

# A Self-boosting Framework for Automated Radiographic Report Generation

Zhanyu Wang  
University of Sydney

zhanyu.wang@sydney.edu.au

Luping Zhou  
University of Sydney

luping.zhou@sydney.edu.au

Lei Wang  
University of Wollongong

leiw@uow.edu.au

Xiu Li

Tsinghua Shenzhen International Graduate School

li.xiu@sz.tsinghua.edu.cn

## Abstract

*Automated radiographic report generation is a challenging task since it requires to generate paragraphs describing fine-grained visual differences of cases, especially for those between the diseased and the healthy. Existing image captioning methods commonly target at generic images, and lack mechanism to meet this requirement. To bridge this gap, in this paper, we propose a self-boosting framework that improves radiographic report generation based on the cooperation of the main task of report generation and an auxiliary task of image-text matching. The two tasks are built as the two branches of a network model and influence each other in a cooperative way. On one hand, the image-text matching branch helps to learn highly text-correlated visual features for the report generation branch to output high quality reports. On the other hand, the improved reports produced by the report generation branch provide additional harder samples for the image-text matching branch and enforce the latter to improve itself by learning better visual and text feature representations. This, in turn, helps improve the report generation branch again. These two branches are jointly trained to help improve each other iteratively and progressively, so that the whole model is self-boosted without requiring external resources. Experimental results demonstrate the effectiveness of our method on two public datasets, showing its superior performance over multiple state-of-the-art image captioning and medical report generation methods.*

## 1. Introduction

Everyday a large amount of medical imaging data are acquired, stored and examined in clinics. This has exerted mounting pressure to radiologists to analyse images and report the findings in time. Automated medical report generation is therefore in demand as it can reduce workload and

diagnostic errors as well as accelerate the clinic workflow.

Automated medical report generation is very challenging and it associates with a broader research topic of image captioning in computer vision. In image captioning, free-form text descriptions are generated to narrate the content of images. A basic deep learning model for image captioning follows the encoder-decoder structure [30], where the visual encoder extracts the visual features from images and the text decoder converts the visual features to text output. Research in this field focuses on advancing encoder and decoder, respectively, by employing carefully-designed attention mechanisms [22, 1, 26], relationship among image regions [40, 38], improved language models [4, 3], or reinforcement learning on language metric [26, 21]. A detailed review could be found in Section 2.

Despite recent achievements in image captioning, when directly applying image captioners for medical report generation, there is often a visible performance decline. This is because compared with generic images, radiographic images are more similar to each other and fine-grained visual differences, such as the findings of clinic importance, need to be narrated. This requires further tightening the visual and text representations. In addition to linking image and text by attention mechanism, some medical report generation methods [10, 41] additionally train classifiers to learn visual representations by predicting tags of medical reports or disease-class labels. This often requires additional annotations and external medical datasets or knowledge, which are task-specific and often unavailable, limiting the generalization of these methods. Moreover, either tags or disease classes only sparsely cover the reports' information, leading to relatively loosely correlated visual and text features.

To bridge this gap, in this paper, we propose a self-boosting framework to promote radiographic report generation. It utilizes an auxiliary task to predict the match of an image-report pair (i.e., image-text matching) by explicitly learning strongly correlated visual and text features. These

features could better serve the fine-grained recognition task in radiographic report generation. More importantly, the auxiliary image-text matching task is deeply coupled with the main report generation task through our proposed self-boosting framework so that these two tasks could mutually and progressively boost each other via the cooperative interactions between them. Specifically, these two tasks are built as two branches of a neural network model and jointly trained. During the training iterations, the image-text matching branch provides better features to the report generation branch to generate ground-truth-like reports, and these improved reports, in turn, serve as additional harder samples to push the image-text matching branch to become stronger by learning better features. In this way, the whole network is gradually self-boosted and improves the ultimate report generation performance.

Our main contributions are summarized as follows.

First, we utilize an auxiliary task of image-text matching to help learn text-correlated visual features that could help capture the fine-grained visual differences of diagnostic importance. This improves the main task of report generation without requiring additional annotations, external medical datasets, or external task-specific knowledge.

Secondly and more importantly, we propose a self-boosting framework that leverages the cooperation between image-text matching and report generation to mutually boost each other progressively. These two tasks are tightly coupled and jointly trained, while the stronger features learned by image-text matching help improve report generation, and the improved reports, as additional harder samples, in turn enforce image-text matching to continue improving feature learning so that the finer mismatch between an image and its generated report could be identified.

Third, additionally utilizing image-text matching also allows us to learn an effective text feature extractor, which is used to evaluate the feature similarity between the generated reports and the ground-truth, providing a new loss term to further promote the report generation.

Fourth, our approach shows promising performance on two benchmarks, generating reports from both classic Chest X-ray images and CT images with COVID-19. It outperforms multiple state-of-the-art methods in image captioning and medical report generation.

## 2. Related Work

**Image Captioning** Natural image captioning task aims at automatically generating sentence description for the given image. A broad collection of methods was proposed in the last few years [11, 30, 6, 34, 42, 39, 22, 26, 21, 1, 3, 44]. Most of them adopt the encoder-decoder architecture, employing CNN as the encoder to extract the visual feature of the image and RNN as the decoder to produce image description. Among these methods, *Show-Tell* [30] is the

most canonical neural image captioning model, which provides an end-to-end framework by feeding the image features extracted by CNN as the input of the LSTM [8] to produce image captions. On the basis of this framework, inspired by human brain's attention, several methods have proposed the attention mechanism [34, 42, 39, 22], allowing the model to automatically learn to fix its gaze on salient objects while generating the corresponding words in the output sequence. For example, in [22] an adaptive attention model was proposed to automatically switch the focus between visual signals and the language model; in [1] a combined bottom-up and top-down attention mechanism calculated attention at the level of objects and salient regions. Also, in a very recent work in [44], it was proposed to learn attention from an additional image-text matching task and use it to regularize the image captioning task. Although this work is close to our work in utilizing image-text matching for image captioning, they are substantially different as explained in Section 3.5. The superiority of our method over [44] has also been verified in our experiment. In addition to developing attention models, other image captioning works also explore how to advance the image encoder and text decoder. To improve the image encoder, the methods [40, 38] explicitly considered the visual relationship of the detected image regions by constructing graphical convolutional network (GCN) or scene graph. To improve the text decoder, in [14] a hierarchically structured RNN was developed to cater for paragraph generation. Recently, Transformer model [28] has been proposed to improve the limited representation power of RNN, and has been used in image captioning tasks to replace RNN as the text decoder [3]. Moreover, recently Reinforcement Learning has also been introduced to train the non-differentiable captioning metrics [26, 21] for improvement.

**Image-text matching** The goal of image-text matching task is to measure the visual-semantic similarity between a text and an image. One of the most common approaches is to project the image and text features into a joint semantic space to compute their similarity by cosine distance. Most methods in this field are dedicated to improving feature extraction techniques and roughly fall into two categories: global representation based [13, 32, 7] and local representation based [15, 20, 31]. Among them, SCAN [15] is a representative attention-based method for local representation, which was also adopted by [44] to facilitate captioning.

**Medical report generation** Due to the characteristics of medical reports, many existing methods are based on the hierarchical structured LSTM network [10, 43, 41] to generate finer detailed text information about the input radiographic images. Jing et al [10] proposed a multi-task hierarchical model with co-attention to automatically predict keywords and generate long paragraphs. Yin et al [41] proposed a topic matching mechanism to project the topic gen-

erated by the sentence RNN and its corresponding ground-truth into the same embedding space, so as to make them share the same semantics. In addition, Xue et al [36, 35] proposed a different network structure which includes a sentence generative model and a recurrent paragraph generative model, utilizing the generated sentence to produce next sentence. Another type of pipeline is called “Describe and Conclude” [9], which first generates the findings and then produces the impression by summarizing the information from the generated findings.

### 3. Method

In this section, we first give an overview of the proposed model. Then two important branches of our model, i.e., image-text matching and medical report generation, are presented. Following that, we describe how these two branches achieve joint training and mutual improvement.

#### 3.1. Overview

A typical radiographic report consists of a conclusive “impression” part and a descriptive “finding” part. Taking a medical image as input, our model generates a long-paragraph diagnostic report with both parts. As mentioned in Section 1, our model improves report generation by taking advantage of the interactions between the main task of report generation and the auxiliary task of image-text matching. Our model is trained in a self-boosting manner.

Fig. 1 provides an overview about our framework. In general, our model consists of two entangled branches: medical report generation (main task) and image-text matching (auxiliary task). As for the main task, our report generation branch (RG-branch) follows the encoder-decoder architecture. The visual encoder detects image regions with an unsupervised method, extracts regional features using CNN, and refines these visual features based on the regional relationships learnt through self-attention. The text-decoder takes the refined visual features as input, passes them through a hierarchical LSTM model catering for both the topic-level and word-level decoding, and generates the final paragraphical report. As for the auxiliary task, taking the image-report pairs as input, our image-text matching branch (ITM-branch) consists of an image encoder and a report encoder, learning the visual and the text features, respectively, in order to predict if the input image and report match each other.

The two branches interact with each other in three ways. First, they share the visual encoder, so that the text-correlated visual features learned by ITM-branch could also be utilized to improve the RG-branch to generate high-quality reports. Second, the report encoder learnt by the ITM-branch is also utilized by the RG-branch to evaluate and minimize the feature-level loss between the generated and the ground-truth reports. Third, during the training,

the improved reports generated by RG-branch are passed to the ITM-branch as additional harder samples, which enforces the ITM-branch to improve itself by enhancing the feature learning. These three interactions last for the training course, making the whole model gradually boost itself. At inference stage, only RG-branch is used, which receives a test image and generates the corresponding report.

#### 3.2. Report Generation (RG) Branch

The ultimate target of our model is to generate ground-truth-like diagnostic reports from radiographic images. Therefore, report generation is our main task. Like most image-captioning methods, our RG-branch consists of a visual encoder and a successive text decoder, with enhanced characteristics. They are elaborated as follows.

**Visual encoder** Although some generic image captioners use Faster R-CNN [25] to encode image regions at object-level, existing medical report generation methods extract visual features from the whole image rather than image regions, due to the lack of supervised information. Differently, given an input image of size  $3 \times H \times W$ , we employ the selective search algorithm to generate region proposals unsupervisedly, and then refine these region proposals by: 1) excluding regions smaller than 2000 pixels; 2) excluding regions where the background ratio exceeds 70%. This leads to roughly  $M = 30$  regions for one image. After that, we adopt ResNet-101 to produce a set of vectors  $v_1, \dots, v_M \in \mathbb{R}^D$  with a dimension  $D = 2048$  for each region. Compared with image-level features, using regional features can “look closer” to the image, thus better suits the fine-grained pattern description. It is found that modeling region relationship is beneficial for visual representation. Unlike some image-captioning methods [40, 38] that rely on the Visual Genome dataset to learn GCN or scene graph for generic images, we employ self-attention [28] to learn the relationships among regions and transform regional visual features  $\mathbf{V}$  into relationship-enhanced ones  $\mathbf{V}_h$  for refinement. These relationship-enhanced regional features are then passed through a region pooling layer to generate the pooled vector  $\mathbf{v}_p$ . We further input the pooled vector  $\mathbf{v}_p$  through a 2048-dimensional fully-connected layer to produce the final representation:  $\mathbf{I}_f = \mathbf{W}_f \mathbf{v}_p + \mathbf{b}_f$ . The output of the visual encoder  $\mathbf{I}_f$  is then sent to the text-decoder in RG-branch for report generation, as well as to ITM-branch for predicting image-text matching.

**Text decoder** To generate paragraphical report, we adopt a hierarchical structured LSTM, which includes a sentence LSTM and a word LSTM [14]. The sentence LSTM takes the visual features as input, produces a *topic vector* for each sentence in the report paragraph, and determines when to stop generating the topic vector. Those topic vectors are then fed into the word LSTM to produce the sentences. More details are as follows.

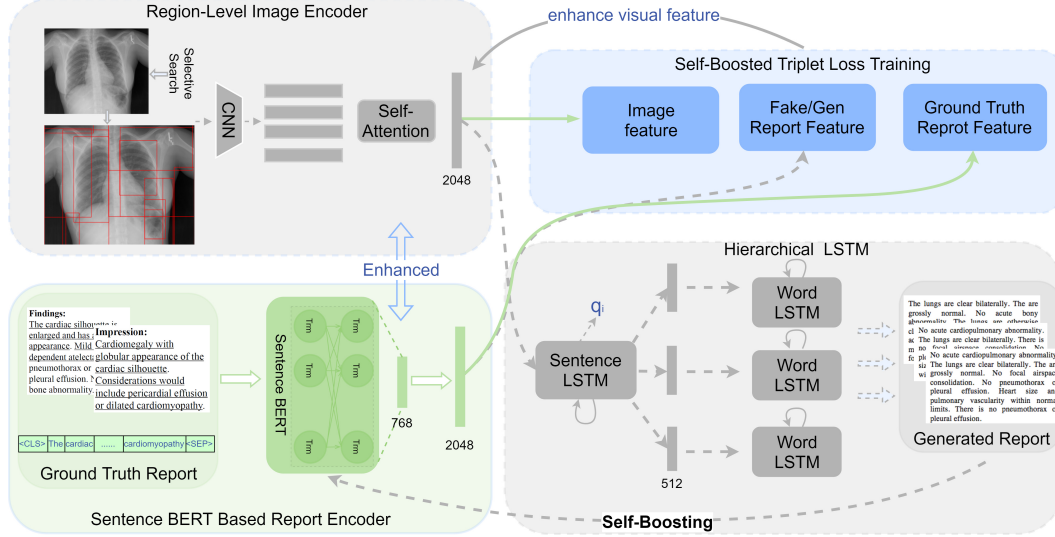


Figure 1. An overview of the proposed framework comprising of a medical report generation branch (RG-branch, the gray part) and an auxiliary image-text matching branch (ITM-branch, the green part). RG-branch consists of two components: a region-level image encoder (shared with ITM-branch) to extract visual features from medical images, and a hierarchical LSTM to generate diagnose reports. ITM-branch consists of an image encoder (shared with RG-branch) and a report encoder based on Sentence-BERT. It compares the similarity of visual and text features for match. The two branches are designed to have cooperative interactions during training. ITM-branch provides RG-branch text-correlated visual features for report generation and its report encoder for measuring the feature level loss of the generated reports. In turn, with the improved generated reports, RG-branch supplies ITM-branch harder and harder samples to force the latter to evolve for better feature learning. In this way, the whole model is self-boosted.

The sentence LSTM is a single layer LSTM with a hidden size  $H = 512$ . It takes the image feature  $\mathbf{I}_f$  as input, and in turn generates a sequence of hidden states  $\mathbf{h}_1 \cdots \mathbf{h}_N \in \mathbb{R}^H$ , one for each sentence in the paragraph:  $\mathbf{h}_{i+1}, \mathbf{c}_{i+1} = \text{LSTM}(\mathbf{I}_f, (\mathbf{h}_i, \mathbf{c}_i))$ ,  $i \in \{0, \dots, N-1\}$ , where  $\mathbf{c}_i$  denotes the memory cells of LSTM, and  $N$  is the number of sentences in the paragraph. Based on the hidden state  $\mathbf{h}_i$ , a topic vector  $\mathbf{t}_i \in \mathbb{R}^P$  is obtained as:  $\mathbf{t}_i = \tanh(\mathbf{W}_t \mathbf{h}_i + \mathbf{b}_t)$ ,  $i \in \{1, \dots, N\}$ . In addition, the hidden state  $\mathbf{h}_i$  is also mapped from  $H$  dimension to two dimension by a linear projection to produce a probability  $q_i$  through a sigmoid function, determining whether the  $i$ th sentence is the last sentence in the paragraph.  $q_i$  can be learnt by minimizing the cross entropy loss for a binary classification of being continued or not. In our work, when the “impression” part is available in the reports, we further use this information to guide the generation of the topic vectors, since “impression” implies a global topic about the report. Specifically, a single layer LSTM is first trained to generate “impression”, and then the attained weights are utilized to initialize the weights of the sentence LSTM.

The word LSTM is also a single layer LSTM with hidden size  $H = 512$  and the initial hidden and cell states of the word LSTM are set to zero. It takes the topic vectors produced by the sentence LSTM as the input and generates detailed sentences. Given the  $i$ th topic  $t_i$ , the hidden and cell states of the word LSTM are updated by  $\mathbf{h}_{i+1}, \mathbf{c}_{i+1} = \text{LSTM}(\mathbf{s}_{j-1}^{(i)}, (\mathbf{h}_i, \mathbf{c}_i))$ , where  $j \in \{0, \dots, M+2\}$ , and  $M$

is the number of word in  $i$ th sentence. It should be noted, when  $j = 0$ ,  $\mathbf{s}_{-1}^{(i)}$  represents the topic vector  $\mathbf{t}_i$ ; when  $j > 0$ ,  $\mathbf{s}_j^{(i)}$  represents the word embedding of the  $j$ th word in the  $i$ th sentence. In particular,  $\mathbf{s}_0$  and  $\mathbf{s}_{M+1}$  represent the word embedding of the special token  $\langle start \rangle$  and  $\langle end \rangle$  respectively. After obtaining the hidden state of word LSTM, the probability  $p_{ij} = \text{Softmax}(\mathbf{W}_p \mathbf{h}_j + \mathbf{b}_p)$  can be learnt, predicting the probability of  $j$ th word in  $i$ th sentence, where  $i \in \{1, \dots, M\}$ ,  $j \in \{1, \dots, N\}$ .

**Loss function** Most image captioning and medical report generation methods construct the loss functions evaluating the difference between the generated and the ground-truth reports at the word-level (cross-entropy based loss). In this paper, we further argue that a generated report with high-quality should also be similar to the ground-truth at the semantic feature levels. That is, they should stay similar after they are passed through a text feature extractor producing features corresponding to high-level semantics. In our work, such a text feature extractor could be readily achieved by utilizing the report encoder learnt in the auxiliary ITM-branch, which is well trained to effectively encode the most critical text features correlated to the visual contents. Specifically, given  $K$  reports, the loss function used in our RG-branch consists of two terms as follows:

$$\mathcal{L}_{RG} = \lambda_{CE} \sum_{k=1}^K \mathcal{L}_{CE}^{(k)} + \lambda_{feat} \sum_{k=1}^K \mathcal{L}_{feat}^{(k)}, \quad (1)$$

where  $\lambda_{CE}$  and  $\lambda_{feat}$  are hyper-parameters balancing the

two terms.

The first loss term  $\mathcal{L}_{CE}^{(k)}$  is the cross-entropy loss used to measure the word-level difference for the  $k$ th report. Since we use a hierarchical LSTM as the text decoder, the  $\mathcal{L}_{CE}^{(k)}$  is the sum of the cross-entropy loss  $\mathcal{L}_S^{(k)}$  of the sentence LSTM and the cross-entropy loss  $\mathcal{L}_W^{(k)}$  of the word LSTM. The loss  $\mathcal{L}_S^{(k)} = -\sum_{i=0}^N [l_i^{(k)} \cdot \log(q_i^{(k)}) + (1 - l_i^{(k)}) \cdot \log(1 - q_i^{(k)})]$ , where  $q_i^{(k)}$  and  $l_i^{(k)}$  are the predicted and the ground-truth labels, respectively, indicating if the  $i$ th sentence is the last sentence in the  $k$ th report. The loss  $\mathcal{L}_W^{(k)} = -\sum_{i=1}^N \sum_{j=1}^M y_{ij}^{(k)} \log(p_{ij}^{(k)})$ , where  $y_{ij}^{(k)}$  refers to the ground-truth and  $p_{ij}^{(k)}$  refers to the prediction of the  $j$ th word in the  $i$ th sentence within the  $k$ th report.

The second loss term  $\mathcal{L}_{feat}^{(k)} = \|\Phi^{ITM-R}(\hat{\mathbf{T}}^{(k)}) - \Phi^{ITM-R}(\mathbf{T}^{(k)})\|_2$  measures the difference at the feature level, where  $\hat{\mathbf{T}}^{(k)}$  and  $\mathbf{T}^{(k)}$  are the predicted paragraph and the ground-truth for the  $k$ th report, respectively. Here the mapping  $\Phi^{ITM-R}(\cdot)$  denotes the embedding function of the report encoder in ITM-branch (see Section 3.3 for more details), which is highly nonlinear.

In addition, following literature [26], we also apply reinforcement learning with CIDER as the reward to further boost the performance of our RG-branch.

### 3.3. Image-Text Matching (ITM) Branch

To promote the main task of report generation, we also introduce an auxiliary task of image-text matching to learn highly correlated visual and text features. Our ITM-branch takes an image-report pair as input and tells whether the image and the report match. It consists of an image encoder and a report encoder, detailed as follows.

**Image encoder** Our ITM-branch shares its image encoder with our RG-branch. As mentioned above, the image encoder extracts visual features from regions, refines these features with regional relationship, and outputs the image representation on top of an additional region pooling. It is worth mentioning that, sharing the image encoder between ITM-branch and RG-branch enables our model to both learn image-text correlated features and fine-tune these features for the main report generation task.

**Report encoder** We considered BERT’s [4] model as the report encoder, which is a multi-layer bidirectional Transformer [28] and has set a new state-of-the-art performance on various NLP tasks. However, we do not directly use the vanilla BERT. As mentioned, the report encoder in ITM-branch is also employed by RG-branch to compare the feature similarity between the generated and the ground-truth reports. Taking this into account, we adapt a Sentence-BERT [24] model to constructing our report encoder, because Sentence-BERT employed a siamese and

triplet network structure [27] for BERT, making it more suitable to derive semantically meaningful sentence embeddings. Sentence-BERT has been well pre-trained on two large broad-coverage corpus [2, 33] for various textual similarity tasks. Moreover, on top of the pretrained Sentence-BERT, we stack another fully connected layer to fine-tune the text features towards our image-text matching task and generate a text feature embedding  $\mathbf{T}_f$  that has the same dimension as the image feature embedding  $\mathbf{I}_f$  output by the image encoder. In sum,  $\mathbf{T}_f = \Phi^{ITM-R}(\mathbf{T}) = \sigma(\mathbf{W}_R \times \text{Sent-Bert}(\mathbf{T}))$ , where  $\sigma(\cdot)$  is the sigmoid function.

**Loss function** Based on the image feature embedding  $\mathbf{I}_f$  and the text feature embedding  $\mathbf{T}_f$ , we minimize a triplet loss similar to that in [7, 15] as follows.

$$\mathcal{L}_{match} = [\alpha - S(\mathbf{I}_f, \mathbf{T}_f) + S(\mathbf{I}_f, \bar{\mathbf{T}}_f)]_{++} + [\alpha - S(\mathbf{I}_f, \mathbf{T}_f) + S(\bar{\mathbf{I}}_f, \mathbf{T}_f)]_{+} \quad (2)$$

where  $S(\mathbf{I}_f, \mathbf{T}_f) = \frac{\mathbf{I}_f \cdot \mathbf{T}_f}{\|\mathbf{I}_f\| \cdot \|\mathbf{T}_f\|}$ , measuring the cosine similarity between the image and the text features. The symbol  $[\cdot]_{+}$  denotes the function  $[x]_{+} = \max(x, 0)$ ; and  $\alpha$  serves as a margin parameter.  $\bar{\mathbf{T}}_f$  and  $\bar{\mathbf{I}}_f$  represent the hardest negative samples [7] in a mini-batch for the positive pair  $(\mathbf{I}_f, \mathbf{T}_f)$ . Minimizing this triple loss requires that the distance from  $\mathbf{I}_f$  to  $\mathbf{T}_f$  be smaller than that from  $\mathbf{I}_f$  to  $\bar{\mathbf{T}}_f$  by a margin  $\alpha$ , and vice versa.

Moreover, it is noticed that negative mining [7] has a crucial impact on the performance of the image-text matching. In our model, we utilize the generated reports from RG-branch as hard negative samples to further improve ITM-branch. As the quality of the generated reports improves over iterations, they become more and more similar to the ground-truth ones, and therefore harder and harder to be differentiated by ITM-branch. When they are used to gradually join the training of image-text matching task, ITM-branch will be enforced to enhance its feature learning so that even finer mismatch between the image and the report could be identified. For this purpose, we additionally introduce a self-boosted triplet loss as follows:

$$\mathcal{L}_{match-gen} = [\alpha - S(\mathbf{I}_f, \mathbf{T}_f) + S(\mathbf{I}_f, \mathbf{T}_f^g)]_{++} + [\alpha - S(\mathbf{I}_f, \mathbf{T}_f^g) + S(\mathbf{I}_f, \bar{\mathbf{T}}_f)] \quad (3)$$

where  $\mathbf{T}_f^g$  refers to the text embedding of the generated report from RG-branch and other symbols remain the same as used in Eqn. 2. Minimizing this self-boosted triplet loss enforces ITM-branch to learn effective feature embedding so that in the embedding space, the generated report  $\mathbf{T}_f^g$  stays closer to its corresponding image  $\mathbf{I}_f$  than other reports  $\bar{\mathbf{T}}_f$ ; at the same time  $\mathbf{T}_f^g$  is still farther than its ground-truth  $\mathbf{T}_f$  to the image  $\mathbf{I}_f$  in order to enforce ITM-branch to differentiate the generated and the ground-truth reports. During our

training, the  $\mathcal{L}_{match-gen}$  is only employed after  $k$  number of epochs since the generated reports from RG-branch are of poor quality at the early iterations and they cannot assist as hard samples.

### 3.4. Self-boosted Training

The overall objective function of our model combines the loss terms from both RG-branch and ITM-branch:

$$\mathcal{L}_{all} = \lambda_{CE}\mathcal{L}_{CE} + \lambda_{feat}\mathcal{L}_{feat} + \lambda_{ITM}\mathcal{L}_{ITM}, \quad (4)$$

where

$$\mathcal{L}_{ITM} = \begin{cases} \mathcal{L}_{match}, & epoch \in (0, k] \\ \mathcal{L}_{match-gen}, & epoch \in (k, N] \end{cases} \quad (5)$$

The hyper-parameters  $\lambda_{CE}$ ,  $\lambda_{feat}$ , and  $\lambda_{ITM}$  are simply set to balance the three loss terms, and their values are given in Section 4.

To simplify the optimization, we update RG-branch and ITM-branch in an alternative way. That is, in some iterations we fix RG-branch and learn ITM-branch, and in next iterations we fix ITM-branch and learn RG-branch. This pattern is repeated throughout the whole training procedure. Please note that, as the two branches are highly entangled in our model, the update of ITM-branch will affect the visual encoder and the loss of RG-branch; while the update of RG-branch will affect the image encoder of ITM-branch and generate harder samples to the latter. In this way, our training ends up with a self-boosting manner: the improvement of ITM-branch enhances RG-branch and the improvement of RG-branch in turn pushes ITM-branch to evolve.

At the first glance, the interaction between our RG-branch and ITM-branch resembles that observed in Generative Adversarial Networks (GANs). Our RG-branch is like the generator that produces realistic reports while our ITM-branch is like the discriminator that differentiates the generated and the ground-truth reports. However, we would like to point out that our self-boosting mechanism is essentially different from the adversarial training in GANs. First, our RG-branch and ITM-branch do not compete but help with each other. On the contrary, GANs play the two-player min-max game between the generator and the discriminator so that the generator minimizes the reward of the discriminator. Such adversarial loss is not used in our approach. Second, in most GANs (if not all), the generator and the discriminator are two separated networks linked by the adversarial loss, while our RG-branch and ITM-branch share their visual/image encoder as one way of communication.

### 3.5. Remarks

It is noticed that in a very recent work [44], the image-text matching was also used to help image captioning, how-

ever, in a substantially different way as ours. In [44] the attentions pretrained by image-text matching are used to regularize the attentions used in report generation, while we communicate the visual and text features between the two tasks throughout the training. Moreover, in [44] the image-text matching and report generation are two separate steps and they are not tightly coupled and interact with each other, which is in a sharp contrast to our self-boosting framework. We also experimentally compare these two methods in Section 4. Our results clearly exhibit the superiority of our learning strategy over that of [44] by communicating both the features and the generated samples between the two tasks for progressive mutual improvement.

## 4. Experiments

### 4.1. Datasets

In this experiment, two datasets are used for the performance evaluation. One (IU-Xray) represents the most widely used benchmark for medical report generation, while the other (COV-CTR) is employed to validate model performance on the newly discovered disease COVID-19.

**IU-Xray** Indiana University Chest X-ray Collection (IU-Xray) [5] is the most widely used publicly accessible dataset in medical report generation task. It contains 3,955 fully de-identified radiology reports, each of which is associated with a frontal and/or lateral chest X-ray images, and 7,470 chest X-ray images in total. Each report is comprised of several sections: Impression, Findings and Indication etc. In this work, we filtered out reports without two complete image views or without complete sections of “findings” and “impression”, resulting in a smaller dataset with 3195 reports associated with 6390 images. We tokenized all the words in “findings” and “impression” and obtained 2,076 unique words. In addition, two special tokens,  $\langle start \rangle$  and  $\langle end \rangle$ , are added to indicate the start and the end of a sentence. We randomly picked 311 (10%) reports to form the test set, which is comparable to the works in the literature [10, 36]. All evaluations are done on the test set.

**COV-CTR** COVID-19 CT Report dataset (COV-CTR)[17] contains lungs CT images paired with their corresponding diagnostic chinese reports, in which the lungs CT images are collected by [37] during the outbreak time of COVID-19 and Li et al.[17] provided the corresponding diagnostic report, constructing a new medical report dataset. The COV-CTR dataset includes 728 images in total, in which 349 for COVID-19 and 379 for Non-COVID. We used the “jieba” software tool<sup>1</sup> to tokenize all report words, leading to 328 unique words or phrases in total. For a fair comparison, we follow [17] to randomly split the data into training, validation, and test sets with the ratio 7:1:2.

<sup>1</sup><https://github.com/fxsjy/jieba>

Dataset	Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	CIDEr
IU-Xray	Show-Tell [30]	0.346	0.214	0.141	0.095	0.320	0.239
	Att2in [26]	0.399	0.249	0.172	0.126	0.321	0.341
	AdaAtt [22]	0.436	0.288	0.203	0.150	0.354	0.265
	Topdown [1]	0.375	0.237	0.167	0.123	0.319	0.336
	Transformer[28]	0.422	0.264	0.177	0.120	0.338	0.268
	M2transformer[3]	0.463	0.318	0.214	0.155	0.335	0.349
	Grounded [44]	0.446	0.301	0.237	0.176	0.343	0.395
	CoAtt <sup>†</sup> [10]	0.455	0.288	0.205	0.154	<b>0.369</b>	0.277
	HGRG-Agent <sup>†</sup> [18]	0.438	0.298	0.208	0.151	0.322	0.343
	KERP <sup>†</sup> [16]	0.482	0.325	0.226	0.162	0.339	0.280
Ours	<b>0.487</b>	<b>0.346</b>	<b>0.270</b>	<b>0.208</b>	0.359	<b>0.452</b>	
COV-CTR	Show-Tell [30]	0.665	0.613	0.578	0.549	0.684	0.631
	Att2in [26]	0.674	0.614	0.574	0.542	0.588	0.686
	AdaAtt [22]	0.689	0.636	0.596	0.563	0.709	0.729
	Topdown [1]	0.705	0.663	0.622	0.586	0.720	0.968
	Transformer[28]	0.683	0.636	0.590	0.558	0.719	0.721
	M2transformer[3]	0.733	0.662	0.620	0.582	0.750	1.289
	Grounded [44]	0.753	0.708	0.665	0.627	0.776	1.381
	CoAtt <sup>†</sup> [10]	0.709	0.645	0.603	0.552	0.718	0.672
	Vision-BERT <sup>†</sup> [4]	0.710	0.653	0.606	0.558	0.747	0.684
	ASGK <sup>†</sup> [17]	0.712	0.659	0.611	0.570	0.746	0.680
	Ours	<b>0.810</b>	<b>0.766</b>	<b>0.721</b>	<b>0.679</b>	<b>0.790</b>	<b>2.371</b>

Table 1. Comparison on IU-Xray (upper part) and COV-CTR datasets (lower part). <sup>†</sup> indicates the results are quoted from the published literature. Specifically, we quote the results from [16] for IU-xray, and from [17] for COV-CTR. For other methods in comparison, their results are obtained by re-running the publicly released codes on these two datasets using the same training-test partition as our method.

## 4.2. Experimental Settings

**Evaluation Metrics** To evaluate the quality of the generated text report, we utilize the widely used BLEU scores[23], ROUGE-L[19] and CIDEr[29] as evaluation metric. We compute those metrics by the standard image captioning evaluation tool<sup>2</sup>.

**Implementation Details** We use the paired data as the input for IU-Xray dataset and concatenate the features of frontal and lateral images. The margin parameter  $\alpha$  in Eqn. 2 and 3 is set to 0.2 and  $\lambda_{CE}, \lambda_{feat}, \lambda_{ITM} = 2, 10, 1$  in Eqn. 4, respectively. The word embedding size and topic vector size were respectively set to 300 and 512. We train our model using Adam optimizer [12] with mini-batch size of 16. The learning rate are set to be 0.0001 and 0.0002, respectively, for RG-branch and ITM-branch with a total 30 epochs. We set  $k = 10$  in Eqn. 3, training ITM branch with the regular triplet loss for the first 10 epochs and the self-boosted triple loss for the rest.

## 4.3. Results and Discussion

**Comparison with SOTA** We compare our model with 7 state-of-the-art image captioning methods. These include the classic model Show-tell [30], different attention-based methods including AdaAtt [22], Att2in [1], and Topdown [26], methods using advanced NLP models including

Dataset	Model	B@4	Rouge	CIDEr
IU-Xray	baseline	0.128	0.307	0.359
	baseline+ITM	0.155	0.321	0.372
	baseline+ITM+FLL	0.169	0.330	0.391
	baseline+ITM+STL	0.176	0.345	0.418
	Ours-RL	0.193	0.352	0.426
	Ours	<b>0.208</b>	<b>0.359</b>	<b>0.452</b>
COV-CTR	baseline	0.573	0.659	0.722
	baseline+ITM	0.611	0.704	1.243
	baseline+ITM+FLL	0.637	0.729	1.617
	baseline+ITM+STL	0.658	0.761	1.953
	Ours-RL	0.672	0.783	2.014
	Ours	<b>0.679</b>	<b>0.790</b>	<b>2.371</b>

Table 2. Ablation studies. “Baseline” refers to RG-branch only. “ITM”, “FLL”, “STL” and “RL” stand for “Image-Text Matching”, “Feature Level Loss” ( $\mathcal{L}_{feat}$  in Eqn. 1), “Self-boosted Triplet Loss” (Eqn. 3), and “Reinforcement Learning”, respectively. B@4 is the BLEU score uses 4-grams.

Transformer [4] and M2transformer [3], and the very recent method Grounded [44] that also takes advantage of the image-text matching task to improve image captioning but in a completely different manner as ours. For these methods, we download the codes released publicly and re-run them on these two datasets with the same experiment setting as ours. Moreover, we also compare with 3 state-of-the-art medical report generation models: CoAtt [10], HGRN-

<sup>2</sup><https://github.com/tylin/coco-caption>



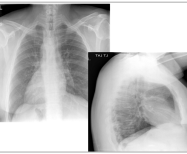
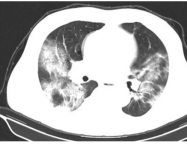
Input Images	Ground Truth	Show-Tell	Grounded	Ours
	<b>Impression:</b> No acute cardiopulmonary finding. <b>Findings:</b> <b>The heart size and cardio mediastinal silhouette are normal.</b> There is no focal air space opacity, pleural effusion, or pneumothorax. <b>The osseous structures are intact with mild degenerative changes in thoracic spine</b>	<b>Impression:</b> no acute cardiopulmonary abnormality. <b>Findings:</b> lungs are clear bilaterally . cardiac and mediastinal silhouettes are normal . pulmonary vasculature is normal . no pneumothorax or pleural effusion . no acute bony abnormality	<b>Impression:</b> no acute cardiopulmonary abnormality . <b>Findings:</b> the lungs are clear and without focal air space opacity . cardio mediastinal silhouette is normal in size and contour and stable . there is no pneumothorax or large pleural effusion	<b>Impression:</b> no acute cardiopulmonary finding . <b>Findings:</b> <b>the cardio mediastinal silhouette is normal in size and contour .</b> the lungs are clear . there is no pleural effusion or pneumothorax . <b>the osseous structures are intact with mild degenerative changes in thoracic spine</b>
	The thorax is symmetrical, mediastinal heart shadow is in the middle, no enlarged lymph nodes are found in mediastinum, <b>bilateral lung markings are enhanced, patchy ground glass shadows are seen in both lungs, the boundary is fuzzy, some density is uneven,</b> lobar bronchus is unobstructed, and no abnormal density shadow is found in bilateral thoracic cavity	The thorax is symmetrical, mediastinal heart shadow is in the middle, no enlarged lymph node shadow is found in mediastinum, no abnormal density shadow is found in both lungs. lobar bronchus is unobstructed, and no abnormal density shadow is found in bilateral thoracic cavity.	The thorax is symmetrical, mediastinal heart shadow is in the middle, no enlarged lymph nodes are found in mediastinum, bilateral lung markings are enhanced, patchy shadows are seen in right lower lobe with clear boundary, lobar bronchus are unobstructed, and no abnormal density shadow is found in bilateral thoracic cavity.	The chest is symmetrical, mediastinal heart shadow is in the middle, no enlarged lymph node shadow is found in mediastinum, <b>bilateral lung markings are enhanced, multiple flake ground glass shadows are seen in both lungs, the boundary is fuzzy, the bilateral hilar is not big, some of the density is uneven,</b> the lobar bronchus are unobstructed, and no abnormal density shadow is found in bilateral chest.

Figure 2. Examples of the generated report on IU-xray(upper part) and COV-CTR dataset (lower part). We compare our results with the ground truth report and the other two methods: Show-Tell [30] and Grounded [44]. Highlighted sentences are the key information in the ground truth report that is only generated by our method.

Agent [18], KERP [16]. Since medical report generation models do not release the codes, we have to quote their performance from [16] that is often referred to in this field.

As shown in Table 1, on both datasets, our self-boosting method achieves the best performance over almost all evaluation metrics among the comparing methods. The two very recent image captioning methods M2transformer [3] and Grounded [44] perform better than other existing methods, but still lose to ours. Especially, our advantages over Grounded [44] fully demonstrate the benefit to jointly train the two tasks of report generation and image-text matching in our proposed self-boosting manner. It is also noted that, our performance improvement on CIDEr is most significant. Cross-referring to Table 2, even without the reinforcement learning on CIDEr (denoted as Ours-RL), our method is still the best performer on CIDEr on both datasets and the advantage is pronounced<sup>3</sup>. This is a positive sign for radiographic reports, since CIDEr focuses on key information by down-weighting the common words in all reports.

In addition, we also conduct qualitative comparison among our method, the classic ShowTell model [30], and the most recent and closest method of Grounded [44]. Two examples of the generated reports are visually compared in Fig. 2, one from each of the two datasets. A wider visual comparison among more examples and methods is provided in our supplement. Please note that COV-CTR reports are generated and evaluated in Chinese. They are simply translated into English by Google Translate<sup>4</sup> for display. As can be observed, our proposed model can significantly improve the quality of the generated reports, consistent with the above quantitative analysis. For example, our generated report correctly describes “the osseous structures” in the first example, and points out “the boundary is fuzzy” and “the density is uneven” in the second example, while the other two methods missed such information.

**Ablation Study** We conduct an ablation study to single

out the contributions of each component in our proposed method, as presented in Table 2. We take the RG branch of our proposed pipeline as the baseline to verify the performance improvements brought by ITM-branch and our self-boosted training strategy. In Table 2, there are four components: ITM, FLL, STL and RL, representing Image-Text matching, Feature-Level Loss, Self-boosted Loss and Reinforcement Learning, respectively. The symbols “+” or “-” indicate the inclusion or exclusion of the following component. The benefit of using the auxiliary image-text matching task can be well reflected by the improvement from “baseline” to “baseline+ITM”. As shown, the performance can be further boosted by additionally utilizing feature-level loss (“baseline+ITM+FLL”) in RG-branch and the self-boosted triplet loss (“baseline+ITM+STL”) in ITM-branch, proving their indispensability in our proposed method. Moreover, as mentioned, even removing the refinement by reinforcement learning, our model still significantly outperforms the comparing methods in Table 1.

## 5. Conclusions

In this work, we improve the diagnostic report generation by additionally utilizing an auxiliary image-text matching task to learn strongly correlated visual and text features to describe fine-grained differences among radiographic images. We show how to achieve this by deeply coupling the two tasks and encouraging the cooperative interactions between them. By communicating the learned features and the newly generated samples, the two tasks could mutually boost each other in a progressive way. As experimentally verified, our generated reports could better capture the subtle but critical information in radiographic images.

## 6. Acknowledgements

Prof Xiu Li was supported by the National Key R & D Program of China (No.2020AAA0108303) and NSFC 41876098.

<sup>3</sup>It is noted that Grounded [44] also uses RL on CIDEr for refinement.

<sup>4</sup>https://translate.google.com/



## 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 *CVPR*, 2018.
- [2] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *EMNLP*, 2015.
- [3] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *CVPR*, 2020.
- [4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *ACL*, 2019.
- [5] Demner Fushman Dina, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Rodriguez Laritza, Antani Sameer, George R Thoma, and Clement J Mcdonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association Jamia*, 2015.
- [6] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Trevor Darrell, and Kate Saenko. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015.
- [7] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. VSE++: improving visual-semantic embeddings with hard negatives. In *BMVC*, 2018.
- [8] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computat.*, 1997.
- [9] Baoyu Jing, Zeya Wang, and Eric Xing. Show, describe and conclude: On exploiting the structure information of chest X-ray reports. In *ACL*, 2019.
- [10] Baoyu Jing, Pengtao Xie, and Eric P. Xing. On the automatic generation of medical imaging reports. In *ACL*, 2018.
- [11] Andrej Karpathy and Fei-Fei Li. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015.
- [12] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.
- [13] Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *CoRR*, 2014.
- [14] Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. A hierarchical approach for generating descriptive image paragraphs. In *CVPR*, 2017.
- [15] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *ECCV*, 2018.
- [16] Christy Y. Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In *AAAI*, 2019.
- [17] Mingjie Li, Fuyu Wang, Xiaojun Chang, and Xiaodan Liang. Auxiliary signal-guided knowledge encoder-decoder for medical report generation. 2020.
- [18] Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. Hybrid retrieval-generation reinforced agent for medical image report generation. In *NeurIPS*, 2018.
- [19] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *ACL*, 2004.
- [20] Chunxiao Liu, Zhendong Mao, An-An Liu, Tianzhu Zhang, Bin Wang, and Yongdong Zhang. Focus your attention: A bidirectional focal attention network for image-text matching. In Laurent Amsaleg, Benoit Huet, Martha A. Larson, Guillaume Gravier, Hayley Hung, Chong-Wah Ngo, and Wei Tsang Ooi, editors, *MM*, 2019.
- [21] Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. Improved image captioning via policy gradient optimization of spider. In *ICCV*, 2017.
- [22] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *CVPR*, 2017.
- [23] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002.
- [24] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP*, 2019.
- [25] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2017.
- [26] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *CVPR*, 2017.
- [27] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015.
- [28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.
- [29] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015.
- [30] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015.
- [31] Haoran Wang, Ying Zhang, Zhong Ji, Yanwei Pang, and Lin Ma. Consensus-aware visual-semantic embedding for image-text matching. In *ECCV*, 2020.
- [32] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *CVPR*, 2016.
- [33] Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *ACL*, 2018.
- [34] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.
- [35] Yuan Xue and Xiaolei Huang. Improved disease classification in chest x-rays with transferred features from report generation. In *IPMI 2019*.
- [36] Yuan Xue, Tao Xu, L. Rodney Long, Zhiyun Xue, Sameer Antani, George R. Thoma, and Xiaolei Huang. Multimodal

- recurrent model with attention for automated radiology report generation. In *MICCAI*, 2018.
- [37] Xingyi Yang, Xuehai He, Jinyu Zhao, Yichen Zhang, Shanghang Zhang, and Pengtao Xie. Covid-ct-dataset: A ct scan dataset about covid-19, 2020.
- [38] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *CVPR*, 2019.
- [39] Zhilin Yang, Ye Yuan, Yuexin Wu, William W. Cohen, and Ruslan Salakhutdinov. Review networks for caption generation. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *NIPS*, 2016.
- [40] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *ECCV*, 2018.
- [41] Changchang Yin, Buyue Qian, Jishang Wei, Xiaoyu Li, Xi-anli Zhang, Yang Li, and Qinghua Zheng. Automatic generation of medical imaging diagnostic report with hierarchical recurrent neural network. In *ICDM*, 2020.
- [42] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *CVPR*, 2016.
- [43] Jianbo Yuan, Haofu Liao, Rui Luo, and Jiebo Luo. Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. In Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali R. Khan, editors, *MICCAI*, 2019.
- [44] Yuanen Zhou, Meng Wang, Daqing Liu, Zhenzhen Hu, and Hanwang Zhang. More grounded image captioning by distilling image-text matching model. In *CVPR*, 2020.