Improving Joint Learning of Chest X-Ray and Radiology Report by Word Region Alignment

Zhanghexuan Ji[1](✉), Mohammad Abuzar Shaikh[1](✉), Dana Moukheiber[1],
Sargur N Srihari[1], Yifan Peng[2], and Mingchen Gao[1]

[1] Department of Computer Science and Engineering, University at Buffalo,
The State University of New York, Buffalo, NY, USA
{zhanghex,mshaikh2,danamouk,srihari,mgao8}@buffalo.edu

[2] Population Health Sciences, Weill Cornell Medicine, New York, NY, USA
yip4002@med.cornell.edu

Abstract. Self-supervised learning provides an opportunity to explore unlabeled chest X-rays and their associated free-text reports accumulated in clinical routine without manual supervision. This paper proposes a Joint Image Text Representation Learning Network (JoImTeR-Net) for pre-training on chest X-ray images and their radiology reports. The model was pre-trained on both the global image-sentence level and the local image region-word level for visual-textual matching. Both are bidirectionally constrained on Cross-Entropy based and ranking-based Triplet Matching Losses. The region-word matching is calculated using the attention mechanism without direct supervision about their mapping. The pre-trained multi-modal representation learning paves the way for downstream tasks concerning image and/or text encoding. We demonstrate the representation learning quality by cross-modality retrievals and multi-label classifications on two datasets: OpenI-IU and MIMIC-CXR. Our code is available at https://github.com/mshaikh2/JoImTeR_MLMI_2021.

Keywords: Self-supervised learning · Multi-modality · Attention

1 Introduction

Chest X-ray is the most common medical imaging study globally for conducting clinical routines to assess chest regions. Because of its popularity, large, labeled datasets such as ChestX-ray14 dataset [24], CheXpert [10], OpenI-IU [5], and MIMIC-CXR [11,12], were collected as benchmarks for data-driven deep learning models to archive expert-level performance in analyzing chest regions. Among these biomedical datasets, OpenI-IU and MIMIC-CXR contain radiology reports along with corresponding radiographs. Given the large size of collected images and manual labeling being impractical, the disease labels are usually derived using natural language processing tools applied to the corresponding radiology reports.

Z. Ji and M.A. Shaikh—Equal contributions.

C. Lian et al. (Eds.): MLMI 2021, LNCS 12966, pp. 110–119, 2021.
https://doi.org/10.1007/978-3-030-87589-3_12

Recently, self-supervised representation learning has been explored to extract underlying information from the data by performing proxy tasks that explore the organization of the data itself. This is a promising direction for learning from a large amount of unlabeled biomedical data, where manual labeling is tedious, time-consuming, subjective, and requires domain knowledge. Self-supervised learning provides a great potential to investigate the biomedical data, including both medical images and their associated reports, accumulated during clinical routines. Ideally, both the modalities of the data encode the same medical condition and should be cross-referable.

Self-Attention mechanism was introduced to find the cross references within the same data modality [23]. This concept has contributed tremendously to the recent success of natural language processing models, such as BERT [6]. These models are pre-trained by predicting masked tokens to learn the underlying semantic representations from unlabeled textual data. Once the representation learning models are pre-trained, they can be fine-tuned and used as a backbone for a wide range of downstream natural language processing tasks.

Motivated by the above discussion, we propose to establish the cross references of the chest radiology images and reports to jointly learn the image-text representations. Learning cross-modal visual and textual representation is an essential task that can combine the semantic information contained within images and their descriptive reports [16,17]. These approaches have also been explored in biomedical image analysis [18]. The proposed representation learning mechanism will provide the foundation for a wide range of biomedical vision-and-language tasks, such as clinical inter-modal and intra-modal image-text retrieval, medical visual question answering [1], and automatic clinical report generation [19].

Contributions: We propose JoImTeRNet - a self-supervised pre-training network trained on multimodal inputs. Our network extracts and fuses the representations of the visual and textual modalities using both global image-sentence matching and local attention-based region-phrase matching. Phrases vary from length of one to three words. The proposed local region-phrase alignment enhances the joint representation learning by performing automatic fine-grained matching between image region-of-interests with phrases in reports. The local region-phrase matching is further enhanced using a soft-attention mechanism in the image encoder, without the need for explicit manual bounding box annotation or object detection on images. The quality of the learned representation is tested on the downstream classification and retrieval tasks.

2 Joint Image Text Representation Learning Network

We propose a **Jo**int **Im**age **Te**xt **R**epresentation Learning **Net**work (JoImTeRNet) shown in Fig. 1. The JoImTeRNet architecture consists of an image and a text encoder. The representations are matched through a list of matching tasks, Text to Image Matching (TIM), Masked Language Modeling (MLM), Phrase to Region Alignment (PRA), and Word to Region Alignment (WRA). The learned

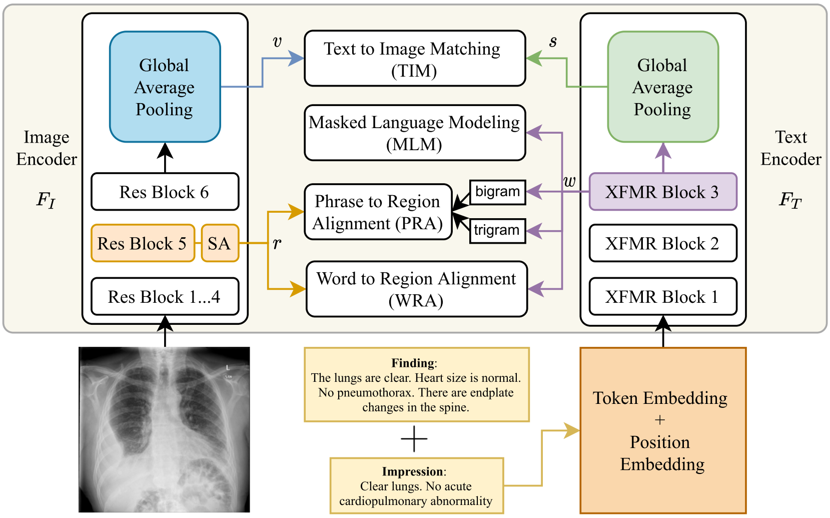
Image Encoder
F_I
Global Average Pooling
Res Block 6
Res Block 5
SA
Res Block 1...4
v
r
Text to Image Matching (TIM)
Masked Language Modeling (MLM)
Phrase to Region Alignment (PRA)
bigram
trigram
Word to Region Alignment (WRA)
s
w
Global Average Pooling
XFMR Block 3
XFMR Block 2
XFMR Block 1
Text Encoder
F_T
Finding: The lungs are clear. Heart size is normal. No pneumothorax. There are endplate changes in the spine.
+
Impression: Clear lungs. No acute cardiopulmonary abnormality
Token Embedding + Position Embedding

Fig. 1. The architecture of the proposed JoImTeRNet.

image and text representation are mapped to a shared feature space, given the hypothesis that the radiographs and their corresponding report contain consistent semantic meaning.

Given an X-ray image I and its corresponding radiology report T, we first encode them with an image encoder F_I and a text encoder F_T. The image encoder contains one input convolution layer, 6 residual blocks [8] and a global average pooling (GAP) layer. In the meantime, we also get the output from the Soft-Attention (SA) [22] block placed after the ResBlock5 to extract the region features $r \in \mathbb{R}^{D \times M}$, such that $v, r = F_I(\mathrm{I})$ where $v \in \mathbb{R}^D$ is the global image features from GAP. Sentence and word level features s and w are extracted using a Transformer [23] based text encoder F_T, such that $w, s = F_T(\mathrm{T})$, where $w \in \mathbb{R}^{D \times N}$ and $s \in \mathbb{R}^D$. Three transformer layers are deployed in F_T to encode the text report with the self-attention mechanism.

2.1 Matching Images and Sentences

To learn the joint representation of image and text pairs, we use the cross-entropy based matching (CEM) loss [26] and the ranking-based triplet matching (TM) loss [3]. Given a batch of image-text pairs $(\mathrm{I}_i, \mathrm{T}_i)_{i=1}^{B}$ (B is the batch size) and their corresponding visual features v and sentence features s from F_I and F_T, probability of T_i matching with I_i using softmax as: the image-to-text CEM loss L_{CEM}^{IT} is defined as the negative log posterior probability of the images being matched with their corresponding texts, *i.e.*,

$$L_{CEM}^{\mathrm{IT}} = -\sum_{i=1}^{B} log(P(\mathrm{T}_i|\mathrm{I}_i)) = -\sum_{i=1}^{B} log(\frac{e^{\gamma S(\mathrm{I}_i,\mathrm{T}_i)}}{\sum_{j=1}^{B} e^{\gamma S(\mathrm{I}_i,\mathrm{T}_j)}}) \quad (1)$$

where γ is the smoothing factor. $P(\mathrm{T}_i|\mathrm{I}_i)$ is the posterior probability of T_i matching with I_i using softmax. Cosine similarity $S(\mathrm{I}_i, \mathrm{T}_i) = (v^T s)/(\|v\|\|s\|)$ is used as the similarity score between image-text pairs. During the training, T_i is the correct match to I_i in the batch and all the other $\mathrm{T}_j (j \neq i)$ are mismatching texts. Considering that image-text joint representation mapping should be bidirectional, we reverse I and T in Eq. (1) and get the symmetric text-to-image CEM loss as L^{TI}_{CEM}. Thus, the bidirectional CEM loss for globally matching image and text is defined as $L^s_{CEM} = L^{\mathrm{IT}}_{CEM} + L^{\mathrm{TI}}_{CEM}$.

Although CEM loss is designed to make the similarity between correct image-text pairs relatively higher than other mismatched pairs, it is difficult to set a hard margin between mismatched features. To solve this problem, TM loss [3], a ranking-based criterion, is added to increase the distance of mismatched pairs in the joint embedding space. Given an image I_i as the anchor, T_i is used as the positive paired sample. We then randomly select a mismatching text $\mathrm{T}_j (j \neq i)$ within the batch as the negative paired sample. Symmetrically, if T_i is used as the anchor, then I_i and I_j would be positive/negative samples. The bidirectional TM loss for global image-text matching is formed as:

$$\begin{aligned} L^s_{TM} = L^{\mathrm{IT}}_{TM} + L^{\mathrm{TI}}_{TM} = \sum_{i,j=1}^{B} \Big[&\max(0, S(\mathrm{I}_i, \mathrm{T}_j) - S(\mathrm{I}_i, \mathrm{T}_i) + \eta_s) \\ &+ \max(0, S(\mathrm{I}_j, \mathrm{T}_i) - S(\mathrm{I}_i, \mathrm{T}_i) + \eta_s) \Big] \end{aligned} \tag{2}$$

where η_s is the hard margin and S is the cosine similarity the same as in Eq. (1).

2.2 Aligning Image Regions and Report Phrases

Both chest X-rays and their corresponding reports contain lots of fine-grained semantic information. We introduce a region-phrase level matching to align different concepts in the text reports with the regions of the images to further improve the joint representation. We apply region-phrase alignment with both CEM loss and TM loss. The length of a phrase is in the range of 1 to 3 words. Features of words, bigram and trigram phrases are denoted as w, p_2, p_3 respectively.

The cosine similarity between regions and words/phrases is not feasible to calculate directly due to the lack of explicit mapping between them. Instead, an attention-based matching score is deployed to overcome this challenge [7,9,26]. For region-word-level matching, given $(\mathrm{I}_i, \mathrm{T}_i)$ and their region-word features (r, w), we first calculate the similarity matrix between all possible pairs of region features and word features using dot-product, *i.e.*, $m = w^T r$, where $m \in \mathbb{R}^{N \times M}$, which is further normalized along N words as $\bar{m} = \mathrm{Softmax}_N(m)$. Next, a context feature c is computed as the weighted sum over region features r, weighted by the region-word attention score α as follows:

$$c = \alpha r^T, \text{ where } \alpha_{i,j} = \frac{e^{\gamma_1 \bar{m}_{i,j}}}{\sum_{k=0}^{M-1} e^{\gamma_1 \bar{m}_{i,k}}} \tag{3}$$

where $c \in \mathbb{R}^{N \times D}$ and $\alpha \in \mathbb{R}^{N \times M}$; γ_1 is a hyper-parameter to tune the required amount of visual attention for a word. Here, the i^{th} vector of c is the attention-weighted representation of all the sub-regions related to the i^{th} word.

The attention-based region-word-level matching score is computed as:

$$S_a(\mathrm{I}, \mathrm{T}) = \log(\sum_{i=1}^{N-1} e^{(\gamma_2 S(c_i, w_i))})^{\frac{1}{\gamma_2}} \tag{4}$$

where $S(c_i, w_i) = (c_i^T w_i)/(\|c_i\|\|w_i\|)$, is the element-wise cosine similarity score between c_i and w_i, γ_2 is the importance magnification hyper-parameter for the most relevant word and context vector pair.

By replacing the cosine similarity score $S(\cdot, \cdot)$ with the region-word matching score $S_a(\cdot, \cdot)$ in Eq. (1) (2), we obtain the bidirectional CEM loss and TM loss for region-word alignment as $L_{CEM}^{p_1} = L_{CEM}^{rw} + L_{CEM}^{wr}$ and $L_{TM}^{p_1} = L_{TM}^{rw} + L_{TM}^{wr}$.

Furthermore, we obtain the phrase features by applying a 1D convolutional layer with kernel size 2 and 3 over w to get bigram $p_2 = \theta_{p_2}^T w$ and trigram $p_3 = \theta_{p_3}^T w$ phrase features respectively [14,27]. Here, $\theta_{p_2}, \theta_{p_3}$ are the convolution kernels of size 2 and 3. Our final cross-entropy with triplet matching (CETM) loss for our image-text joint representation learning is designed as:

$$L_{CETM} = \lambda_{CEM}(L_{CEM}^{s} + \sum_{i=1}^{3} L_{CEM}^{p_i}) + \lambda_{TM}(L_{TM}^{s} + \sum_{i=1}^{3} L_{TM}^{p_i}) \tag{5}$$

where λ_{CEM} and λ_{TM} are the loss weight hyper-parameters.

2.3 Downstream Task

In order to demonstrate the performance of joint representation learning, we use the pre-trained image and text encoders as the backbone and test the learned features on multi-label classification. We add projection layers followed by two fully connected layers for multi-label classification. Cross-entropy loss balanced with positive/negative ratio and class-wise weights [25] are used for training.

3 Experiments

3.1 Datasets

MIMIC-CXR v2.0 [11], is a large public dataset consisting of 377,110 chest X-rays associated with 227,835 radiology reports. We limit our study to the frontal-view images and only keep one frontal view image for each report. Following the pre-processing scheme in [3], we extract the *impressions, findings, conclusion* and *recommendation* sections from the raw report, normalized by SciSpaCy [20], and concatenate them. If none of these sections are present, we use the *final report* section. 14 CheXpert labels provided in MIMIC-CXR are used for classification task, where label 1 is considered as positive and all the other labels (−1, 0) and

Table 1. Ablation for selecting the best loss setting. The matching score for OpenI-IU and MIMIC-CXR is computed on 1000 and 1000/3000 test samples respectively. Subscript s, w, p stand for image-text, region-word and region-phrase level matching.

Model setting	MIMIC-CXR												OPENI-IU					
	I2T (1K)			T2I (1K)			I2T (3K)			T2I (3K)			I2T (1K)			T2I (1K)		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
TM_s[3]	5.37	19.43	30.73	5.40	20.23	30.23	2.37	9.70	15.63	2.23	10.20	16.37	1.83	5.70	9.13	1.50	5.67	9.30
TM_{ws}	6.30	21.73	32.23	6.00	20.97	30.90	2.77	10.67	18.07	2.83	10.97	17.83	1.93	6.37	10.17	1.97	6.53	10.30
CEM_s[7]	18.60	43.10	56.10	18.13	43.20	55.97	12.20	31.27	41.80	11.80	30.87	41.10	4.70	12.83	17.73	4.83	12.30	17.87
CEM_{ws}[26]	18.60	44.20	56.27	18.87	43.40	55.67	12.60	31.67	41.80	12.83	31.57	41.27	4.87	13.00	18.00	5.37	13.40	18.33
CETM_{ws}	**19.07**	45.33	57.20	**19.07**	44.70	56.73	**12.77**	31.90	43.00	**12.97**	31.97	42.03	**5.13**	**13.07**	18.73	5.50	13.73	19.20
CETM_{wps}	18.93	**46.20**	**58.67**	**19.07**	**45.27**	**58.50**	12.67	**33.20**	**44.07**	12.83	**32.43**	**43.40**	5.07	**13.07**	**18.83**	**5.67**	**13.83**	**19.20**

missing labels are merged as negative. This results in 222,252 image-report pairs with 14 binary labels. We split the dataset into 217,252, 2,000 and 3,000 samples for training, validation and testing respectively.

OpenI-IU. [5] is a public dataset with 3,996 radiology reports and 8,121 associated chest X-ray images, which are manually annotated by human experts using MeSH words. Similar to TieNet [25], only unique frontal images and their corresponding reports which contain either *findings* and/or *impressions* are selected. This yields 3,643 image-report pairs, which are only used as external evaluation sets. For comparison and evaluation purposes, we select the 7 common labels in both OpenI and MIMIC-CXR (Table 3) from the MeSH domain.

3.2 Implementation Details

JoImTeRNet is implemented in Pytorch [21] and all the experiments are carried out on NVIDIA GTX 1080 Ti GPUs. For F_I, we use the basic residual blocks proposed in [8]. We employ 3 layers of Transformer blocks with 8 heads in F_T. The input image is encoded into 256 regions (r) flattened from 16×16 feature map output from Res Block 5 as shown in Fig. 1. The input image is cropped or padded to 2048×2048 then normalized to $[-1, 1]$. Random crop, rotation and color jitter are used for data augmentation. Report input is tokenized by a word-level tokenization scheme, where we collect all the words that appear more than twice in MIMIC-CXR dataset, which results in vocabulary size of $8,410$. The input reports are truncated or padded to the max length of $N = 160$.

Parameter Settings. We pretrain F_I and F_T on MIMIC-CXR training set using the image-text matching task explained in Sect. 2.1 and 2.2 to generate the joint image and text representations. The maximum epoch is set as 30. We employ AdamW [15] optimizer with an initial learning rate of 10^{-4}, which is dropped by 10 times after 20 epochs. L2 weight decay is set as 10^{-4}. For the downstream classification task in Sect. 2.3, we set up two different settings for comparison: randomly initializing the backbone and fine-tuning the pre-trained backbone. The learning rate for the classification head in both settings is set to 10^{-4}. For the randomly initialized setting, we train the backbone using the same learning rate as the classification block for 20 epochs, whereas the pre-trained backbone is fine-tuned with a smaller learning rate of 10^{-5} for only 10

Table 2. Classification AUCs on MIMIC-CXR [11] dataset. "FS" stands for training from scratch (FS). "FT" stands for fine-tuned model. Other comparison experiments are Visualbert [17], Uniter [4], and ClinicalBert [2].

Findings	Image		Image+Report				Report		
	FS	FT	[17]	[4]	FS	FT	[2]	FS	FT
EC	0.738	**0.763**	0.981	0.979	0.989	**0.996**	0.966	0.986	**0.980**
Cardiomegaly	0.794	**0.820**	0.991	0.989	0.989	**0.993**	0.979	0.988	**0.989**
Airspace opacity	0.749	**0.759**	0.991	0.989	0.988	**0.992**	0.978	0.985	**0.986**
Lung lesion	0.696	**0.756**	0.985	0.981	0.996	**0.998**	0.972	**0.989**	0.987
Edema	0.883	**0.895**	0.991	0.990	0.995	**0.996**	0.979	0.993	**0.994**
Consolidation	0.794	**0.810**	0.989	0.988	0.994	**0.998**	0.979	**0.996**	0.995
Pneumonia	0.733	**0.738**	0.977	0.974	0.981	**0.985**	0.962	0.980	**0.984**
Atelectasis	0.808	**0.824**	0.988	0.987	0.988	**0.995**	0.976	0.991	**0.992**
Pneumothorax	0.845	**0.855**	0.992	0.991	0.989	**0.993**	0.979	0.993	**0.994**
Pleural effusion	0.898	**0.904**	0.993	0.992	0.986	**0.997**	0.981	0.989	**0.990**
Pleural others	0.812	**0.839**	0.981	0.973	0.996	**0.999**	0.964	**0.998**	0.993
Fracture	0.641	**0.714**	0.976	0.977	**0.997**	0.990	0.958	**0.997**	**0.997**
Support devices	0.901	**0.913**	**0.995**	0.994	0.992	**0.995**	0.983	0.988	**0.993**
No findings	0.865	**0.874**	–	–	0.985	**0.989**	–	**0.980**	**0.980**
Avg	0.792	**0.815**	0.987	0.985	0.991	**0.994**	0.974	**0.990**	**0.990**

epochs. Our model pre-trained with the full loss setting CETM$_{wps}$ is used as the backbone for fine-tuning. The batch size is set to 32 for all the experiments. We select the loss hyper-parameters as $\gamma, \gamma_1, \gamma_2 = 2, 1, 1$, $\eta_s, \eta_w = 0.5, 0.5$ and $\lambda_{CEM}, \lambda_{TM} = 2.0, 1.0$.

3.3 Performances

Evaluation Metric. We evaluate the performance of JoImTeRNet by cross-modality retrieval task: given one image (text) as a query, we rank a subset of text (image), including the paired one, based on cosine similarity between the image and text features from JoImTeRNet. Recall@K (R@K) [13] is reported, where $K \in \{1, 5, 10\}$, which measures the fraction of times the correct matching is retrieved among the top K results in the test set. We compute R@K on a subset of 1000 image-text pairs and on the full 3000 samples in our MIMIC-CXR test set. We also report R@K on a subset of 1000 samples in OpenI-IU in order to evaluate JoImTeRNet on the external dataset (Table 1).

Ablation Study for Loss Settings. Ablation studies for different combinations of our losses are listed in Table 1. As we can see, full loss setting CETM$_{wps}$ achieves the highest R@5 and R@10 scores on all the test set, which shows the effectiveness of our multilevel phrase matching loss. In addition, the matching

Table 3. Classification AUCs on OpenI-IU [5] dataset. “FS” stands for training From Scratch (FS). “FT” stands for Fine-tuned model. Other comparison experiments are ChestX-ray14 [24], TieNet [25], Visualbert [17], Uniter [4], and ClinicalBert [2].

Findings	Image			Image+Report					Report				No. of samples
	[24]	FS	FT	[25]	[17]	[4]	FS	FT	[25]	[2]	FS	FT	
Cardiomegaly	0.803	0.924	**0.937**	0.962	0.977	0.978	0.956	**0.985**	0.944	0.969	0.966	**0.987**	315
Edema	0.799	0.937	**0.953**	**0.995**	0.982	0.989	0.922	0.962	**0.984**	0.976	0.947	0.964	40
Consolidation	0.790	0.951	**0.951**	0.989	0.996	**0.998**	0.954	0.975	0.969	**0.982**	0.938	0.975	28
Pneumonia	0.642	0.863	**0.934**	**0.994**	0.990	0.988	0.877	0.949	**0.983**	0.982	0.880	0.943	36
Atelectasis	0.702	0.829	**0.858**	0.972	**0.992**	0.982	0.947	0.978	**0.981**	0.947	0.952	0.971	293
Pneumothorax	0.631	0.926	**0.936**	0.960	0.988	0.983	0.962	**0.989**	0.960	0.973	0.951	**0.989**	22
Pleural effusion	0.890	0.938	**0.957**	0.977	**0.985**	0.983	0.922	0.971	0.968	**0.976**	0.926	0.968	140
No finding	–	0.844	**0.851**	–	–	–	0.883	**0.961**	–	–	0.898	**0.930**	2789
Avg	0.751	0.910	**0.932**	0.978	**0.987**	0.986	0.934	0.973	0.970	**0.972**	0.932	0.971	
W. Avg	0.771	0.893	**0.915**	0.971	**0.985**	0.982	0.943	0.978	0.965	0.964	0.949	**0.975**	

performance degrades when the model is trained on global matching loss only without the region-phrase(word)-level matching, i.e. CEM_s performs worse than CEM_{ws}. Similar results are found when comparing TM_s with TM_{ws}. This result shows that our proposed method for assisting joint representation learning using region-word matching is able to improve the representation ability of the image-text encoder. Moreover, the CETM combination consistently gains performance compared with only CEM loss or TM loss settings, which is just as we expected in Sects. 2.1 and 2.2. Notice that the matching scores are much lower on OpenI, since OpenI contains a large amount of similar reports, e.g. ‘No acute disease.’, which can have very similar feature representation from our model and thus largely degrade the matching score.

Downstream Image Classification Results. The AUCs from our two settings on both datasets along with other SOTA performances are shown in Table 2 and 3. We can see that the classifier performance finetuned on JoimTerNet backbone (FT) is always higher than training from scratch (FS), which shows the advance of our pre-training method. As shown in Table 2, FT achieves the highest AUCs on most tasks and labels on MIMIC-CXR test set (internal evaluation), even better than some SOTA models [2,4,17] on image-text and text classification. For the external evaluation on OpenI in Table 3, our FT setting extremely improves average AUC on image classification by 18% compared with TieNet [25], and also gains 1% on wAvg AUC than ClinicalBERT [2] on report classification. For the image-text classification, our model is still comparable with other SOTA models, even though our text encoder only contains 3 transformer layers compared with [4,17] which has a 12 layer BERT encoder as the backbone.

4 Conclusion

We propose a joint image-text representation learning network and show its performance on cross-modality retrieval and multi-label classification. We demonstrate the potential of self-supervised learning when it meets the continuously generated biomedical images and reports. We also leverage and show the importance of information contained within the relationship of words, phrases and image regions. Future work includes more complicated downstream tasks regarding both images and text.

Acknowledgment. This research was supported in part by NSF through grant IIS-1910492. It also was supported by the National Library of Medicine under Award No. 4R00LM013001.

References

1. Abacha, A.B., Hasan, S.A., Datla, V.V., Liu, J., Demner-Fushman, D., Müller, H.: VQA-Med: overview of the medical visual question answering task at ImageCLEf 2019. In: CLEF (Working Notes) (2019)
2. Alsentzer, E., et al.: Publicly available clinical BERT embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop, pp. 72–78. Association for Computational Linguistics (2019)
3. Chauhan, G., et al.: Joint modeling of chest radiographs and radiology reports for pulmonary edema assessment. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12262, pp. 529–539. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59713-9_51
4. Chen, Y.-C., et al.: UNITER: UNiversal image-TExt representation learning. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12375, pp. 104–120. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58577-8_7
5. Demner-Fushman, D., et al.: Preparing a collection of radiology examinations for distribution and retrieval. J. Am. Med. Inform. Assoc. **23**(2), 304–310 (2016)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186. Association for Computational Linguistics (2019)
7. Fang, H., et al.: From captions to visual concepts and back. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1473–1482 (2015)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
9. Huang, P.S., He, X., Gao, J., Deng, L., Acero, A., Heck, L.: Learning deep structured semantic models for web search using clickthrough data. In: Proceedings of the 22nd ACM International Conference on Information & Knowledge Management, pp. 2333–2338 (2013)
10. Irvin, J., et al.: CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 590–597 (2019)
11. Johnson, A.E., et al.: MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci. Data **6**(1), 1–8 (2019)
12. Johnson, A.E., et al.: MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042 (2019)

13. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3128–3137 (2015)
14. Kim, Y.: Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1746–1751. Association for Computational Linguistics (2014)
15. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
16. Li, G., Duan, N., Fang, Y., Gong, M., Jiang, D.: Unicoder-VL: a universal encoder for vision and language by cross-modal pre-training. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 11336–11344 (2020)
17. Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J., Chang, K.W.: VisualBERT: a simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557 (2019)
18. Li, Y., Wang, H., Luo, Y.: 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), pp. 1999–2004. IEEE (2020)
19. Li, Y., Liang, X., Hu, Z., Xing, E.P.: Hybrid retrieval-generation reinforced agent for medical image report generation. In: Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, Montréal, Canada, 3–8 December 2018, pp. 1537–1547 (2018)
20. Neumann, M., King, D., Beltagy, I., Ammar, W.: ScispaCy: fast and robust models for biomedical natural language processing. In: Proceedings of the 18th BioNLP Workshop and Shared Task, pp. 319–327. Association for Computational Linguistics (2019)
21. Paszke, A., et al.: Pytorch: an imperative style, high-performance deep learning library. Adv. Neural Inf. Process. Syst. **32**, 8026–8037 (2019)
22. Shaikh, M.A., Duan, T., Chauhan, M., Srihari, S.: Attention based writer independent verification. In: 2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR), pp. 373–379 (2020)
23. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
24. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: 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, pp. 2097–2106 (2017)
25. Wang, X., Peng, Y., Lu, L., Lu, Z., Summers, R.M.: 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, pp. 9049–9058 (2018)
26. Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., He, X.: AttnGAN: fine-grained text to image generation with attentional generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1316–1324 (2018)
27. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.: Stacked attention networks for image question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 21–29 (2016)