Visual-Textual Attentive Semantic Consistency for Medical Report Generation

Yi Zhou *[1], Lei Huang[2], Tao Zhou[3], Huazhu Fu[4], and Ling Shao[4]

[1]School of Computer Science and Engineering, Southeast University, Nanjing, China
[2]SKLSDE, Institute of Artificial Intelligence, Beihang University, Beijing, China
[3]School of Computer Science and Technology, Nanjing University of Science and Technology, China
[4]Inception Institute of Artificial Intelligence, Abu Dhabi, UAE

Abstract

Automatic report generation on medical radiographs have recently gained interest. However, identifying diseases as well as correctly predicting their corresponding sizes, locations and other medical description patterns, which is essential for generating high-quality reports, is challenging. Although previous methods focused on producing readable reports, how to accurately detect and describe findings that match with the query X-Ray has not been successfully addressed. In this paper, we propose a multi-modality semantic attention model to integrate visual features, predicted key finding embeddings, as well as clinical features, and progressively decode reports with visual-textual semantic consistency. First, multi-modality features are extracted and attended with the hidden states from the sentence decoder, to encode enriched context vectors for better decoding a report. These modalities include regional visual features of scans, semantic word embeddings of the top-K findings predicted with high probabilities, and clinical features of indications. Second, the progressive report decoder consists of a sentence decoder and a word decoder, where we propose image-sentence matching and description accuracy losses to constrain the visual-textual semantic consistency. Extensive experiments on the public MIMIC-CXR and IU X-Ray datasets show that our model achieves consistent improvements over the state-of-the-art methods.

1. Introduction

Chest X-Rays are highly important radiological examinations. However, interpreting chest X-Ray images requires the strong expertise and experience of radiologists and is prone to mistakes. Therefore, automatic diagnosis of diseases [41, 35, 19, 21, 31, 7, 51] has been a rising research topic in the medical imaging community. The common tasks include identification of different chest lesions [42, 45, 52] and their corresponding positions and sizes, and generating human-readable reports [26, 24, 20, 53] which contain detailed descriptions such as lesion shape and type.

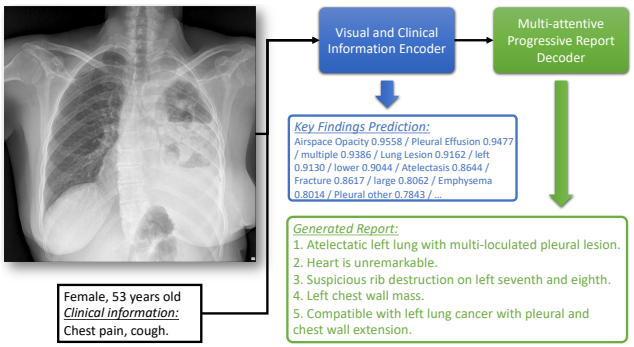

Figure 1. Illustration of automatic medical report generation. Given a chest X-Ray and corresponding clinical information, our model predicts key findings and generates a diagnostics report.

The basic framework of medical report generation is similar to that of image captioning [16]. Currently, most image captioning models [39, 43, 1, 27, 9] adopt deep learning due to its recent breakthroughs in many tasks [15, 14, 37, 40, 50, 11]. However, medical report generation is more challenging than image captioning for two main reasons. First, compared to general images, abnormal lesion appearances in medical images are not always obvious and sometimes even difficult for radiologists to identify. Public benchmarks with paired image-report data are scarce. However, the objects in a general image and corresponding relations among them are very clear and easy to describe. Large-scale matched image-sentence training datasets are available, such as MS COCO [29] and Visual Genome [23]. Second, the target of image captioning is usually to generate one sentence for each image or several sentences with similar descriptions. For medical reports, multiple sentences need to be generated to focus on different diseases in vari-

*Corresponding author: Yi Zhou (yizhou.szcn@gmail.com)

ous regions. Previous methods [30, 47, 49] presented hierarchical decoders to generate different topics, but only used word-level supervision without any constraints on the accuracy and suitability of sentence-level (topic-level) topics.

In this paper, as shown in Fig. 1, given a radiograph and the corresponding clinical information, we propose an automatic diagnosis method to predict key findings and generate detailed descriptions. Clinical information is combined into the inputs of the model because it is closely related to the disease diagnosis and always available together with X-Rays in clinical scenarios. **The main contributions are highlighted as follow: (1)** A multi-modality semantic attention (MMSA) model is proposed to combine different modality features into context vectors for the decoder. Be different from previous attention modules [43, 18], in this work, the regional visual features are both self-attended and correlated with the hidden states of the sentence decoder to obtain semantic attentions at different topic steps. Thus, the MMSA learns both the image-level and topic-level attentions. Moreover, the clinical features and word embeddings of predicted key findings are also integrated for multi-attention learning. **(2)** To optimize the sequential sentence and word decoders, in addition to the word-level supervision, we introduce two topic-level losses at the top of the sentence decoder. An image-sentence matching loss is designed to link the paired image features and generated sentence embeddings, while punishing unpaired ones. Besides, a description accuracy loss is presented to ensure that the generated global report embedding contains correct semantic information for the predicted key findings. **(3)** Extensive experiments are conducted to show the effectiveness of the proposed multi-attention model, matching and description accuracy losses. A new metric, normalized Key Term Distance (nKTD), is also introduced to more reasonably evaluate the medical report generation performance.

2. Related Work

Image Captioning. Recent state-of-the-art image captioning methods are based on generative models [9, 27, 46, 12, 25], which achieve better performance than template-based [8] and retrieval-based [10] models. A general framework for this category is to first encode the visual content of an image and then adopt recurrent language decoders to generate descriptions [39]. Attention mechanisms [43, 1, 18] are introduced to select important regions and focus on the dominating visual objects for better captioning. Though most image captioning approaches only produce one sentence, a few of works [22, 2] have introduced paragraph captioning, which can generate multiple sentences. Krause et al. [22] first proposed a hierarchical framework to generate descriptive image paragraphs, which can tell detailed stories. In [2], coherence vectors and inherence ambiguity of associating paragraphs with images are modeled through a variational auto-encoder. These methods can better fit the medical report generation task, but still do not address the generation accuracy and diversity well.

Medical Report Generation methods can be categorized into automatic generation models and template-based retrieval models. TieNet [42] was first proposed to combine image and text modalities using the CNN-RNN architecture, but mainly focuses on taking both of them as inputs to improve disease classification accuracy. Jing et al. [20] presented a basic co-attention model to implement automatic report generation without enough thoughtful design. Our method enhances this model from many aspects and largely improves the performance. Li et al. [26] combined a template-based method with the generation framework in a reinforcement learning fashion. However, their method requires careful selection of the templates, so the performance varies with different datasets. Finally, knowledge-driven encoding [24] adds an abnormality graph but lacks attention learning for the images. Similarly, in [49], a pre-constructed graph embedding based on multiple disease terms, is adopted to improve generation of medical reports.

3. Proposed Methods

As illustrated in Fig. 2, the proposed method mainly consists of three parts: feature encoding, mutli-modality semantic attention learning, and progressive report decoding.

3.1. Visual and Clinical Feature Encoding

The visual feature encoding module is based on the backbone of DenseNet-201 [17]. The regional feature $\{\mathbf{r}_n\}_{n=1}^{N}$ ('region' denotes each grid in certain feature maps, and N is the number of regions) is extracted from the last dense block, on which a self-attention model and a semantic attention model are applied to weight important regions in the given image. After modeling the long-range dependencies in the image using the self-attention mechanism [48], the global visual feature is obtained and supervised by multi-label classification and global report embedding regression.

To train an image encoder for thorax disease classification, conventional methods [41, 13, 42] only adopt disease labels. In addition to predicting disease categories, we also introduce another type of labels, called description patterns, that contain richer information such as the lesion location, size and shape. The motivation behind this design is that the combination of disease and description pattern labels can enrich the detail and accuracy of generated reports. For example, a nodular found in a scan can be accurately described as "Small nodular opacities in left upper lobe." The two kinds of labels are extracted in two different ways. The disease labels extractor is built on an automated rule-based labeler [19], with necessary modifications to the pre-negation uncertainty, negation and post-negation uncertainty according to the sentences in the two datasets we use. On the other

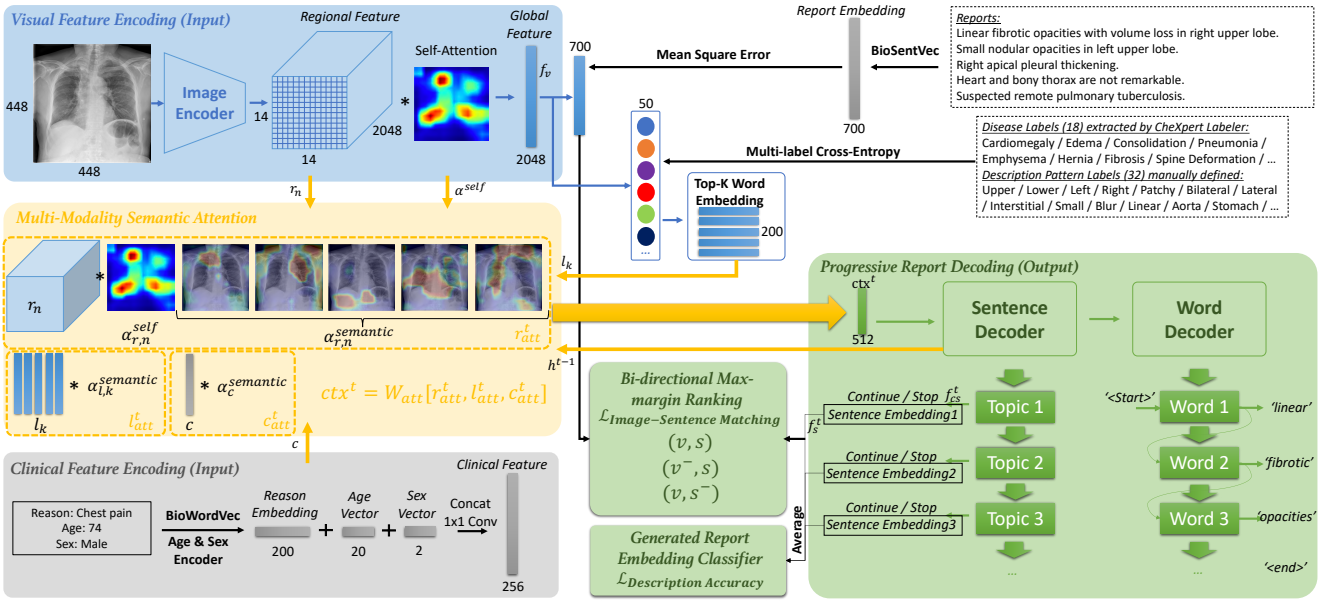

Figure 2. Pipeline of the proposed method. A multi-modality semantic attention model learns correlations across different modality features to compute context vectors for different topics, collaborating with a progressive report decoder. The sentence decoder is optimized under the constraints of matching generated topics with accurate images and report labels, and the word decoder parses each topic into detailed sentences. Dashed lines denote the supervision components that are only available during training.

side, the description pattern labels are positive if the word appears in the report, and negative otherwise. The binary cross-entropy loss is adopted for training. Moreover, the labels predicted for the top-K probabilities are selected and the corresponding word-level embeddings are extracted using BioSentVec [3] for further multi-attention learning.

The discrete labels extracted are noisy and thus cannot preserve all the information from radiology reports, particularly for those useful words that appear less frequently. A report embedding, is also extracted using BioSentVec to train the image encoder, which exploits relationships among words with more semantic information. BioSentVec [3] is trained over 30 million scholarly articles from PubMed and clinical notes from the MIMIC-III database so that the extracted embeddings can accurately represent medical concepts from the original sentences. Therefore, we adopt report embeddings to co-train the image encoder and learn more semantic visual features f_v. This supervision is only available in the training phase. Moreover, inspired by real clinical scenarios, radiologists usually assess an X-Ray image in consideration of its corresponding clinical reason information, such as "chest pain", "fever" and "cough", and the age and sex information. Thus, we encode the reason embedding, age and sex vectors, and integrate them into a clinical feature embedding. The word-level clinical reason embedding is extracted by BioSentVec. The age is encoded as a 20-dimensional one-hot vector for ages ranging from 1-100 years old, with each class covering five years. The sex is encoded as a 2-dimensional vector.

3.2. Multi-Modality Semantic Attention

Traditional visual attention mechanisms usually focus on finding important regions in an image, with the aim of addressing recognition tasks in a data-driven manner. This is helpful but lacks exploration of semantic information. In the medical report generation task, multiple sentences need to be generated based on the different regions that the attention model focuses on for various organs and diseases. In this paper, a multi-modality semantic attention model is proposed to correlate regional image features, semantic embeddings of key findings prediction and clinical information, with the hidden state at each topic step from the sentence decoder. Thus, different findings and exclusions of diseases can be correctly described one by one.

To obtain a concise context vector focusing on one topic at each time step t, three semantic attentions are computed to select relevant features for that topic. For each attention, linear projections are operated over one specific modality feature and the hidden state embedding $\mathbf{h}^{t-1}$ at step $t-1$ of the sentence decoder. $\mathbf{h}^0$ is zero initialized. Then, the soft attention mechanism is adopted as the following functions:

$$\alpha_{r,n}^{semantic} = \exp(\mathbf{W}_{visual} \tanh(\mathbf{W}_r \mathbf{r}_n + \mathbf{W}_{r,h}\mathbf{h}^{t-1})), \quad (1)$$

$$\alpha_{c}^{semantic} = \exp(\mathbf{W}_{clinical} \tanh(\mathbf{W}_c \mathbf{c} + \mathbf{W}_{c,h}\mathbf{h}^{t-1})), \quad (2)$$

$$\alpha_{l,k}^{semantic} = \exp(\mathbf{W}_{label} \tanh(\mathbf{W}_l \mathbf{l}_k + \mathbf{W}_{l,h}\mathbf{h}^{t-1})), \quad (3)$$

where $\mathbf{c}$ denotes clinical features and $\{\mathbf{l}_k\}_{k=1}^{K}$ is the semantic embedding of the top-K predicted labels. $\{\mathbf{W}_{visual}, \mathbf{W}_r, \mathbf{W}_{r,h}\}$, $\{\mathbf{W}_{clinical}, \mathbf{W}_c, \mathbf{W}_{c,h}\}$ and

$\{\mathbf{W}_{label}, \mathbf{W}_l, \mathbf{W}_{l,h}\}$ are trainable parameters for semantic attention learning on the visual, clinical feature and key label embedding, respectively. The corresponding biases are omitted in the equations. Once the attention weights are obtained, the corresponding context vectors for different modalities at each time step t are computed as:

$$\mathbf{r}_{att}^t = \sum_{n=1}^{N} \left[\gamma^{semantic}\alpha_{r,n}^{semantic}\mathbf{r}_n + \gamma^{self}\alpha_{r,n}^{self}\mathbf{r}_n\right], \quad (4)$$

$$\mathbf{c}_{att}^t = \alpha_c^{semantic}\mathbf{c}, \quad (5)$$

$$\mathbf{l}_{att}^t = \sum_{k=1}^{K} \alpha_{l,k}^{semantic}\mathbf{l}_k. \quad (6)$$

For regional visual features, in addition to the semantic attention, a self attention map $\alpha_{r,n}^{self}$ is also learned [48] to better model long-range dependencies for capturing global lesions. The parameters $\gamma^{semantic}$ and γ^{self} are learned to automatically balance the two kinds of attention maps.

Finally, the three attended features $\mathbf{r}_{att}^t$, $\mathbf{c}_{att}^t$ and $\mathbf{l}_{att}^t$ are concatenated and one more linear transformation $\mathbf{W}_{att}$ is used to fuse them as $\mathbf{ctx}^t = \mathbf{W}_{att}\left[\mathbf{r}_{att}^t, \mathbf{c}_{att}^t, \mathbf{l}_{att}^t\right]$. The dimension of the joint context vector is set to 512. A diagram of the whole multi-attention model is shown in yellow in Fig. 2. At different topic steps from the decoder, the semantic attention model produces high responses in different regions of the image to generate the corresponding sentences.

3.3. Report Decoding with Semantic Consistency

3.3.1 Sentence Decoder

The sentence decoder plays a crucial role in determining which part of an image should be described in each sentence. The decoder is based on a two-layer long short-term memory (LSTM) unit that recurrently decodes multiple sentence embeddings. At each topic step t, the LSTM takes as input the current context vector $\mathbf{ctx}^t$ and the hidden states $\mathbf{h}^{t-1}$ and $\mathbf{h}^t$, then models them into two outputs – a predicted sentence embedding $\mathbf{f}_s^t$ and a continue-stop vector $\mathbf{f}_{cs}^t$ – using the following functions:

$$\mathbf{f}_s^t = \mathbf{ReLU}(\mathbf{W}_{ctx}\mathbf{ctx}^t + \mathbf{W}_{s,t}\mathbf{h}^t), \quad (7)$$

$$\mathbf{f}_{cs}^t = \mathbf{Sigmoid}(\mathbf{W}_{cs,t}\mathbf{h}^t + \mathbf{W}_{cs,t-1}\mathbf{h}^{t-1}), \quad (8)$$

where $\{\mathbf{W}_{ctx}, \mathbf{W}_{s,t}, \mathbf{W}_{cs,t}, \mathbf{W}_{cs,t-1}\}$ are additional learnable weights, excluding LSTM weights, in the sentence decoder. The hidden state dimension is set to 512. To optimize the sentence decoder, four different losses are carefully designed for its supervision. First, ground-truth sentence embeddings extracted using BioSentVec are directly adopted and connected to $\mathbf{f}_s^t$ for mean square error learning $\mathcal{L}_{sent}$. Second, the continue-stop vector $\mathbf{f}_{cs}^t$ is supervised by $\mathcal{L}_{stop}$ using ground-truth $\{0, 1\}$ where 0 indicates that the current sentence is not the last sentence, and 1 is adopted otherwise. In the testing phase, if the prediction is over a threshold of 0.5, the sentence decoder will end the unrolling and stop generating sentences. In addition to these two basic losses, an image-sentence matching loss and a description loss are proposed to constrain visual-textual semantic consistency.

3.3.2 Image-Sentence Matching Loss $\mathcal{L}_{ISM}$

Given an extracted visual feature $\mathbf{f}_v$ and the corresponding generated sentence embeddings $\mathbf{f}_s^t$ at topic step t, we map them into a latent space using a non-linear transformation (via the fully-connected layer with ReLU activation, dimension reduced to 128) for feature selection, and create positive pairs $(\mathbf{v}, \mathbf{s}_i)$, $i \in (1, S)$, where S is the number of sentences for the query image. Moreover, we sample two types of negative pairs $(\mathbf{v}^-, \mathbf{s}_i)$ and $(\mathbf{v}, \mathbf{s}_j^-)$, $j \in (1, S)$. $\mathbf{v}^-$ and $\mathbf{s}_j^-$ denote incorrectly matched images and sentences, respectively. Here, we also sample S negative sentences from other reports. Then, the image-sentence matching loss $\mathcal{L}_{ISM}$ is defined by learning a bi-directional max-margin ranking [33] as the following function:

$$\mathcal{L}_{ISM} = \sum_{(\mathbf{v},\mathbf{s})} \Big\{ \max\left[0, m - \frac{1}{S}\sum_i(\mathbf{v}\cdot\mathbf{s}_i) + \frac{1}{S}\sum_i(\mathbf{v}^-\cdot\mathbf{s}_i)\right] + \max\left[0, m - \frac{1}{S}\sum_i(\mathbf{v}\cdot\mathbf{s}_i) + \frac{1}{S}\sum_j(\mathbf{v}\cdot\mathbf{s}_j^-)\right]\Big\}, \quad (9)$$

where m is the margin constraint. This function aims to link the query radiograph with matching diagnosis descriptions by optimizing the sentence decoder to generate sentence embeddings that are close to the visual feature of the input image in the latent space. To compose more effective negative pairs, we sample abnormal images and their sentences if the given query image is normal, and vice versa.

3.3.3 Description Accuracy Loss $\mathcal{L}_{DA}$

As mentioned, the labels for training the image encoder are extracted from the ground-truth reports. To ensure that the generated reports provide accurate descriptions, we constrain the generated sentence embeddings map to correct disease and description pattern labels. Thus, an additional description accuracy net, which consists of three fully-connected layers (512-512-Num of Labels), is constructed for multi-label classification. The input of the generated report embedding classifier is the averaged representation of all sentence embeddings in each report. This net is optimized with the sentence decoder simultaneously.

3.3.4 Word Decoder

Given a sentence embedding $\mathbf{f}_s^t$ of one topic generated by the sentence decoder, a following word decoder will sequentially decode the corresponding words. The word decoder is based on a one-layer LSTM with a dimension of

512, which takes the concatenation of $\mathbf{f}_s^t$ and every word token as inputs. For each topic, the word decoder works similarly to conventional image captioning decoders that only predict one sentence. The $START$ and END tokens are used as the first input and the last output of the decoder, respectively. The output of each step is connected to the next word token and cross-entropy $\mathcal{L}_{word}$ is adopted. The overall loss function of the report generation is defined as:

$$\mathcal{L}_{report} = \lambda_{ISM}\mathcal{L}_{ISM} + \lambda_{DA}\mathcal{L}_{DA} + \lambda_{sent}\mathcal{L}_{sent} + \lambda_{stop}\mathcal{L}_{stop} + \lambda_{word}\mathcal{L}_{word}, \quad (10)$$

where $\{\lambda_{ISM}, \lambda_{DA}, \lambda_{sent}, \lambda_{stop}, \lambda_{word}\}$ are configured as $\{10, 10, 1, 100, 1\}$ for balancing losses to the same scale.

3.4. Implementation Details

In the training phase, the visual feature encoding model is first pre-trained to predict accurate self-attention maps and labels. Then, we fix the parameters of this component and start to train the multi-attention and report decoding models. Otherwise, end-to-end training from the beginning stage prevents the decoders from converging.

In our implementation, the sentence-level and report-level embedding dimensions are both 700, extracted by BioSentVec [3]. The dimension of the word-level embedding for clinical information and top-K label prediction is 200. Moreover, K is set as 20. If the probability of a predicted label in the top-20 is lower than 0.5, the corresponding semantic embedding will be discarded as zero.

To pre-train the image encoder, the ADAM optimizer is adopted with an empirical base learning rate of 0.001 and momentum of 0.5. The mini-batch size is set to 128 for training over 20 epochs. Then, we configure the base learning rate to 0.0002 to train the multi-attention model and report decoders with a batch size of 32.

4. Experimental Results

4.1. Datasets and Evaluation Metrics

MIMIC-CXR. MIMIC [21] is the largest public dataset for chest radiographs, with more than 140k pairs of chest X-Ray images and reports. The images include anteroposterior, posteroanterior and lateral views. The findings section in reports is used as the ground-truth sentences. The indications are used as clinical information. Tokenization is conducted over the corpus and only words with a frequency over 5 are kept, resulting in 5,348 unique tokens in total.

IU X-Ray. Indiana University Chest X-ray Collection [4] is a public dataset containing 7,470 pairs of images and corresponding diagnostics reports. Each study has one frontal and one lateral view and is associated with a report that consists of impression, findings, comparison and indication sections. We select sentences in the findings section as ground-truths. Since IU X-ray has incomplete age and sex information, we only use reason embedding for clinical feature embedding. Due to the dataset's small size, we filter tokens by a minimum frequency of 3, keeping 1,042 tokens.

Following the advice of expert radiologists who were asked to read the reports from the two datasets, we manually define two categories of image labels. The first category has 18 disease labels, such as "No Finding", "Cardiomegaly", "Airspace Opacity" and "Fibrosis". We carefully define the negation and uncertainty language patterns and adopt the CheXpert Labeler [19] to extract the labels. The second group has 32 description pattern labels, such as "Upper/Lower", "Left/Right", "Patchy" and "Blur", which are annotated as positive if they appear in the report, and negative otherwise. For both datasets, the data indexed by patients is randomly split into training, validation and testing sets with a ratio of 7:1:2, followed by [26, 20, 24]. Moreover, we use the MIMIC dataset to pre-train the image encoder in all experiments due to its large scale.

The common image captioning evaluation metrics, including BLEU [34], CIDEr [38], ROUGE [28] and METEOR [5], only focus on word-level fluency or recall, which are insufficient to evaluate generated medical reports. Thus, we propose a new metric called normalized Key Term Distance (**nKTD**). The aim is to judge whether or not the generated sentences contain all the observed diseases and their detailed descriptive information. We extract all the labels using the CheXpert Labeler [19] both from generated reports and ground-truths, as $\mathbf{b}_{ge}$ and $\mathbf{b}_{gt}$, and compute the Hamming distance which is defined as:

$$\mathcal{S}_{nKTD} = \frac{d_{hamming}(\mathbf{b}_{ge}, \mathbf{b}_{gt})}{N}, \quad (11)$$

where N denotes the number of labels. The smaller the score is, the more accurate the key findings contained in the generated reports.

4.2. Ablation Studies

4.2.1 Ablation Studies on Report Generation

To evaluate the proposed report generation model, five baselines are compared for analysis. **Effectiveness of Multi-Modality Semantic Attention (Ours-wo-MMSA):** MMSA is able to combine regional visual features with predicted key finding embeddings and clinical features to model correlations among them and extract high-level semantic context information. By detaching the MMSA, the global visual feature f_v is simply concatenated with the clinical feature. A fully-connected layer is configured to map the combined vector to a context vector for decoding. **Effectiveness of Description Pattern Labels (Ours-wo-DPL):** Previous methods only adopt disease labels to train the image encoder. To evaluate whether or not the additional description pattern labels can contribute to the accuracy of generated sentences, with more details on the detected diseases, we drop them and only adopt the disease labels in this baseline. **Effectiveness of Matching and De-**

Table 1. Quantative evaluation of automatic report generation. "wo" is the abbreviation of without. The best result of our model is shown in **red**, while the result of the best state-of-the-art method is shown in blue.

Dataset	Conference	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	ROUGE	METEOR	nKTD
MIMIC-CXR	CVPR 2015 [39]	CNN-RNN	0.303	0.198	0.135	0.090	0.879	0.296	0.149	0.256
	CVPR 2015 [6]	LRCN	0.316	0.207	0.137	0.091	0.861	0.295	0.146	0.238
	CVPR 2017 [32]	AdaAtt	0.306	0.202	0.137	0.089	0.883	0.298	0.150	0.246
	CVPR 2017 [36]	Att2in	0.309	0.204	0.136	0.090	0.885	0.299	0.149	0.232
	CVPR 2018 [42]	TieNet	0.329	0.215	0.138	0.093	0.993	0.294	0.153	0.217
	ACL 2018 [20]	Co-Attention	0.346	0.226	0.152	0.112	0.859	0.324	0.179	0.185
	NeurIPS 2018 [26]	HRGR-Agent	0.342	0.224	0.155	0.111	0.934	0.311	0.170	0.193
	IPMI 2019 [44]	IDCTF	0.347	0.229	0.156	0.115	0.916	0.320	0.176	0.179
	AAAI 2019 [24]	KERP	0.352	0.225	0.154	0.109	0.894	0.307	0.168	0.162
	MICCAI 2019 [47]	MvH+AttL+MC	0.355	0.228	0.157	0.113	0.907	0.321	0.174	0.154
	Ours	Ours-wo-MMSA	0.353	0.224	0.156	0.111	0.909	0.315	0.171	0.158
		Ours-wo-DPL	0.361	0.232	0.161	0.115	1.032	0.323	0.177	0.170
		Ours-wo-CF	**0.373**	0.239	0.168	0.121	**1.125**	0.319	0.175	0.117
		Ours-wo-$\mathcal{L}_{ISM}$	0.351	0.223	0.158	0.113	0.954	0.320	0.173	0.155
		Ours-wo-$\mathcal{L}_{DA}$	0.364	0.233	0.162	0.115	1.029	0.325	0.179	0.132
		Ours	0.372	**0.241**	**0.168**	**0.123**	1.121	**0.335**	**0.190**	**0.106**
IU X-Ray	CVPR 2015 [39]	CNN-RNN	0.309	0.208	0.137	0.090	0.115	0.274	0.157	0.233
	CVPR 2015 [6]	LRCN	0.358	0.214	0.142	0.096	0.198	0.283	0.151	0.219
	CVPR 2017 [32]	AdaAtt	0.313	0.210	0.138	0.092	0.113	0.273	0.159	0.225
	CVPR 2017 [36]	Att2in	0.316	0.211	0.139	0.092	0.114	0.274	0.158	0.216
	CVPR 2018 [42]	TieNet	0.389	0.252	0.195	0.147	0.324	0.319	0.180	0.198
	ACL 2018 [20]	Co-Attention	0.502	0.363	0.285	0.231	0.313	0.425	0.213	0.169
	NeurIPS 2018 [26]	HRGR-Agent	0.483	0.359	0.287	0.232	0.319	0.394	0.204	0.175
	IPMI 2019 [44]	IDCTF	0.498	0.362	0.289	0.234	0.317	0.413	0.209	0.157
	AAAI 2019 [24]	KERP	0.511	0.368	0.293	0.237	0.312	0.356	0.195	0.145
	MICCAI 2019 [47]	MvH+AttL+MC	0.518	0.374	0.296	0.240	0.315	0.411	0.198	0.137
	Ours	Ours-wo-MMSA	0.515	0.371	0.290	0.238	0.320	0.394	0.196	0.142
		Ours-wo-DPL	0.524	0.379	0.301	0.244	0.327	0.423	0.210	0.145
		Ours-wo-CF	0.533	**0.392**	0.312	0.250	0.337	0.412	0.205	0.104
		Ours-wo-$\mathcal{L}_{ISM}$	0.512	0.372	0.293	0.236	0.324	0.416	0.198	0.130
		Ours-wo-$\mathcal{L}_{DA}$	0.521	0.380	0.305	0.248	0.328	0.431	0.215	0.116
		Ours	**0.536**	0.391	**0.314**	**0.252**	**0.339**	**0.448**	**0.228**	**0.097**

scription Loss (Ours-wo-$\mathcal{L}_{ISM}$ and wo-$\mathcal{L}_{DA}$): Both the image-sentence matching loss and description accuracy loss are designed to push the generated sentences to match the query X-Ray image and contain correct key findings. We set two baselines by dropping the bi-directional max-margin ranking module and the generated report embedding classifier to evaluate the two designs, respectively. **Effectiveness of Clinical Features (Ours-wo-CF):** Clinical features contain some basic indications of relevant diseases, and are usually considered by radiologists for preliminary diagnosis. We also want to explore how much these can benefit our model. In this baseline, we detach the clinical feature input in the multi-attention model.

The bottom rows of each dataset in Table 1 show the comparison results. On the MIMIC dataset, without the multi-modality semantic attention design, the performance is largely decreased over all the evaluation metrics. To compute the nKTD for Ours-wo-DPL, we drop the description pattern labels in the training phase but still extract them in the testing phase for fair comparison. On average, compared to Ours, 3.2 more labels are wrongly predicted in total by Ours-wo-DPL, where 2.7 out of 3.2 are description pattern labels. Clinical features can moderately increase the results, particularly in terms of ROUGE and METEROR, which focus more on the recall of key findings. Both the bi-directional max-margin ranking module and the generated report embedding classifier can enhance the model, though the former is more effective.

For the ablation studies on the IU X-Ray dataset, the effectiveness of each designed module has been validated with similar result comparisons of the MIMIC datasets. Detaching either the multi-modality semantic attention module or topic-level image-sentence matching losses, largely reduces the performance. Training the image encoder with the extra description pattern labels under the constraint of the generated report embedding classifier, can also moderately contribute to the final generated reports. Besides, as evaluated by our proposed metric nKTD, only 4.85 out of 50 labels (key terms) are wrongly described in reports generated by our method, on average. Therefore, our model achieves consistent improvements on both the datasets.

4.2.2 Ablation Studies on Disease Prediction

The performance of report generation moderately depends on the accuracy of the intermediate key finding predictions in our model. Thus, we explore several specific designs to train the image encoder for better multi-label classification results. Three baselines are compared for this task. **Effectiveness of Self-Attention (Ours-wo-SA):** To better

Table 2. Evaluation of disease prediction on MIMIC. The best result is shown in **red**.

Disease	Ours	Ours-wo-SA	Ours-wo-RE	Ours-w-CF
Enlarged Cardiom.	**0.766**	0.753	0.748	0.765
Cardiomegaly	**0.856**	0.839	0.841	0.852
Lung Lesion	0.773	0.782	0.761	**0.785**
Airspace Opacity	0.845	0.841	0.836	**0.852**
Edema	0.943	0.926	0.935	**0.958**
Consolidation	0.909	0.912	0.915	**0.923**
Pneumonia	0.894	0.857	0.880	**0.908**
Atelectasis	0.889	0.876	0.868	**0.895**
Pneumothorax	0.799	**0.808**	0.804	0.806
Pleural Effusion	0.958	0.936	0.951	**0.966**
Pleural Other	0.848	**0.855**	0.831	0.853
Fracture	**0.837**	0.821	0.812	0.830
Emphysema	0.846	0.837	0.848	**0.851**
Hernia	0.754	**0.756**	0.752	0.753
Fibrosis	**0.955**	0.931	0.940	0.948
Spine Degen. Cha.	**0.778**	0.763	0.776	0.773
Support Devices	0.823	**0.829**	0.817	0.819
No Finding	0.805	**0.818**	0.782	0.793
mean	0.849	0.841	0.838	**0.852**

Table 3. Evaluation of description pattern labels prediction on the MIMIC dataset. The best result is shown in **red**.

Description Pattern	Ours	Ours-wo-SA	Ours-wo-RE	Ours-w-CF	Description Pattern	Ours	Ours -wo-SA	Ours-wo-RE	Ours -w-CF
upper	**0.788**	0.767	0.763	0.784	esophagus	0.776	0.773	0.771	**0.779**
lower	**0.817**	0.803	0.798	0.812	stomach	0.814	0.806	0.809	**0.816**
left	**0.859**	0.842	0.836	0.850	aorta	0.821	0.815	0.808	**0.825**
right	**0.816**	0.803	0.801	0.814	tortuous	**0.791**	0.782	0.774	0.787
patchy	**0.881**	0.876	0.879	0.878	diaphragm	0.754	**0.761**	0.750	0.742
bilateral	0.756	0.739	0.732	**0.759**	elevated	0.871	**0.874**	0.871	0.867
lateral	0.749	0.751	0.746	**0.752**	mitral	**0.728**	0.713	0.715	0.720
volume	0.837	0.835	**0.838**	0.830	multiple	0.784	**0.785**	0.777	0.782
small	**0.810**	0.802	0.797	0.808	solitary	-	-	-	-
interstitial	**0.903**	0.895	0.898	0.899	hypertension	0.837	0.845	0.832	**0.846**
basal	0.951	**0.953**	0.948	0.946	large	**0.815**	0.798	0.785	0.803
blur	0.618	**0.625**	0.612	0.605	diffuse	0.904	0.886	0.889	**0.907**
linear	0.778	0.774	**0.781**	0.775	central	**0.825**	0.808	0.804	0.812
peribronchial	0.860	0.858	0.852	**0.863**	coronary	**0.884**	0.877	0.879	0.882
density	0.821	**0.827**	0.819	0.820	calcification	0.760	**0.762**	0.757	0.755
enlarged	0.854	0.853	0.845	**0.858**	reticular	**0.943**	0.931	0.933	0.928
mean	**0.819**	0.813	0.809	0.816					

compute the visual attention in the multi-attention module, in addition to the semantic attention, we also build a self-attention purely based on the regional visual features. We evaluate how much this self-attention module improves the multi-label classification results, compared to the baseline without it. **Effectiveness of Report Embedding Supervision (Ours-wo-RE):** Since the labels are not annotated by domain experts but automatically extracted by the CheXpert labeler, they are not absolutely correct, nor keep all the important information from the original reports. The report embedding can compensate this drawback to some extent and is adopted to in addition to discrete labels to train the image encoder. Its effectiveness is also evaluated. **Effectiveness of Clinical Feature (Ours-w-CF):** The motivation behind exploiting clinical features to help report generation is our observation that it can benefit thorax disease classification. To validate this point, the encoded clinical feature is directly concatenated with the global visual feature, and then connected to the report embedding and multi-label classification supervision. Thus, improvement on disease prediction can also be achieved.

Following [41, 19], we adopt the area-under-the-curve (AUC) of the Receiver Operator Characteristic (ROC) to evaluate performance of identification over 18 disease labels, on the MIMIC dataset, in Table 2. The experiment is conducted on the MIMIC dataset due to its large size, which is more convincing. The slightly decreased average result obtained by Ours-wo-SA illustrates that most lesions can benefit from the self-attention design. Minor lung lesions, pneumothoraxes and normal findings are exceptions since abnormal visual patterns are not clear for these disease labels. The report embedding supervision facilitates the image encoder training over all the labels except pneumothorax, with a mean AUC increase of 1.1%. Moreover, the clinical features moderately enhance the performance, particularly for lesions in the lung area, since most clinical indications are related to lung lesions.

The results of description pattern labels prediction are also shown in Table. 3. (Please note 'solitary' is not available in the MIMIC dataset.) Compared with the mean AUC performance of key findings, the results of description pattern labels are 3.0% lower. The main reason is that definitions of those labels are usually ambiguous and noisy such as "basal", "blur", "multiple", and "diffuse". Moreover, our fully-functioned final model achieves the best results compared to all three baselines. Therefore, the improvements on the key label predictions inspired us to enhance our report generation model by such designs.

4.3. Comparison with State-of-the-arts

To further validate our method, we compare it with four state-of-the-art image captioning models and six medical report generation models, which are listed in the upper part of each dataset in Table 1. All the models adopt DenseNet-201 as the backbone for the visual feature encoder. For the reinforcement learning-based methods HRGR-Agent [26] and KERP [24], which require a template data pool to be built, sentences from the training set of each dataset that occur with high frequency are selected by a threshold as template candidates. Then, the candidates with similar meanings but various linguistic descriptions are further grouped into the same template. Thus, 132 and 29 sentence templates are selected for MIMIC-CXR and IU X-Ray, respectively. Moreover, we use the same 50 labels which are defined in our models for the medical concepts in MvH+AttL+MC [47]. As shown in Table 1, reinforcement learning does not demonstrated clear superiority over pure cross-entropy optimization. Our final model achieves consistent improvements over all the evaluation metrics. For the reports generated on MIMIC-CXR, our method makes 2.4 less wrong label predictions, on average, compared to the best performing prior work MvH+AttL+MC, which demonstrates that descriptions of key findings by our model are more complete and precise.

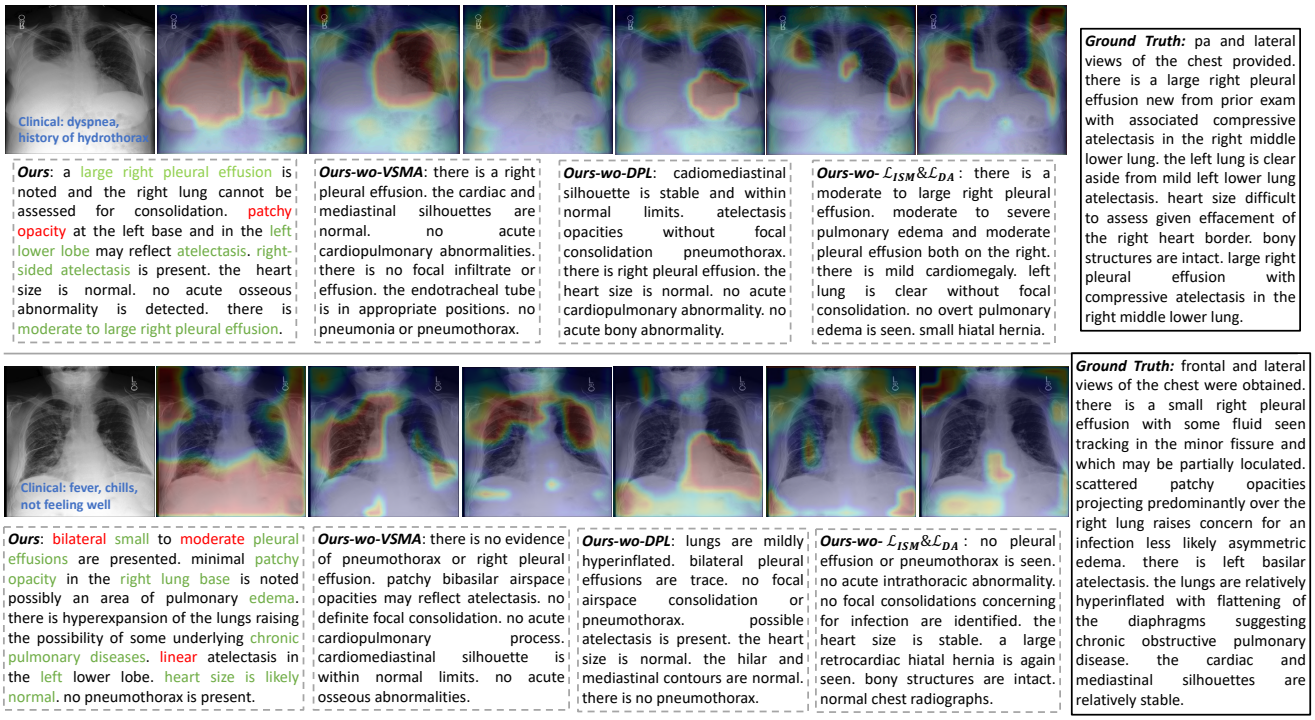

Figure 3. Illustration of generated sentences by our method and comparisons with baselines. For reports by **Ours**, the key findings correctly mentioned in the report are highlighted by green, and those wrongly described are in **red**. The blue texts are input clinical information.

Overall, we obtain the following observations and analysis. Conventional image captioning models CNN-RNN, LRCN, AdaAtt, and Att2in, get much lower results over all the metrics than the medical generation methods TieNet, Co-Attention, HRGR-Agent, IDCTF, KERP, and MvH+AttL+MC. This is because the report generators for medical images are proposed for predicting multiple sentences describing separate and different topics. The reinforcement learning-based methods HRGR-Agent and KERP use selected templates for a retrieval manner. However, the ROUGE results are not satisfactory compared to others due to the low recall of key medical terms. Although MvH+AttL+MC achieves the best performance among all the previous works since it combines the multi-view image features, attention learning and medical concepts, it doesn't consider matching generated sentences with the query X-Ray image for better description accuracy.

4.4. Qualitative Results

To better demonstrate the model performance, the reports generated by **Ours** are compared with ground-truths and some baselines in Fig. 3 (More results are shown in the supplementary file). Semantic attention maps at different topic steps of the sentence decoder are visualized as well. Compared with the ground-truths, the thorax diseases and their corresponding description patterns can usually be detected and correctly mentioned in our reports. For example, in the first case, pleural effusion is found in the scan, and it is also accurately described as "large" and "right". Atelectasis is also mentioned with its position "left lower lobe". However, inappropriate descriptions, such as "patchy opacity", sometimes occur, which are not observed in the scan. Moreover, examples of generated sentences by some baselines are also compared. Ours-wo-MMSA can detect very obvious abnormal regions but fails to detect unclear diseases. Most identified lesions are written in the reports by Ours-wo-DPL but they are not as correctly or fully described in detail as by **Ours**. To illustrate the effectiveness of the loss constraints on the sentence decoder, we drop the $\mathcal{L}_{ISM}$ and $\mathcal{L}_{DA}$ together for much clearer comparison. We find that Ours-wo-$\mathcal{L}_{ISM}$&$\mathcal{L}_{DA}$ sometimes wrongly describes the findings and tends to generate normal descriptions even when lesions are clearly present.

5. Conclusion

In this paper, we proposed a medical report generation method with a multi-modality semantic attention module and progressive decoder, optimized by image-sentence matching and description accuracy constraints. Extensive experiments showed that our method achieves significant improvements compared to the state-of-the-art methods.

Acknowledgements: This work was supported by the National Natural Science Foundation of China (62106043).

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on CVPR*, pages 6077–6086, 2018.

[2] Moitreya Chatterjee and Alexander G Schwing. Diverse and coherent paragraph generation from images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 729–744, 2018.

[3] Qingyu Chen, Yifan Peng, and Zhiyong Lu. Biosentvec: creating sentence embeddings for biomedical texts. *arXiv preprint arXiv:1810.09302*, 2018.

[4] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2015.

[5] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380, 2014.

[6] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *The IEEE Conference on CVPR*, June 2015.

[7] Deng-Ping Fan, Tao Zhou, Ge-Peng Ji, Yi Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Inf-net: Automatic covid-19 lung infection segmentation from ct images. *IEEE Transactions on Medical Imaging*, 39(8):2626–2637, 2020.

[8] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In *ECCV*, pages 15–29. Springer, 2010.

[9] Yang Feng, Lin Ma, Wei Liu, and Jiebo Luo. Unsupervised image captioning. In *The IEEE Conference on CVPR*, June 2019.

[10] Yunchao Gong, Liwei Wang, Micah Hodosh, Julia Hockenmaier, and Svetlana Lazebnik. Improving image-sentence embeddings using large weakly annotated photo collections. In *ECCV*, pages 529–545. Springer, 2014.

[11] Jiuxiang Gu, Jianfei Cai, Shafiq R. Joty, Li Niu, and Gang Wang. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In *The IEEE Conference on CVPR*, June 2018.

[12] Jiuxiang Gu, Shafiq Joty, Jianfei Cai, Handong Zhao, Xu Yang, and Gang Wang. Unpaired image captioning via scene graph alignments. In *The IEEE Conference on ICCV*, October 2019.

[13] Sebastian Guendel, Sasa Grbic, Bogdan Georgescu, Siqi Liu, Andreas Maier, and Dorin Comaniciu. Learning to recognize abnormalities in chest x-rays with location-aware dense networks. In *Iberoamerican Congress on Pattern Recognition*, pages 757–765. Springer, 2018.

[14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE ICCV*, pages 2961–2969, 2017.

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

[16] MD Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CSUR)*, 51(6):118, 2019.

[17] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on CVPR*, pages 4700–4708, 2017.

[18] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *The IEEE Conference on ICCV*, November 2019.

[19] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *arXiv preprint arXiv:1901.07031*, 2019.

[20] Baoyu Jing, Pengtao Xie, and Eric Xing. On the automatic generation of medical imaging reports. In *Proceedings of the annual meeting of the Association for Computational Linguistics*, pages 2577–2586, 2018.

[21] Alistair EW Johnson, Tom J Pollard, Seth Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr: A large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

[22] Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. A hierarchical approach for generating descriptive image paragraphs. In *Proceedings of the IEEE Conference on CVPR*, pages 317–325, 2017.

[23] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

[24] Christy Y Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In *AAAI*, 2019.

[25] Guang Li, Linchao Zhu, Ping Liu, and Yi Yang. Entangled transformer for image captioning. In *The IEEE Conference on ICCV*, October 2019.

[26] Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. Hybrid retrieval-generation reinforced agent for medical image report generation. In *Advances in Neural Information Processing Systems*, pages 1530–1540, 2018.

[27] Yehao Li, Ting Yao, Yingwei Pan, Hongyang Chao, and Tao Mei. Pointing novel objects in image captioning. In *The IEEE Conference on CVPR*, June 2019.

[28] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014.

[30] Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. Clinically accurate chest x-ray report generation. *arXiv preprint arXiv:1904.02633*, 2019.

[31] Jingyu Liu, Gangming Zhao, Fei Yu, Ming Zhang, Yizhou Wang, and Yizhou Yu. Align, attend and locate: Chest x-ray diagnosis via contrast induced attention network with limited supervision. In *The IEEE Conference on ICCV*, November 2019.

[32] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *The IEEE Conference on CVPR*, July 2017.

[33] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. In *Proceedings of the IEEE Conference on CVPR*, pages 299–307, 2017.

[34] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

[35] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

[36] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *The IEEE Conference on CVPR*, July 2017.

[37] Yunhang Shen, Rongrong Ji, Yan Wang, Yongjian Wu, and Liujuan Cao. Cyclic guidance for weakly supervised joint detection and segmentation. In *The IEEE Conference on CVPR*, June 2019.

[38] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on CVPR*, pages 4566–4575, 2015.

[39] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *The IEEE Conference on CVPR*, June 2015.

[40] Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In *The IEEE Conference on CVPR*, June 2019.

[41] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald 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 CVPR*, pages 2097–2106, 2017.

[42] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald 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 CVPR*, pages 9049–9058, 2018.

[43] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, pages 2048–2057, 2015.

[44] Yuan Xue and Xiaolei Huang. Improved disease classification in chest x-rays with transferred features from report generation. In *International Conference on Information Processing in Medical Imaging*, pages 125–138. Springer, 2019.

[45] Li Yao, Eric Poblenz, Dmitry Dagunts, Ben Covington, Devon Bernard, and Kevin Lyman. Learning to diagnose from scratch by exploiting dependencies among labels. *arXiv preprint arXiv:1710.10501*, 2017.

[46] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Hierarchy parsing for image captioning. In *The IEEE Conference on ICCV*, November 2019.

[47] Jianbo Yuan, Haofu Liao, Rui Luo, and Jiebo Luo. Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. In *MICCAI*, 2019.

[48] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*, 2018.

[49] Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan Yuille, and Daguang Xu. When radiology report generation meets knowledge graph. In *AAAI*, volume 34, pages 12910–12917, 2020.

[50] Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng. Deep supervised cross-modal retrieval. In *The IEEE Conference on CVPR*, June 2019.

[51] Yi Zhou, Xiaodong He, Lei Huang, Li Liu, Fan Zhu, Shanshan Cui, and Ling Shao. Collaborative learning of semi-supervised segmentation and classification for medical images. In *The IEEE Conference on CVPR*, pages 2079–2088, 2019.

[52] Yi Zhou, Lei Huang, Tianfei Zhou, and Ling Shao. Many-to-one distribution learning and k-nearest neighbor smoothing for thoracic disease identification. In *AAAI*, volume 35, pages 768–776, 2021.

[53] Yi Zhou, Tianfei Zhou, Tao Zhou, Huazhu Fu, Jiacheng Liu, and Ling Shao. Contrast-attentive thoracic disease recognition with dual-weighting graph reasoning. *IEEE Transactions on Medical Imaging*, 40(4):1196–1206, 2021.