)	1.57E-05	0.021
Non-crossing	0.531(0.015)	0.734(0.017)	5.58E-12	0.003	0.474(0.083)	0.732(0.03)	3.94E-08	0.043
Fine-tuning Only	0.232(0.019)	0.649(0.026)	2.98E-43	5.38E-11	0.124(0.014)	0.606(0.035)	1.08E-42	1.50E-08
CNN & Transformer	0.24(0.029)	0.667(0.015)	7.66E-46	2.70E-9	0.124(0.067)	0.523(0.033)	7.60E-32	1.62E-12
MEVF [27]	0.407	0.741	-	-	-	-	-	-

TABLE V: Report generation performance in terms of Perplexity and Label Accuracy, Precision Recall and F1 and BLEU4. Inference time(ms) on MIMIC-CXR: MedViLL(32.81), Bi&S2S(33.55), Bi(32.54), S2S(33.04), Non-crossing(32.71), Fine-tuning Only(32.92).

Dataset	Models	Perplexity (↓)	Accuracy (↑)	Precision (↑)	Recall (↑)	F1 Score (↑)	BLEU4 (↑)	p-value
MIMIC	MedViLL	4.185(0.022)	**0.841(0.003)**	**0.698(0.002)**	**0.559(0.004)**	**0.621(0.002)**	0.066(0.001)	-
	Bi&S2S	6.515(0.12)	0.786(0.007)	0.619(0.003)	0.435(0.009)	0.511(0.006)	0.066(0.001)	4.17E-43
	Bi	849.67(5.225)	0.637(0.004)	0.283(0.007)	0.07(0.024)	0.11(0.032)	0.015(0.004)	1.11E-36
	S2S	4.258(0.069)	0.797(0.007)	0.662(0.004)	0.448(0.01)	0.534(0.007)	0.043(0.001)	2.32E-38
	Non-crossing	718.122(9.484)	0.634(0.005)	0.277(0.013)	0.076(0.004)	0.12(0.005)	0.007(0.001)	2.30E-75
	Fine-tuning Only	224.343(0.204)	0.664(0.003)	0.417(0.012)	0.305(0.006)	0.352(0.005)	0.009(0.004)	4.14E-65
	TieNet	**4.132(0.033)**	0.687(0.003)	0.487(0.003)	0.380(0.006)	0.426(0.006)	**0.123(0.002)**	7.17E-54
Open-I	MedViLL	5.637(0.259)	0.734(0.001)	0.512(0.002)	0.594(0.001)	0.55(0.001)	0.049(0.001)	-
	Bi&S2S	15.97(1.071)	0.712(0.003)	0.497(0.003)	0.369(0.006)	0.423(0.004)	0.024(0.01)	4.02E-52
	Bi	787.66(55.492)	0.686(0.004)	0.356(0.025)	0.103(0.006)	0.16(0.008)	0.015(0.004)	2.14E-52
	S2S	**4.732(0.537)**	**0.736(0.003)**	**0.517(0.002)**	0.538(0.004)	0.527(0.002)	0.043(0.002)	1.76E-44
	Non-crossing	217.27(12.139)	0.693(0.003)	0.337(0.025)	0.085(0.005)	0.135(0.007)	0.002(0.001)	1.46E-57
	Fine-tuning Only	292.60(19.858)	0.684(0.003)	0.291(0.023)	0.073(0.035)	0.112(0.047)	0.006(0.002)	7.02E-30
	TieNet	7.901(0.483)	0.732(0.007)	**0.517(0.013)**	**0.610(0.017)**	**0.553(0.013)**	**0.189(0.005)**	0.2181

given pair, which is trained by the binary cross-entropy loss. At inference, in each trial, a model is given 100 image-report pairs, and it must use the predicted scores to rank the positive pair as highest as possible. The evaluation metrics are Hit@K, Recall@K, Precision@K (K = 5), and mean reciprocal rank (MRR).

3) Medical Visual Question Answering: We perform VQA on the VQA-RAD dataset [1], which contains 3,515 question-answer pairs on 315 images (104 head CTs or MRIs, 107 Chest X-rays, and 104 abdominal CTs). As our models are pre-trained on Chest X-ray images, VQA-RAD provides a unique opportunity to study whether the pre-trained models would generalize well beyond the single image domain. Given a pair of an image and a free-text question, we use the final representation $\overline{CLS}$ to predict a one-hot encoded answer (all possible answers are treated as a single token). The performance was evaluated with accuracy, but separately for the closed questions (i.e. short-form answers such as yes/no) and the open-ended questions (i.e. long-form answers), following the original VQA-RAD paper.

4) Radiology Report Generation: The fine-tuning process is same as the MLM pre-training task, except that we fix the self-attention mask to S2S for all models. At inference, reports can be generated by sequentially recovering the MASK tokens; given visual features followed by a single MASK token, the model can predict the first language token. Then we can replace the first MASK with the sampled token, and a new MASK token is appended. This process is repeated until the model predicts the SEP token as the stop sign. The performance is measured with three metrics: perplexity, clinical efficacy, and BLEU score. Clinical efficacy metrics is obtained by applying the Chexpert labeler on both the original matching report and the generated report. Based on the extracted labels, we can calculate accuracy, precision, recall, and F1 accordingly. Perplexity is used to evaluate the linguistic fluency of the model, while the clinical efficacy is used to evaluate if the model can capture the semantics of the given image. We also report 4-gram BLEU score to evaluate how similar the generated report is to the reference report.

D. Downtream Task Result

MedViLL is compared to four pre-training models trained with different attention masks. We also include two more baselines: 1) Fine-tuning Only, which follows the same model architecture as MedViLL, but directly fine-tuned on each downstream task without any pre-training. 2) CNN & Transformer, which uses the CNN module for encoding image only, and the Transformer module (same size as MedViLL) for encoding report only, and the outputs from each module are used for downstream tasks. CNN & Transformer also does not use pre-training. For all tasks, we conduct 30 random multiple-bootstrap experiments and report the mean performance and its standard deviation. Also, we perform a statistical hypothesis test. Based on the average values of the various metrics obtained for each model, we conduct an independent t-test with a significance value of 0.05 to identify the significant pairwise difference of our method against multiple baseline models. In summary, MedViLL achieved the best or second-best performance by

analyzing statistical significance with various baseline models in the VLU and VLG tasks. In addition, MedViLL shows superior generalization ability by outperforming most of the models in out-of-domain evaluations.

1) Diagnosis Classification: All model performance is shown in TABLE II. For the Open-I images, we use Chexpert labeler on their MeSH annotations to extract the same set of diagnosis categories as the MIMIC-CXR dataset. Specifically, Bi and S2S outperform MedViLL with a statistically significant difference in both micro-averaged AUROC and F1 scores, indicating the null hypothesis can be rejected (t-test produced a p-value lower than 0.05). Also, although MedViLL (0.9805) achieves a higher score than Non-crossing (0.9801), it is not statistically meaningful with a p-value of 0.254 in micro-averaged AUROC. However, MedViLL outperforms all other baselines with a statistically meaningful difference in MIMIC-CXR. Moreover, MedViLL outperforms all models statistically significantly when transferred to Open-I. It is also noteworthy that Bi&S2S, which is aimed to take advantage of both the bidirectional mask and the S2S mask demonstrates much better generalization capability compared to the two individual masks.

2) Medical Image-Report Retrieval: TABLE III shows the performance of Image-to-Report and Report-to-Image retrieval. We report the p-value of MRR for the performance of both tasks since MRR is a rank-aware evaluation metrics compared to other metrics. We can observe that all pre-trained models significantly outperform naive baselines (Fine-tune Only and CNN & Transformer) for the MIMIC-CXR dataset. In Report-to-Image retrieval, while MedViLL achieves lower performance than Bi and S2S with a statistically significant difference in MIMIC-CXR, MedViLL outperforms all baselines when fine-tuned on the unseen Open-I dataset except for Bi. However, although Bi outperforms MedViLL in Open-I, there is no statistically significant difference between both models with a p-value of 0.843. In Image-to-Report retrieval, S2S statistically outperforms MedViLL in MIMIC-CXR, but MedViLL is superior to all baselines with a statistically meaningful difference in Open-I. It is notable that the naive baselines, while severely underperforming for MIMIC-CXR, show substantially increased performance for Open-I. We believe this is due to the Open-I being a significantly smaller dataset than MIMIC-CXR, with only two Chexpert labels (Others, and Cardiomegaly) mostly dominating the label space.

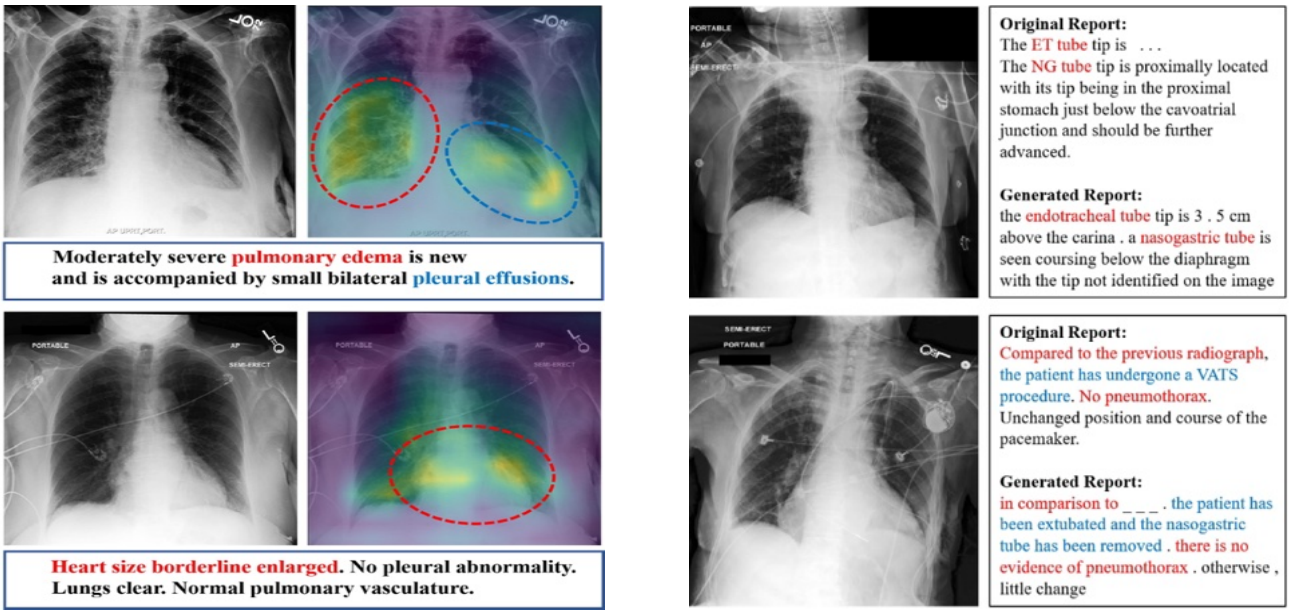
3) Medical Visual Question Answering: TABLE IV shows the VQA accuracy when models were fine-tuned with all image types ('ALL'), and with only the Chest X-ray images ('CHEST'). We can see that MedViLL significantly outperforms MEVF [27], the state-of-the-art model for the VQA-RAD dataset, indicating the effectiveness of the multi-modal pre-training for this complex multi-modal reasoning task. We can also see that MedViLL shows comparable performance as the bidirectional mask which allows unrestricted interaction between all text and vision features. In a statistical analysis, MedViLL shows significantly higher performance than all models for O.E. (open-ended questions) of 'ALL' and 'CHEST'. However, there was no statistically significant difference for Bi&S2S of 'CHEST' and Bi of 'ALL', (p-values of 0.164, and 0.124, respectively). For C.E. (close-ended questions) of 'ALL', Bi performed the best of all the models, followed by MedViLL. However, this result is not statistically meaningful, obtaining a p-value of 0.643 between Bi and MedViLL. Also, for C.E. of 'CHEST', MedViLL outperforms all baselines, but it is not statistically significant against Bi&S2S (p-value of 0.519) and Bi (p-value of 0.549).

4) Radiology Report Generation: TABLE V shows the report generation performance of all models. For this task, we also implemented TieNet [42] as a baseline, which is a widely used CNN-RNN based attention model for report generation. We can see that the models pre-trained with auto-regressive manners (MedViLL, Bi&S2S, S2S) all significantly outperform the other models in terms of both perplexity (except for TieNet) and clinical efficacy metrics for the MIMIC-CXR dataset. Among the clinical efficacy metrics, we report the p-value of F1 score because the F1 score is a balanced measure of both precision and recall and allows us to better capture the true performance here in light of the strong class imbalance of dataset. MedViLL achieved the best performance on the MIMIC dataset with a statistically significant difference. MedViLL seems to best capture the semantics embedded in the image given the highest F1 score. When fine-tuned on the unseen Open-I dataset, TieNet performed the best of all the models, followed by MedViLL. However, this result is not statistically significant between TieNet and MedViLL with a p-value of 0.2181, indicating that the performance of the two models is similar. As opposed to its generally favorable performance in VLU tasks, the bidirectional mask seems to be evidently harmful for the VLG task, most likely due to its incompatibility with the auto-regressive nature of VLG. Interestingly, we found N-gram based measures to be a suboptimal measure for report generation by showing that all models except TieNet achieved low BLEU scores; where the original report contains abbreviated terms (e.g. "ET tube") models would generate expanded terms (e.g. "endotracheal tube") and vice versa.

E. Qualitative results and analysis

1) Attention Map Visualization: We visualize the attention maps of the intermediate Transformer layers of MedViLL in order to gain qualitative insight of the cross-modality alignment between text tokens and image features as shown in Fig. 6 (a). Although MedViLL uses only report and images for training without any annotations, it can well attend to the disease-discriminatory regions written in the reports, as confirmed by a professional cardiologist. This suggests the potential of MedViLL's capability to explain its learned representations to human users for smoother real-world deployment.

2) Radiology Report Generation: We compare the generated reports with the original report. As confirmed by a professional cardiologist, Fig. 6 (b) shows that MedViLL is able to generate clinically appropriate reports. Specifically, the left image-report pair of Fig. 6 (b) shows that MedViLL generates a report with the abbreviated medical terms (i.e. "ET tube", "NG tube") expanded to the original terms (i.e. "endotracheal tube" or "nasogastric tube"). In the right image-report pair, the blue text in the original report describes the completion of VATS (i.e. video-assisted thoracic surgery). Interestingly, the generated



(a) Attention map visualization. (b) Radiology report generation analysis.

Fig. 6: **Qualitative results and analysis.** We visualize the attention regions extracted from the MedViLL (a). Also, we compare the generated report with the original report on the same chest X-ray image (b).

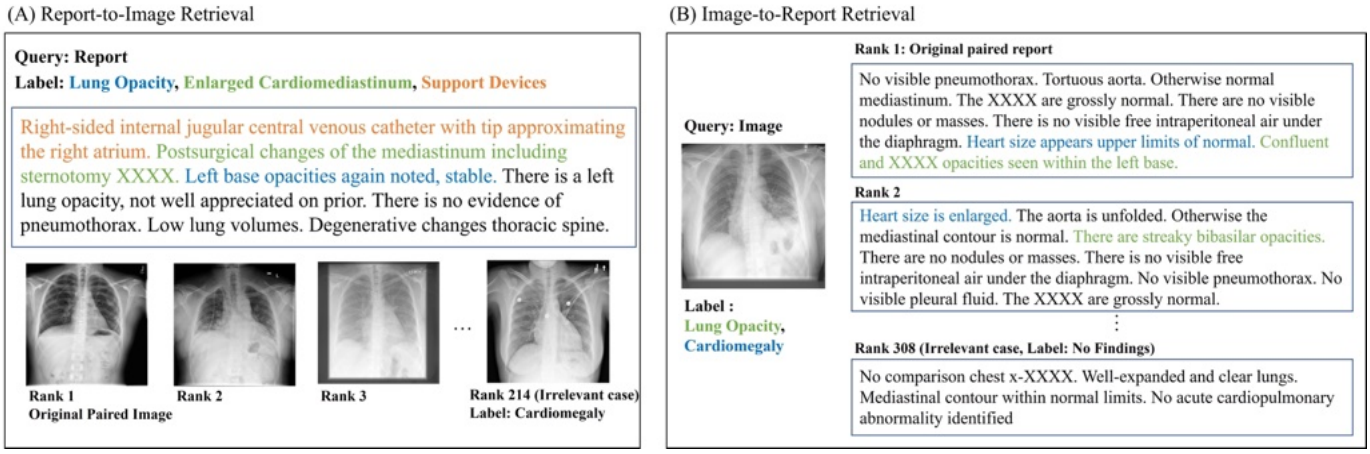


Fig. 7: **Case study of Medical Image-Report Retrieval.** Given a report or image as a query, we show the retrieved images or reports in (a) and (b) respectively. In (a), the given report is annotated by the Chexpert labeler with the diagnosis labels Lung Opacity (blue), Enlarged Cardiomediastinum (green), and Support Devices (orange). The top three images also contain the same labels as the given query while the last image (Rank 214) is irrelevant to the give query. In (B), the top three reports also contain the same labels recognized by the Chexpert labeler, Lung Opacity (green) and Cardiomegaly (blue). Note that the last report (Rank 308) does not contain any label, and is irrelevant to the given query image.

report describes extubation and nasogastric tube removal, which is a part of VATS. This indicates that the BLEU score is not an appropriate measure to evaluate report generation especially in the medical domain.

3) Medical Image-Report Retrieval: We retrieved the cases pooled from 1,536 studies in the test set (Fig. 7). As confirmed by a cardiologist, Fig. 7 demonstrate the clinical understanding of MedViLL. Specifically, we can observe that the results in the top-3 retrieved samples all share the same diagnosis labels as the given query; all top three images in Fig. 7 (a) are labeled with "Lung Opacity", "Enlarged Cardiomediastinum" and "Support Devices" as the query report, and all top three reports in Fig. 7 (b) contain the same labels "Cardiomegaly" and "Lung Opacity" as the query image. Note that the samples in the low rank contain labels irrelevant to the given query.

V. Conclusion

In this study, we propose a multi-modal pre-training model MedViLL, which uses a novel self-attention scheme to flexibly adapt to multiple downstream tasks of vision-language understanding and generation. By statistically and rigorously evaluating MedViLL on all four downstream tasks with three radiographic image-report datasets, we empirically demonstrated the superior performance of MedViLL against various baselines including task-specific architectures. Despite the impressive performance of MedViLL, this is just the beginning of the vision-language representation learning in the medical domain, and we plan to expand this approach to more diverse settings

such as multi-view Chest X-ray studies or a sequence of studies over time.

ACKNOWLEDGMENT

This work was supported by Samsung Electronics (No.IO201211-08109-01), Institute of Information & Communications Technology Planning & Evaluation (IITP) grant (No.2019-0-00075, Artificial Intelligence Graduate School Program(KAIST)), and National Research Foundation of Korea (NRF) grant (NRF-2020H1D3A2A03100945) funded by the Korea government (MSIT).

REFERENCES

[1] R. Alizadehsani, M. Roshanzamir, M. Abdar, A. Beykikhoshk, A. Khosravi, M. Panahiazar, A. Koohestani, F. Khozeimeh, S. Nahavandi, and N. Sarrafzadegan. A database for using machine learning and data mining techniques for coronary artery disease diagnosis. *Scientific data*, 6(1):1–13, 2019.

[2] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019.

[3] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.

[4] S. Biswal, P. Zhuang, A. Pyrros, N. Siddiqui, S. Koyejo, and J. Sun. Emixer: End-to-end multimodal x-ray generation via self-supervision. *arXiv preprint arXiv:2007.05597*, 2020.

[5] E. Bugliarello, R. Cotterell, N. Okazaki, and D. Elliott. Multimodal pretraining unmasked: Unifying the vision and language berts. *arXiv preprint arXiv:2011.15124*, 2020.

[6] W. W. Chapman, M. Fizman, B. E. Chapman, and P. J. Haug. A comparison of classification algorithms to automatically identify chest x-ray reports that support pneumonia. *Journal of biomedical informatics*, 34(1):4–14, 2001.

[7] Y.-C. Chen, L. Li, L. Yu, A. E. Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu. Uniter: Learning universal image-text representations. *arXiv preprint arXiv:1909.11740*, 2019.

[8] V. Cheplygina, M. de Bruijne, and J. P. Pluim. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical image analysis*, 54:280–296, 2019.

[9] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. Antani, G. R. Thoma, and C. J. McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016.

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

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[12] T.-M. H. Hsu, W.-H. Weng, W. Boag, M. McDermott, and P. Szolovits. Unsupervised multimodal representation learning across medical images and reports. *arXiv preprint arXiv:1811.08615*, 2018.

[13] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

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

[15] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597, 2019.

[16] A. E. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, and S. Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6, 2019.

[17] C. E. Kahn Jr, C. P. Langlotz, E. S. Burnside, J. A. Carrino, D. S. Channin, D. M. Hovsepian, and D. L. Rubin. Toward best practices in radiology reporting. *Radiology*, 252(3):852–856, 2009.

[18] J. Kalpathy-Cramer and W. Hersh. Multimodal medical image retrieval: image categorization to improve search precision. In *Proceedings of the international conference on Multimedia information retrieval*, pages 165–174, 2010.

[19] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

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

[21] Y. Li, H. Wang, and Y. Luo. 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)*, pages 1999–2004. IEEE, 2020.

[22] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

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

[24] G. Liu, T.-M. H. Hsu, M. McDermott, W. Boag, W.-H. Weng, P. Szolovits, and M. Ghassemi. Clinically accurate chest x-ray report generation. *arXiv preprint arXiv:1904.02633*, 2019.

[25] H. Liu, Z. Zhang, Y. Xu, N. Wang, Y. Huang, Z. Yang, R. Jiang, and H. Chen. Use of bert (bidirectional encoder representations from transformers)-based deep learning method for extracting evidences in chinese radiology reports: development of a computer-aided liver cancer diagnosis framework. *Journal of medical Internet research*, 23(1):e19689, 2021.

[26] J. Lu, D. Batra, D. Parikh, and S. Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23, 2019.

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

[28] D. Qi, L. Su, J. Song, E. Cui, T. Bharti, and A. Sacheti. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. *arXiv preprint arXiv:2001.07966*, 2020.

[29] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio. Transfusion: Understanding transfer learning for medical imaging. In *Advances in neural information processing systems*, pages 3347–3357, 2019.

[30] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.

[31] T. Schlegl, S. M. Waldstein, W.-D. Vogl, U. Schmidt-Erfurth, and G. Langs. Predicting semantic descriptions from medical images with convolutional neural networks. In *International Conference on Information Processing in Medical Imaging*, pages 437–448. Springer, 2015.

[32] L. H. Schwartz, D. M. Panicek, A. R. Berk, Y. Li, and H. Hricak. Improving communication of diagnostic radiology findings through structured reporting. *Radiology*, 260(1):174–181, 2011.

[33] S. Sharif, R. A. Naqvi, and M. Biswas. Learning medical image denoising with deep dynamic residual attention network. *Mathematics*, 8(12):2192, 2020.

[34] S. Sharif, R. A. Naqvi, M. Biswas, and W.-K. Loh. Deep perceptual enhancement for medical image analysis. *IEEE Journal of Biomedical and Health Informatics*, 2022.

[35] A. Smit, S. Jain, P. Rajpurkar, A. Pareek, A. Y. Ng, and M. P. Lungren. Chexbert: combining automatic labelers and expert annotations for accurate radiology report labeling using bert. *arXiv preprint arXiv:2004.09167*, 2020.

[36] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019.

[37] C. Sun, F. Baradel, K. Murphy, and C. Schmid. Learning video representations using contrastive bidirectional transformer. *arXiv preprint arXiv:1906.05743*, 2019.

[38] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7464–7473, 2019.

[39] H. Tan and M. Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

[40] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[41] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.

[42] X. Wang, Y. Peng, L. Lu, Z. Lu, and R. M. Summers. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9049–9058, 2018.

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

[44] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

[45] Q. Xia, H. Huang, N. Duan, D. Zhang, L. Ji, Z. Sui, E. Cui, T. Bharti, and M. Zhou. Xgpt: Cross-modal generative pre-training for image captioning. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 786–797. Springer, 2021.

[46] X. Yang, M. Ye, Q. You, and F. Ma. Writing by memorizing: Hierarchical retrieval-based medical report generation. *arXiv preprint arXiv:2106.06471*, 2021.

[47] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. J. Corso, and J. Gao. Unified vision-language pre-training for image captioning and vqa. In *AAAI*, pages 13041–13049, 2020.

VI. Appendix

A. Ablation studies

1) Visual Features: To demonstrate the effectiveness of visual features, we performed additional experiments with visual feature sampling methods (full sample vs random 80% sample) and various visual feature extractors (ResNet-50 vs ResNet-101 [11] vs DenseNet-121 [13]). The detailed training method for each model is the same as that of MedViLL. We report the mean and standard deviation through 5 times random bootstrap experiments. As shown in TABLE VI, TABLE VII, TABLE VIII, and TABLE IX, all results reached poor performance in both in-domain and out-of-domain evaluation when using all the visual features of the CNN. These results show the difficulty of reaching good performance by over-fitting the MIMIC-CXR training set. Also, there are performance differences for each downstream task when using various visual feature extractor. We believe that better performance can be obtained with various experiments on the model architecture or hyper-parameter tuning (e.g., various visual feature extractors, BERT transforms, etc.)

2) Self-attention mask: We explore the remaining 3 types of self-attention masks in this section. Bidirectional attention mask ((a) of Fig. 4 and Fig. 8) allows the attention calculation of all possible combinations (Eq. (10) - (13)) that encourage unconstrained context learning between vision and language modalities. This type of mask drives learning to understand the input correlation between the two vision-language modalities. On the other hand, Sequence-to-Sequence attention mask ((d) of Fig. 4 and Fig. 8) allows the attention calculation only to key from the visual modality (Eq. (10), (12)) and restricts all visual queries to attend any language key so that attention can be computed on input features sequentially. A model trained in this way can generate future features based on given input features in an auto-regressive manner. To restricts the interaction between two modalities, the Non-crossing attention mask ((c) of Fig. 4 and Fig. 8) only allows the attention calculation of each modality (Eq. (10), (12)).

TABLE VI: AUROC and F1 scores for the diagnosis classification task on MIMIC-CXR and Open-I. * indicates using all visual features.

Dataset	Metrics	MedViLL	MedViLL *	MedViLL (Resnet101)	MedViLL (Densenet121)
MIMIC-CXR	avg AUROC	**0.982 (0.00)**	0.926 (0.0)	0.918 (0.0)	0.962 (0.0)
	avg F1	**0.841 (0.01)**	0.699 (0.0)	0.666 (0.01)	0.771 (0.0)
Open-I	avg AUROC	0.894 (0.00)	0.886 (0.00)	0.894 (0.00)	**0.897 (0.00)**
	avg F1	0.408 (0.01)	0.397 (0.00)	0.404 (0.00)	**0.409 (0.00)**

TABLE VII: Medical Image-Report Retrieval performance on MIMIC-CXR and Open-I. * indicates using all visual features.

Task	Models	MIMIC-CXR				OpenI			
		MRR	H@5	R@5	P@5	MRR	H@5	R@5	P@5
Report-to-Image	MedViLL	**58.1 (0.4)**	**78.4 (0.7)**	**47.8 (0.5)**	**20.0 (0.1)**	**56.5 (1.3)**	**74.8 (1.4)**	13.3 (0.3)	**37.0 (0.6)**
	MedViLL *	49.0 (0.1)	69.7 (0.1)	41.9 (0.1)	17.6 (0.1)	54.8 (0.1)	74.7 (0.1)	13.4 (0.1)	35.6 (0.1)
	MedViLL (Resnet101)	47.5 (0.1)	69.3 (0.1)	41.5 (0.1)	17.4 (0.1)	53.1 (0.1)	73.2 (0.1)	12.8 (0.1)	35.8 (0.1)
	MedViLL (Densenet121)	49.3 (0.1)	70.0 (0.1)	42.2 (0.1)	17.75 (0.1)	56.4 (0.1)	74.5 (0.1)	**14.0 (0.1)**	36.7 (0.1)
Image-to-Report	MedViLL	**56.1 (0.2)**	**76.2 (0.6)**	**48.1 (0.7)**	**19.9 (0.6)**	**49.3 (0.9)**	**62.2 (0.7)**	12.4 (0.2)	34.2 (0.5)
	MedViLL *	48.6 (0.1)	68.0 (0.1)	42.4 (0.1)	17.4 (0.1)	48.5 (0.1)	**62.2 (0.1)**	12.5 (0.1)	**34.3 (0.1)**
	MedViLL (Resnet101)	46.3 (0.1)	65.5 (0.1)	39.8 (0.1)	17.1 (0.1)	47.2 (0.1)	62.0 (0.1)	11.8 (0.1)	32.5 (0.1)
	MedViLL (Densenet121)	48.6 (0.1)	69.4 (0.1)	42.7 (0.1)	18.38 (0.1)	46.5 (0.1)	60.9 (0.1)	**12.8 (0.1)**	31.7 (0.1)

TABLE VIII: Model accuracy on the VQA-RAD dataset. * indicates using all visual features.

Models	ALL		CHEST	
	Open-ended	Close-ended	Open-ended	Close-ended
MedViLL	**0.597(0.038)**	0.782(0.022)	0.608(0.051)	**0.783(0.033)**
MedViLL *	0.512(0.012)	0.743(0.009)	0.572(0.021)	0.736(0.006)
MedViLL (Resnet101)	0.548(0.019)	0.781(0.002)	0.602(0.012)	0.773(0.022)
MedViLL (Densenet121)	0.593(0.001)	**0.787(0.004)**	**0.612(0.01)**	0.781(0.003)

TABLE IX: Report generation performance in terms of Perplexity and Label Accuracy, Precision Recall and F1 and BLEU4. * indicates using all visual features.

Dataset	Models	Perplexity (↓)	Accuracy (↑)	Precision (↑)	Recall (↑)	F1 Score (↑)	BLEU4 (↑)
MIMIC	MedViLL	4.19(0.03)	**0.841(0.003)**	**0.698(0.002)**	**0.558(0.004)**	**0.620(0.010)**	0.066(0.001)
	MedViLL *	5.04(0.01)	0.780(0.016)	0.643(0.038)	0.417(0.026)	0.505(0.027)	0.072(0.009)
	MedViLL (Resnet101)	**3.91(0.003)**	0.833(0.02)	0.652(0.004)	0.472(0.024)	0.526(0.003)	0.116(0.02)
	MedViLL (Densenet121)	4.02(0.002)	0.81(0.003)	0.628(0.02)	0.425(0.004)	0.484(0.005)	**0.126(0.003)**
Open-I	MedViLL	5.58(0.32)	0.734(0.001)	0.512(0.003)	**0.594(0.001)**	**0.550(0.009)**	0.049(0.001)
	MedViLL *	5.06(0.081)	0.790(0.018)	0.577(0.015)	0.294(0.089)	0.382(0.081)	0.044(0.010)
	MedViLL (Resnet101)	**3.815(0.002)**	0.812(0.003)	**0.675(0.006)**	0.367(0.003)	0.466(0.001)	**0.075(0.031)**
	MedViLL (Densenet121)	3.998(0.013)	**0.818(0.011)**	0.582(0.03)	0.385(0.012)	0.472(0.002)	0.071(0.009)

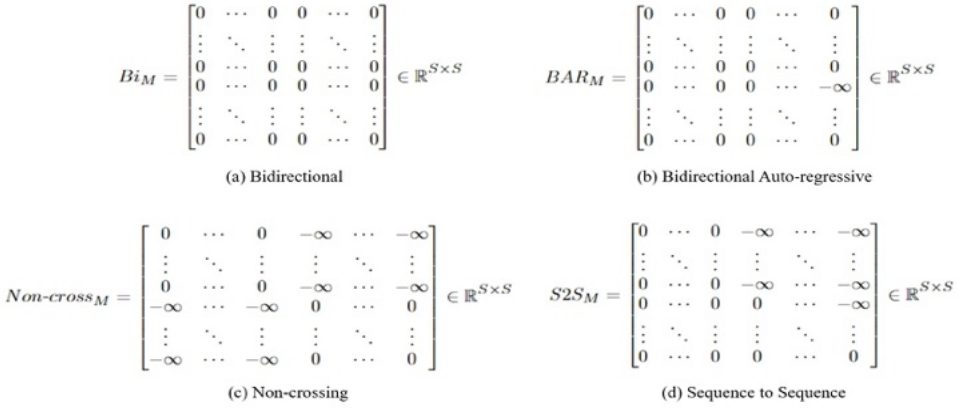


Fig. 8: **Self-attention mask schemes.** Four types of the self-attention masks in matrix. The self-attention mask $M \in R^{S \times S}, S = visual\ features + language\ features + 3special\ tokens$ consists of 0s and negative infinities.