ECAI 2023
K. Gal et al. (Eds.)
© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms
of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230389

Multi-Modal Fusion with Semantic Supervision for Radiology Report Generation

Xing Jia[a], **Yun Xiong**[a;*], **Yao Zhang**[a] **and Li Luo**[b]

[a]Shanghai Key Laboratory of Data Science,
School of Computer Science,
Fudan University, China
[b]School of Public Health, Fudan University, China

Abstract. Radiology report generation, one way of analyzing radiology images, is to generate a textual report automatically for the given image, and it is of great significance to assist diagnosis and alleviate the workload of radiologists. Some report generation methods have been therefore proposed. However, these methods suffer from the problem of low-quality generation, because of the visual and textual bias and training with text similarity oriented objective. To solve this problem, we propose a novel radiology report generation model with multi-modal fusion and semantic supervision, namely MS-Gen. MS-Gen consists of two main components, i.e., the semantic-visual fusion module and the semantic weighted contrastive loss. Specifically, the main idea of the semantic-visual fusion module is to make use of the domain-specific prior knowledge contained in a large pre-trained visual-language model and also the complementary nature between the image and text modalities. Moreover, a novel optimization term, i.e., the semantic weighted contrastive loss, is proposed to guide the optimization process with semantic similarity objective, and further enforce the generated reports with higher clinical accuracy. Extensive experiments conducted on two real datasets of IU X-Ray and MIMIC-CXR demonstrate the effectiveness of MS-Gen.

1 Introduction

Analyzing radiology images, as the most common task of radiologists, plays an important role in various diagnoses in recent years. Automatic radiology report generation, one research on radiology images, is therefore in demand, since it can assist diagnosis and alleviate the workload of radiologists. It is to interpret a medical image by generating a corresponding diagnostic report automatically, as the case shown in Fig. 1.

With the release of image-report datasets and the advances in deep learning, some studies toward automatic radiology report generation have been proposed. [6, 41, 36, 21, 40, 43, 22, 18, 20, 19, 10, 3, 17, 24, 15, 8, 38, 12, 45]. The mainstream of report generation methods has been developed under the encoder-decoder paradigm. Specifically, a radiology report generation method consists of two components: (i) an image encoder that produces informative representations of the given image, and (ii) a decoder that produces the report based on the representations from the encoder. In general, the encoder is the convolutional neural networks (CNNs), and the decoder can be the recurrent neural networks (RNNs) [6, 36, 40, 43, 22, 10, 17, 38, 12, 45, 11, 36] or Transformers [41, 21, 18, 19, 15, 24, 9]. In addition, a few methods, based on large pre-trained language models, like GPT-3 [1] or ChatGPT [26], have also been recently proposed, which generate reports in a multi-stage way [33]. Although continuously improved performance has been obtained, these models still suffer from the problem of low-quality generation. Two main challenges are:

Impression: Interval removal of xxxx stent without acute cardiopulmonary abnormality.

Findings: Compared to prior examination, xxxx stent has been removed. cardiomediastinal silhouette is stable and within normal limits. stable mild atherosclerotic calcifications of the aortic xxxx are noted. there are mildly low lung volumes without focal consolidation, pneumothorax, or effusion identified. no acute bony abnormality seen.

Figure 1: A case of chest X-ray image along with its report. A report mainly consists of *Impression* part and *Findings* part. *Findings* part describes the detailed information of normal and abnormal findings. *Impression* part provides a summary statement.

- **Visual and textual bias.** Existing benchmark datasets of the report generation have the characteristic of visual and textual bias. That is, normal images and their corresponding descriptions dominate the dataset over the abnormal ones, which results in the problem that report generation methods tend to generate plausible general reports with no prominent abnormal narratives. However, in clinical practice, accurate detection and description of abnormalities are more helpful to radiologists. With this in mind, some studies incorporated prior knowledge in an implicit way, e.g., constructing knowledge graph [45, 15, 10, 19, 8], or in an explicit way, e.g., making use of the semantic labels obtained by multi-label classification [12, 43], the clinical history document [25, 8], or a template database [17]. However, these methods suffer from problems, like limited scope of knowledge because of the limited number of disease nodes in the graph, semantic noise because of incorrect classification results, redundant information in the clinical document, or insufficient coverage of potential diagnoses in the template.
- **Text similarity oriented optimization.** Clinical accuracy in the generated reports is of critical importance, i.e., the generated

* Corresponding Author. Email:yunx@fudan.edu.cn.

reports should be with correct semantic information, which is the point that radiologists are more concerned about. However, most report generation methods are trained with text similarity-oriented objectives, i.e., training a report generation model in a teacher-forcing way with cross-entropy taken as the loss function [6, 41, 36, 21, 40, 43, 20, 19, 10, 3, 15, 8, 38, 12, 45], or via reinforcement learning guided by evaluation metrics-based rewards [22, 18, 17, 24]. This results in the problem that these methods tend to generate the tokens with the text similarity objective, rather than the important semantic information for clinical accuracy.

To handle the challenges, we propose a novel radiology report generation model with multi-modal fusion and semantic supervision, namely MS-Gen. MS-Gen consists of two main components:

- **Semantic-visual fusion module.** Recently we have witnessed the superior effectiveness of large pre-trained visual-language models on various downstream tasks, which inspires us to adopt a large pre-trained visual-language to benefit the report generation. In this paper, we propose the semantic-visual fusion module. The key insight for designing this module is to make use of the domain-specific prior knowledge contained in a large pre-trained visual-language model and also the complementary nature between the image and text modalities, so as to alleviate the challenge of visual and textual bias and further benefit the report generation.
- **Semantic weighted contrastive loss.** Contrastive learning is to learn robust representations by contrasting similar (positive) and dis-similar (negative) samples, in which essential differences among samples are learned. In view of this, we propose a novel optimization term, i.e., the semantic weighted contrastive loss, and combine it with cross-entropy loss during the training phase. The main idea of this optimization term is to learn the essential semantic differences of medical images, by contrasting the positive sample with hard negative samples identified by the semantic similarity evaluation. Thus, reports with higher clinical accuracy are generated during the inference phase.

Overall, the main contributions of this paper are:

- A radiology report generation model, i.e., MS-Gen, is proposed. It improves the quality of the generated reports, which may push automatic radiology report generation closer to being applied to the practice.
- A multi-modal fusion module, i.e., the semantic-visual fusion module, is proposed, which inspires a more feasible line of alleviating the challenge of visual and textual bias of the benchmark datasets.
- A optimization term, i.e., the semantic weighted contrastive loss, is proposed. It inspires a way of integrating a semantic-oriented objective into the optimization process, so as to enforce the generated reports with higher clinical accuracy.
- Extensive experiments are conducted on real medical datasets with the evaluation metrics of both natural language generation and clinical efficacy. The results show that MS-Gen outperforms the state-of-the-art methods.

2 Related work

2.1 Image Caption and Paragraph Generation

Image caption is the task of generating a short textual description given an image and has attracted extensive interest in recent years. The dominant architecture of the caption task is based on the encoder-decoder framework provided by Show-Tell [30], which feeds the image features extracted by CNN as the input of RNN to produce image captions. On the basis of this framework, various attention mechanisms inspired by the human brain's attention are integrated, allowing the model to fix attention on salient visual or language signals [37, 44, 42]. Also, some studies explored the visual relationship among image regions by introducing scene graphs or graphical convolutional networks (GCNs) [39]. Since RNNs are with a low capability of long-range dependency modeling, Transformer [29] based image caption models have been recently introduced [4, 16]. Considering the limited ability to describe an image at a fine granularity of the image caption methods, paragraph generation task has been therefore introduced [14, 32]. Paragraph generation aims to generate a long and semantically coherent paragraph given an input image, in which rich information about the image is displayed in text form. To perform this task, the hierarchical RNN structure is widely used. In detail, the hierarchical RNN employs a two-level RNN to generate a paragraph based on the image representations extracted by CNN, where the topic-RNN is used to generate topic vectors and sentence-RNN further takes each topic vector as input to produce its corresponding description.

2.2 Radiology Report Generation

Compared with the image caption and paragraph generation tasks, existing report generation methods also almost follow the encoder-decoder paradigm [6, 41, 36, 21, 40, 43, 22, 18, 20, 19, 10, 3, 17, 24, 15, 8, 38, 12, 45], but the most urgent goal of the report generation task is clinical accuracy in the resulting generated reports. These methods typically tackled some aspects of the differences between the image caption and the radiology report generation tasks. For example, Yin *et al.* [40] introduced a topic-matching mechanism to solve the problem of semantic repetition among the sentences of a report. That is, topic vectors generated by topic-RNN and corresponding ground-truth reports are projected to the same embedding space, so as to make generated reports more accurate and diverse. Xue *et al.* [38] proposed a multimodal report generation model containing an iterative decoder with visual and semantic attention to improve the coherence between sentences in a recurrent way. Chen *et al.* [3] integrated a relational memory into the Transformer, in which writing patterns shared by relevant medical images are captured. Li *et al.* [17] proposed a hybrid model with template retrieval for the normal sentence generation and a generation module for the abnormal sentence generation respectively, so as to enhance the ability of the model in describing abnormalities. Besides, some studies also explored injecting prior knowledge into the generation model to improve the quality of the report generation, e.g., KG [45], PPKED [19], KERP [15] and RareGen [10]. However, most of these methods still suffer from the problem of low-quality generation, due to the visual and textual bias of the benchmark datasets and text similarity-oriented optimization. A few works studied this problem by designing semantic-based rewards and then optimizing a model via reinforcement learning, e.g., [24, 22], while their performance is limited by the instability of reinforcement learning itself.

2.3 Contrastive Learning

Contrastive learning, designed as a self-supervised technique, is to learn general-purpose representations by contrasting similar (positive) and dis-similar (negative) samples. That is, similar samples are

mapped closely together while dis-similar samples are mapped far apart, in which feature representations can be reused to benefit downstream tasks. Currently, contrastive learning has been widely applied to medical domains, e.g., electronic health records [35] and medical image analysis [31, 2], etc. On the report generation task, Liu *et al.* [20] leveraged the contrastive information between the input image and the normal images, so as to benefit the model to capture and further describe the abnormalities in medical images. In this paper, we apply contrastive learning in a supervised way. That is, we define the positive and negative samples by semantic labels, in which semantic information is integrated into the training process.

3 Method

3.1 Problem formulation

Prior to introducing the proposed MS-Gen, we provide the problem formulation first. We use the dataset $\mathcal{D}$ ={$(I_1, S_1, T_1), (I_2, S_2, T_2), ..., (I_M, S_M, T_M)$}, where M is the number of samples. I, S and T are the medical images, their corresponding semantic tags, and reports, respectively. Our goal is to train a report generation model $f : (I_i, S_i) \mapsto T_i$ to project I_i and its corresponding S_i to T_i.

In a nutshell, the main idea of MS-Gen is to solve the problem of low-quality generation by alleviating the visual and textual bias and enforcing the generated reports with higher clinical accuracy. Specifically, we propose the semantic-visual fusion module. It makes use of the domain-specific prior knowledge contained in a large pre-trained visual-language model and also the complementary nature between the image and text modalities. Moreover, we propose the semantic weighted contrastive loss, and combine it with cross-entropy loss during the training phase. Fig. 2 shows the diagram of MS-Gen. In the following sections, we generalize how a vanilla Transformer is used for report generation in the first subsection, since MS-Gen is a Transformer-based report generation model. Then, we elaborate on the semantic-visual fusion module in subsection 3.3. The semantic weighted contrastive loss is further introduced in subsection 3.4.

3.2 Basic architecture

Given an image I_i, a visual extractor is firstly used to extract its visual feature maps $\mathcal{V}_i \in R^{c\times h\times w}$. $\mathcal{V}_i$ is then reshaped into the visual matrix $\mathbf{V}_i \in R^{c\times(h\cdot w)}$. c, h, and w are the number of channels, height, and width, respectively. This process is formulated as:

$$\mathbf{V}_i = f_{cnn}(I_i), \tag{1}$$

where f_{cnn} represents the visual extractor, i.e., DenseNet121 [7] with fully connected layers removed in this paper.

Then, $\mathbf{V}_i$ is fed into the Transformer based encoder-decoder to generate a report $\hat{R}$. In particular, the encoder of Transformer maps $\mathbf{V}_i$ into the hidden matrix $\mathbf{H}_i^v \in R^{c\times(h\cdot w)}$. The decoder of Transformer further takes $\mathbf{H}_i^v$ and $\{\hat{y}_i\}_{i<t}$ as source inputs, to predict the token to be generated at the current time step t in an auto-regressive manner. This process is formulated as:

$$\mathbf{H}_i^v = f_e(\mathbf{V}_i), \tag{2}$$

$$\hat{y}_t = f_d(\mathbf{H}_i^v, \{\hat{y}_i\}_{i<t}), \tag{3}$$

where f_e and f_d are the encoder and decoder of the Transformer, respectively. $\{\hat{y}_i\}_{i<t} = \{\hat{y}_1, \hat{y}_2, ..., \hat{y}_{t-1}\}$ is the word-token sequence generated before time step t. The generated report is represented as $\hat{R} = \{\hat{y}_1, \hat{y}_2, ..., \hat{y}_{|\hat{R}|}\}$.

The core of the Transformer is the Multi-Head Attention (MHA), formulated as:

$$Att_m(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = softmax(\frac{\mathbf{Q}\mathbf{W}_m^Q(\mathbf{K}\mathbf{W}_m^K)^\top}{\sqrt{p/q}})\mathbf{V}\mathbf{W}_m^V, \tag{4}$$

$$MHA(\mathbf{S}, \mathbf{V}) = [Att_1(\mathbf{S}, \mathbf{V}, \mathbf{V}) \sqcup ... \sqcup Att_q(\mathbf{S}, \mathbf{V}, \mathbf{V})]\mathbf{W}^O, \tag{5}$$

where $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$ refer to the query, key, and value, respectively. $\mathbf{W}_m^Q$, $\mathbf{W}_m^K$, $\mathbf{W}_m^V$ and $\mathbf{W}^O$ are the parameter matrices of the mth head to be learned. p is the dimension of the input feature of each head, and q is the head number of MHA. $\sqcup$ represents the concatenation operation. The subsequent layer of MHA is the feed-forward network, and also the residual connection and layer normalization are used after the aforementioned sub-layers. Note that, the last MHA in the decoder is also followed by softmax. Different from the MHA in other modules where the query, key, and value are the same, e.g, the visual matrix in the encoder or the masked word-token sequence in the first attention layer of the decoder, the second attention layer in MHA of the decoder takes the visual matrix as key and value, with the masked word-token sequence as query.

3.3 Semantic-visual fusion module

The key insight for proposing the semantic-visual fusion module is to make use of the domain-specific prior knowledge contained in a large pre-trained visual-language model and also the complementary nature between the image and text modalities, in which the challenge of visual and textual bias is alleviated. Specifically, this module consists of *semantic-tag extraction* and *multi-modal fusion*.

3.3.1 Semantic-tag extraction

For the semantic-tag extraction, we first provide an overview of MedCLIP [34], a large pre-trained vision-language model in the medical domain. MedCLIP is a CLIP-like [28] model and is contrastively trained on unpaired medical images and text. Specifically, MedCLIP covers massive unpaired image and text datasets and scales usable training data in a random combinatorial manner. Also, MedCLIP replaces the InfoNCE loss of CLIP with the semantic matching loss, in which medical semantic similarity between each image and report is used as the supervision signal. In this way, MedCLIP owns the ability to capture the semantic information that describes a given image.

Specifically, we customize a set of semantic tags $\{t_n\}_{n=1}^N$ based on the datasets, and the tags with top-K similarities between an image are taken as the semantic tags of the image. This process is formulated as:

$$S_i = \{t_n\}_{n=1}^K = \underset{0\le n\le N}{topK} < f_v^c(I_i), f_t^c(t_n^p) >, \tag{6}$$

where $f_v^c(\cdot)$ and $f_t^c(\cdot)$ represent the image encoder and text encoder of MedCLIP, and their parameters are frozen. t_n^p represents the prompt of the semantic tag t_n. More in detail, for the extraction of customized semantic tags $\{t_n\}_{n=1}^N$, we employ the natural language process spaCy library [23] to extract an adjective and noun modifiers of a given disease, e.g., mild pneumonia, focal airspace. We obtained 320 tags in total, and each tag occurs at least 100 times in the dataset. Then, we design prompt templates to process each semantic tag into a textual description. For example, "focal airspace" is processed into

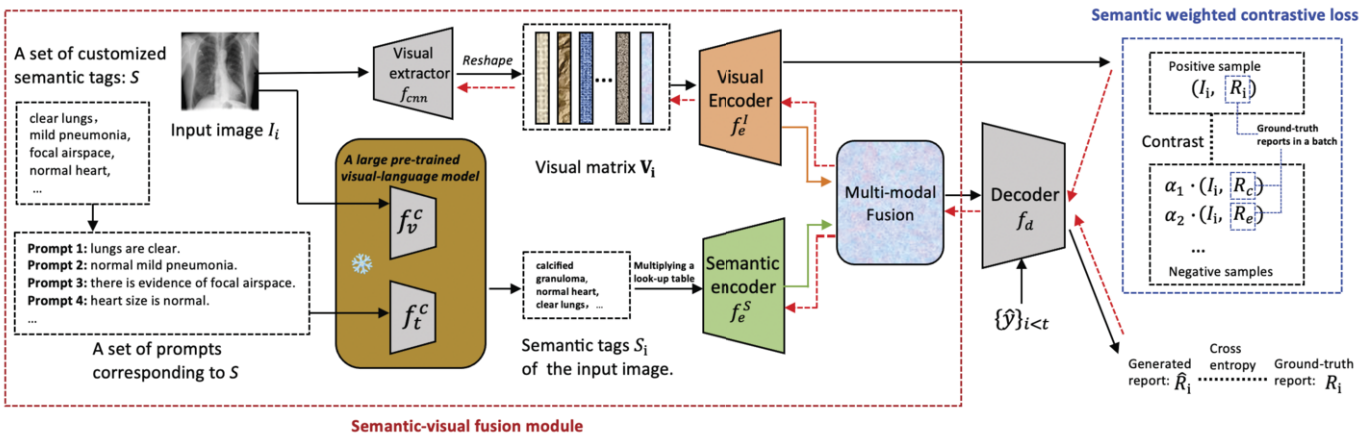


Figure 2: The diagram of MS-Gen. MS-Gen consists of two main components, i.e., the semantic-visual fusion module and the semantic weighted contrastive loss. Dotted red arrows denote the path of the gradients via back-propagation.

"there is evidence of focal airspace". Specifically, we extract descriptions for each semantic tag using regular expressions. Subsequently, the most commonly used expressions for each semantic tag were selected as prompts. Meanwhile, each image is centra-cropped to 224 $\times$ 224, to match the input of MedCLIP.

3.3.2 *Multi-modal fusion*

The motivation for performing multi-modal fusion is to take the complementary nature between the image and text modalities, which can benefit the representation learning and further the report generation.

Specifically, following the way of basic architecture dealing with an image, its visual feature maps $\mathcal{V}_i \in R^{c\times h\times w}$ extracted by a CNN is reshaped to its corresponding visual matrix $\mathbf{V}_i \in R^{c\times(h\cdot w)}$. For the semantic tag matrix, we first tokenize S_i and then embed it into the semantic matrix $\mathbf{S}_i \in R^{K\times d}$ by multiplying a look-up table, where K and d are the number of semantic tags and the dimension of each semantic representation. Then, $\mathbf{S}_i$ is further transformed to the hidden semantic matrix $\mathbf{H}_i^s$ by using the encoder of Transformer. This process is formulated as:

$$\mathbf{V}_i = f_{cnn}(I_i), \tag{7}$$

$$\mathbf{H}_i^v = f_e^I(\mathbf{V}_i), \tag{8}$$

$$\mathbf{H}_i^s = f_e^S(\mathbf{S}_i), \tag{9}$$

where $f_e^I(\cdot)$ and $f_e^S(\cdot)$ are the visual and semantic encoders, respectively. $\mathbf{V}_i$ and $\mathbf{S}_i$ are the visual and semantic matrices, respectively. Each column in $\mathbf{V}_i$ and $\mathbf{S}_i$, i.e., a 1-D vector, represents the feature of a region and a semantic representation of I_i and S_i. Note that both $\mathbf{V}_i$ and $\mathbf{S}_i$ are further passed through a layer normalization for stable training.

Once we obtain the visual hidden matrix $\mathbf{H}_i^v$ and the semantic hidden matrix $\mathbf{H}_i^s$, we get the aligned representations $\mathbf{V}^{'}$ by fusing image and semantic matrices. Then, the decoder of Transformer takes $\mathbf{V}^{'}$ and $\{\hat{y}_i\}_{i<t}$ as source inputs, to predict the token to be generated at the current time step t. This process is formulated as:

$$\mathbf{V}_i^{'} = MHA(\mathbf{H}_i^s, \mathbf{H}_i^v), \tag{10}$$

$$\hat{y}_t = f_d(\mathbf{V}_i^{'}, \{\hat{y}_i\}_{i<t}), \tag{11}$$

where (10) can be interpreted in the following way that the most relevant region features can be found given the hidden semantic matrix $\mathbf{H}_i^s$, and report generation is therefore benefited following (11).

3.4 *Semantic weighted contrastive loss*

The main idea for designing the semantic weighted contrastive loss is to enforce MS-Gen to pay more attention to the semantic differences of medical images, so as to make the generated reports with higher quality on clinical accuracy.

We achieve this goal by contrasting the positive sample with hard negative samples identified by the semantic label. Specifically, given a batch, the image-report pair (I_i, R_i) is taken as the positive sample. I_i combined with the ground-truth reports $\{R_j\}$ in the batch whose semantic labels are not equal to the semantic label of I_i are taken as the negative samples, and they are represented as $\{(I_i, R_j)|l_i \neq l_j\}$, where l_i and l_j are the semantic labels of R_i and R_j. The hard level of a negative sample is evaluated by the semantic similarity between l_i and l_j. The more similar the semantics are, the harder the sample (I_i, R_j) is. The process is formulated as:

$$L_{SSC} = \sum log \frac{exp(s_{i,i})}{\alpha \sum_{l_i \neq l_j} exp(s_{i,j})}, \tag{12}$$

$$s_{(i,j)} = sim(\mathbf{h}_i^v, \mathbf{h}_j^R)/\tau, \tag{13}$$

$$\mathbf{h}_i^v = \psi_v(\mathbf{H}_i^v), \tag{14}$$

$$\mathbf{h}_j^R = \psi_s(\mathbf{H}_j^R), \tag{15}$$

where $\psi_v(\cdot)$ and $\psi_s(\cdot)$ are operation sets consisting of average pooling and non-linear mapping, and they do not share parameters. sim represents the cosine similarity function. τ is the temperature parameter. $\mathbf{H}_i^v$ is the visual hidden matrix of the image I_i. $\mathbf{H}_j^R$ is the semantic hidden matrix of R_j. α is the semantic weight factor, and negative samples in the different hard levels will be set to different α. The harder a negative sample is, the larger the value of α is. That is, more weights will be assigned to these harder negative samples.

Overall, our report generation model is optimized by the semantic weighted contrastive loss combined with the cross-entropy loss.

$$L = \lambda \cdot L_{SWC} + L_{CE}, \tag{16}$$

where λ is a hyperparameter that weights the two losses. L_{SWC} and L_{CE} represent the semantic weighted contrastive loss and the cross-entropy loss, respectively.

4 Experiments and Results

4.1 Datasets

Table 1: The statistics of IU X-Ray and MIMIC-CXR. # Image, # Report, # Patient and # Avg. Len. represent the number of images, reports and patients, and the average length of reports.

Dateset	IU X-Ray			MIMIC-CXR		
	Training	Validation	Testing	Training	Validation	Testing
# Image	5,978	745	745	368,960	2,991	5,159
# Report	3,163	396	396	222,758	1,808	3,269
# Patient	3,163	396	396	64,586	500	293
# Avg. Len.	38.42	36.56	35.87	53.00	53.05	66.40

The experiments are conducted on IU X-Ray [5] and MIMIC-CXR [13], two benchmark datasets of the report generation task. Specifically, IU X-Ray is a public dataset, and it contains 7,470 chest images and 3,955 corresponding reports. MIMIC-CXR is a large-scale public dataset, and it contains 377,110 images and 227,835 corresponding reports. We preprocess the reports in both IU X-Ray and MIMIC-CXR by tokenizing, converting the tokens into lower cases, and removing the tokens with less than 3 occurrences. Four special tokens are added to a report, i.e., <*start*>, <*end*>, <*ukn*>, and <*pad*>. <*start*> and <*end*> are used to indicate the start and end of a report. <*pad*> token is used to fill the report whose length is less than the pre-defined maximum length of a report. <*ukn*> is to represent the excluded tokens. For the data-splitting, we apply the same way as current studies, which randomly divide the entire data into training, validation, and testing with a ratio of 8:1:1. For MIMIC-CXR, its official splitting is adopted, i.e., 368,960 in the training set, 2,991 in the validation set and 5,159 in the testing set. Table 1 shows the statistics of IU X-Ray and MIMIC-CXR.

4.2 Metrics

To evaluate the performance of report generation, we follow the most studies [6, 41, 36, 21, 40, 43, 22, 18, 20, 19, 10, 3, 17, 24, 15, 8, 38, 12, 45] and adopt the natural language generation (NLG) metrics, including BLEU-{1, 2, 3, 4} and ROUGE-L. These metrics are to compare the word-level overlap between the generated reports and the ground-truth reports, which are dubious for examining the semantic meanings of the generated reports. Therefore, we further train a report classification BERT-like [29] model to evaluate the clinical efficacy of the generated reports with the metrics of Accuracy, F-1, Precision, Recall, and Area Under the ROC Curve (AUC). The macro-average metrics can be quite noisy, as the per-class metric may depend on just a few samples. Therefore, the micro-average of AUC, Precision, Recall, and F-1 is adopted.

4.3 Model setting

For image pre-processing, resize and center-crop data augmentation techniques are employed during the training phase. An input image is center-cropped to 224×224, and no other data augmentation techniques are used during the inference phase. The visual extractor we

Table 2: PERFORMANCE COMPARISON OF REPORT GENERATION ON IU X-RAY AND MIMIC-CXR.

Dataset	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L
IU X-Ray	Soft-Att [30]	0.363	0.257	0.183	0.135	0.342
	Att-RK [37]	0.344	0.251	0.168	0.116	0.358
	Co-Att [12]	0.450	0.278	0.189	0.136	0.355
	KG [45]	0.446	0.300	0.211	0.152	0.359
	RareGen [10]	0.448	**0.343**	0.255	0.178	0.371
	R2Gen [3]	0.474	0.312	0.221	0.168	0.375
	CMM+RL [27]	0.492	0.323	0.231	0.177	0.371
	MS-Gen	**0.496**	0.330	**0.237**	**0.183**	**0.382**
MIMIC-CXR	Soft-Att [30]	0.311	0.193	**0.166**	0.112	0.287
	Att-RK [37]	0.325	0.219	0.159	0.120	**0.314**
	Co-Att [12]	0.331	0.220	0.147	0.117	0.276
	KG [45]	0.341	0.221	0.136	0.119	0.272
	RareGen [10]	0.356	0.220	0.147	0.122	0.281
	R2Gen [3]	0.356	0.223	0.151	0.110	0.283
	CMM+RL [27]	**0.370**	0.228	0.157	0.107	0.285
	MS-Gen	0.369	**0.231**	0.157	**0.124**	0.295

Table 3: PERFORMANCE COMPARISON OF CLASSIFICATION ON IU X-RAY AND MIMIC-CXR.

Dataset	Methods	Accuracy	F-1	Precision	Recall	AUC
IU X-Ray	Soft-Att [30]	0.3115	0.3526	0.4002	0.3107	0.7115
	Att-RK [37]	0.3008	0.3461	0.3972	0.3008	0.7083
	Co-Att [12]	0.3651	0.3868	0.4511	0.3519	0.7484
	KG [45]	0.3507	0.3707	0.4227	0.3301	0.7423
	RareGen [10]	0.3646	0.3973	0.4562	0.3512	0.7522
	R2Gen [3]	0.3568	0.3773	0.4262	0.3328	0.7451
	CMM+RL [27]	0.3638	0.4088	0.4652	0.3510	0.7387
	MS-Gen	**0.3762**	**0.4183**	**0.4782**	**0.3593**	**0.7592**
MIMIC-CXR	Soft-Att [30]	0.3623	0.3664	0.4302	0.3026	0.7328
	Att-RK [37]	0.3428	0.3503	0.4216	0.2792	0.6971
	Co-Att [12]	0.3712	0.4037	0.4563	0.3512	0.7501
	KG [45]	0.3853	0.4136	0.4581	0.3691	0.7562
	RareGen [10]	0.3877	0.4212	0.4611	0.3812	0.7683
	R2Gen [3]	0.3734	0.4168	0.4549	0.3788	0.7573
	CMM+RL [27]	0.3712	0.4191	0.4534	0.3822	0.7623
	MS-Gen	**0.4121**	**0.4353**	**0.4702**	**0.4005**	**0.7792**

used in this paper is DenseNet-121 with fully connected layers removed. Each token in reports is embedded as a vector with a dimension of 256. We set the number of MHAs of both visual and semantic encoders to be 3. The dimensions of both the input layer and hidden layers of the visual and semantic encoder are 256. We conduct a grid-based search to choose the weight factor $\lambda \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ by evaluating the model's performance on the validation sets of the two benchmark datasets. The λ we set in this paper is 0.3. The decoder of MS-Gen is the standard decoder of the Transformer, and the number of MHAs of the decoder is 3. The dimensions of the hidden layer of the decoder are 256. For the setting of the hard levels of negative samples, we set 4 levels. That is, for the semantic similarity in [-1, -0.5), [-0.5, 0), [0, 0.5), or [0.5, 1), the weight factor α is set to 1, 1.5, 2, or 2.5, respectively. We set the number of semantic tags K to be 7, considering the sentences' number in a report. Considering the average length of IU X-Ray and MIMIC-CXR, we set the maximum length of a generated report of IU X-Ray and MIMIC-CXR to be 40 and 70, respectively. Experiments are conducted on 2 Nvidia Tesla V100 GPUs, and each GPU is with 32GB VRAM.

Table 4: ABLATION STUDY OF CLASSIFICATION ON IU X-RAY AND MIMIC-CXR.

Dataset	Methods	Accuracy	F-1	Precision	Recall	AUC
IU X-Ray	MS-Gen	**0.3762**	**0.4183**	**0.4782**	**0.3593**	**0.7592**
	w/o SVF	0.3251	0.3570	0.4106	0.3105	0.7201
	w/o SWC	0.3431	0.3823	0.4438	0.3211	0.7427
	VC	0.3562	0.4021	0.4611	0.3437	0.7410
MIMIC-CXR	MS-Gen	**0.4121**	**0.4353**	**0.4702**	**0.4005**	**0.7792**
	w/o SVF	0.3629	0.3942	0.4318	0.3566	0.7461
	w/o SWC	0.3891	0.4256	0.4635	0.3877	0.7681
	VC	0.3851	0.4103	0.4502	0.3679	0.7510

(w/o SVF) and (w/o SWC) represent MS-Gen without the semantic-visual fusion module and the semantic weighted contrastive loss. VC represents MS-Gen equipped with vanilla contrastive loss, instead of SA.

4.4 Baselines

We consider the following representative models as baselines: Soft-Att [30], Att-RK [37], Co-Att [12], KG [45], RareGen [10], and R2Gen [3]. Among these baselines, Soft-Att and Att-RK are classic CNN-RNN based models without prior knowledge incorporated. Co-Att, KG, and RareGen are hierarchical report generation models, where the report generation process is divided into two steps in the decoding stage, including topic vector generation by topic decoder and sentence generation by sentence decoder. Co-Att assists the report generation by making use of the semantic information, i.e., the semantic labels obtained from the multi-label classification. KG and RareGen incorporate prior knowledge by introducing a knowledge graph. Also, we compare MS-Gen with R2Gen, a recently published state-of-the-art method. It proposes to use an extra component, i.e., relational memory, to enhance Transformer to learn from the patterns, which can also be considered as a way of knowledge incorporation. Meanwhile, we compare MS-Gen with [27]. It uses reinforcement learning over a cross-modal memory to align visual and textual features without relying on annotation.

4.5 Experimental Results and Analysis

Table 5: ABLATION STUDY OF REPORT GENERATION ON IU X-RAY AND MIMIC-CXR.

Dataset	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L
IU X-Ray	MS-Gen	**0.496**	**0.330**	**0.237**	**0.183**	**0.382**
	w/o SVF	0.453	0.298	0.211	0.146	0.325
	w/o SWC	0.469	0.313	0.220	0.165	0.344
	VC	0.478	0.316	0.220	0.171	0.367
MIMIC-CXR	MS-Gen	**0.369**	**0.231**	**0.157**	**0.124**	**0.295**
	w/o SVF	0.310	0.165	0.126	0.091	0.222
	w/o SWC	0.335	0.194	0.140	0.109	0.268
	VC	0.351	0.222	0.143	0.117	0.282

(w/o SVF) and (w/o SWC) represent MS-Gen without the semantic-visual fusion module and the semantic weighted contrastive loss. VC represents MS-Gen equipped with vanilla contrastive loss, instead of SA.

(1) Quantitative Results Table 2 shows the quantitative experimental results of MS-Gen and its baselines on NLG metrics. We can find that our proposed method MS-Gen outperforms baselines on almost metrics, e.g., MS-Gen achieves markedly better results on both IU X-Ray and MIMIC-CXR compared with the state-of-the-art method R2Gen, which demonstrates the effectiveness of our proposed method MS-Gen for the radiology report generation, in terms of word overlap. Meanwhile, although MS-Gen achieves slightly lower NLG scores than some baselines, e.g., slightly lower BLEU-3 and ROUGE-L scores on MIMIC-CXR and BLEU-2 score on IU X-Ray, this does not indicate MS-Gen is with worse performance. As we have mentioned before, NLG metrics are to measure the word overlap between the generated reports and the ground-truth reports, but cannot evaluate whether the generated reports own the correct semantic information or not compared with the ground-truth reports. And due to data bias, a model can achieve considerable BLEU and ROUGE-L scores even when it just repeats the most frequent sentences. For this reason, we further show the report classification results of MS-Gen and baselines in Table 3. We can observe that MS-Gen achieves the highest scores on all classification metrics of both IU X-Ray and MIMIC-CXR compared with the baselines, which demonstrates its great ability to generate reports with more clinical accuracy. The potential reason may come from two aspects. On the one hand, semantic tags introduced by applying MedCLIP and multi-modal fusion alleviate the visual and textual bias to a large extent, which guides MS-Gen to correctly generate reports with prominent abnormal descriptions. On the other hand, training MS-Gen with the semantic weighted contrastive loss enforces the generated reports with correct semantic information. In addition, we can also find that all the methods are with low accuracy scores, and the reason can be inferred that an instance is correctly predicted only if all the semantic labels are

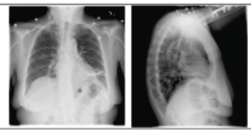
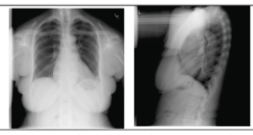
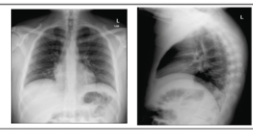
Two-view images			
Ground truth	patchy airspace disease in the left lingula. no significant effusion. clear right lung. normal heart size. granulomatous mediastinal calcifications. right chest xxxx tip at svc.	heart size moderately enlarged, stable mediastinal contours. xxxx xxxx opacity in the left lung base. otherwise, no focal alveolar consolidation, no definite pleural effusion seen. no typical findings of pulmonary edema	the heart size is within normal limits. trachea is midline. no pleural effusions or pneumothorax. there is focal consolidation in the posterior segment of the right lower lobe. no bony or soft tissue abnormalities.
Co-Att[11]	lungs are clear without areas of focal consolidation. there is airspace disease. the heart size appear within normal limits size . no pneumothorax is seen . no pneumothorax is seen .	the heart pulmonary and mediastinum are within normal limits. no focal consolidation . no pneumothorax is identified . no pneumothorax is identified.	heart size is within normal size. no pleural effusion and pleural fluid . no pneumothorax no focal airspace disease. no pneumothorax no focal airspace disease. no pleural effusion and pleural fluid .
R2Gen[4]	there is evidence of focal airspace disease. heart is within normal limits for low lung . there is no hilar lymph . visualized osseous structures are unremarkable in appearance .	no focal consolidation is seen . the lungs are clear. there is no pleural effusion or pneumothorax. no acute cardiopulmonary abnormality.	the heart size is normal . no pneumothorax or effusion identified. there is on pleural effusion . there are lung volumes with focal consolidation. no acute bony abnormality seen .
KG[44]	airspace disease appears . heart size and mediastinal contours appear within normal limits. there is no pneumothorax . no focal consolidation . no focal consolidation .	the heart is normal in size. no large pleural effusions. no evidence of pneumothorax . no evidence of pneumothorax . there is no focal air space opacity .	the heart is normal size with normal appearance . there is focal airspace opacity . there is focal airspace opacity . the mediastinum is unremarkable . the mediastinum is unremarkable .
MS-Gen	there is evidence of focal airspace disease. there is no pneumothorax . the heart size is normal . no lobar consolidation or effusion. no pneumothorax.	enlarged heart size is identified. the lungs are free of focal airspace disease. suspicious pulmonary opacity . lungs are without areas of focal consolidation . no pleural effusion or pneumothorax .	the heart is normal size with normal appearance . no abnormal airspace opacities . focal consolidation are noted. pulmonary is within normal limits.

Figure 3: 3 cases of ground-truth reports and reports generated by our method MS-Gen and the state-of-the-art baselines. The tokens marked in blue represent the abnormalities or diseases of the ground-truth reports. The tokens marked in red represent the abnormalities or diseases correctly detected in the generated reports.

correctly classified. The importance of reporting the accuracy metric is that it could evaluate the completeness and correctness of a generated report to a large extent.

(2) Ablation study

Moreover, we also conduct an ablation study to confirm the contributions of each component in our proposed method MS-Gen, i.e., the semantic-visual fusion module and the semantic weighted contrastive loss. We compare our full model MS-Gen with (w/o SVF) and (w/o SWC). (w/o SVF) represents the MS-Gen without the semantic-visual fusion module, i.e., taking the image as input only. (w/o SWC) represents the MS-Gen without the semantic weighted contrastive loss, i.e., taking the cross-entropy loss only. Also, we also compare our full model with VC which represents MS-Gen equipped with vanilla contrastive loss. As the report generation and report classification results shown in Table 6 and Table 5, the benefit of using these two components can be well reflected by the improvement by comparing MS-Gen with (w/o SVF) and (w/o SWC). For example, on IU X-Ray, removing the semantic-visual fusion module leads to a performance reduction by over 5% on Accuracy and 4% on Recall. This demonstrates that the semantic-visual fusion module is an effective means to solve the problem of visual and textual bias. Meanwhile, by comparing (w/o SWC) with MS-Gen, we find that removing the semantic weighted contrastive loss results in the performance reduction on both IU X-Ray and MIMIC-CXR, which verifies that the semantic weighted contrastive loss term plays an essential role in generating reports with clinical accuracy. To further investigate the effectiveness of MS-Gen, we show the qualitative analysis of 3 cases by comparing the ground-truth reports with the generated reports from the state-of-the-art methods, as the results shown in Fig. 3. We can find that MS-Gen can correctly generate the descriptions of abnormalities or diseases in all the 3 cases. Note that, in the second column, other methods cannot detect both the enlarged heart and opacity, while MS-Gen can identify both of them, which confirms its superior performance of generating reports with higher quality.

5 Conclusion

In this paper, we propose a novel radiology report generation model, namely MS-Gen. It boosts the quality of the generated reports by alleviating the visual and textual bias and improving clinical accuracy, which may push automatic radiology report generation closer to being applied to the practice.

6 Acknowledgements

This work is partially supported by the National Key Research and Development Plan Project 2022YFC3600901, the National Natural Science Foundation of China Projects No. U1936213, and the Major Key Project of PCL (PCL2021A06).

References

[1] Tom Brown and et al. Mann, 'Language models are few-shot learners', in *Advances in Neural Information Processing Systems*, eds., H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, volume 33, pp. 1877–1901. Curran Associates, Inc., (2020).

[2] Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu, 'Contrastive learning of global and local features for medical image segmentation with limited annotations', *Advances in Neural Information Processing Systems*, **33**, 12546–12558, (2020).

[3] Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan, 'Generating radiology reports via memory-driven transformer', *arXiv preprint arXiv:2010.16056*, (2020).

[4] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara, 'Meshed-memory transformer for image captioning', in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10578–10587, (2020).

[5] 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, (2016).

[6] Philipp Harzig, Yan-Ying Chen, Francine Chen, and Rainer Lienhart, 'Addressing data bias problems for chest x-ray image report generation', *arXiv preprint arXiv:1908.02123*, (2019).

[7] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger, 'Densely connected convolutional networks', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, (2017).

[8] Xin Huang, Fengqi Yan, Wei Xu, and Maozhen Li, 'Multi-attention and incorporating background information model for chest x-ray image report generation', *IEEE Access*, **7**, 154808–154817, (2019).

[9] Xing Jia, Yun Xiong, Jiawei Zhang, Yao Zhang, Blackley Suzanne, Yangyong Zhu, and Chunlei Tang, 'Radiology report generation for rare diseases via few-shot transformer', in *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1347–1352. IEEE, (2021).

[10] Xing Jia, Yun Xiong, Jiawei Zhang, Yao Zhang, and Yangyong Zhu, 'Few-shot radiology report generation for rare diseases', in *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 601–608. IEEE, (2020).

[11] Xing Jia, Yun Xiong, Jiawei Zhang, Yao Zhang, Yangyong Zhu, and S Yu Philip, 'Few-shot radiology report generation via knowledge transfer and multi-modal alignment', in *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1574–1579. IEEE Computer Society, (2022).

[12] Baoyu Jing, Pengtao Xie, and Eric Xing, 'On the automatic generation of medical imaging reports', *arXiv preprint arXiv:1711.08195*, (2017).

[13] Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng, 'Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs', *arXiv preprint arXiv:1901.07042*, (2019).

[14] 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 computer vision and pattern recognition*, pp. 317–325, (2017).

[15] Christy Y Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing, 'Knowledge-driven encode, retrieve, paraphrase for medical image report generation', **33**(01), 6666–6673, (2019).

[16] Guang Li, Linchao Zhu, Ping Liu, and Yi Yang, 'Entangled transformer for image captioning', in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8928–8937, (2019).

[17] Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing, 'Hybrid retrieval-generation reinforced agent for medical image report generation', *Advances in neural information processing systems*, **31**, (2018).

[18] Fenglin Liu, Shen Ge, and Xian Wu, 'Competence-based multimodal curriculum learning for medical report generation', *arXiv preprint arXiv:2206.14579*, (2022).

[19] Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou, 'Exploring and distilling posterior and prior knowledge for radiology report generation', 13753–13762, (2021).

[20] Fenglin Liu, Changchang Yin, Xian Wu, Shen Ge, Ping Zhang, and Xu Sun, 'Contrastive attention for automatic chest x-ray report generation', *arXiv preprint arXiv:2106.06965*, (2021).

[21] Fenglin Liu, Chenyu You, Xian Wu, Shen Ge, Xu Sun, et al., 'Auto-encoding knowledge graph for unsupervised medical report generation', *Advances in Neural Information Processing Systems*, **34**, 16266–16279, (2021).

[22] 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', in *Machine Learning for Healthcare Conference*, pp. 249–269. PMLR, (2019).

[23] I. Montani M. Honnibal, 'spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing.', p. https://github.com/explosion/spaCy/issues/5863., (2017).

[24] Yasuhide Miura, Yuhao Zhang, Emily Bao Tsai, Curtis P Langlotz, and Dan Jurafsky, 'Improving factual completeness and consistency of image-to-text radiology report generation', *arXiv preprint arXiv:2010.10042*, (2020).

[25] Hoang TN Nguyen, Dong Nie, Taivanbat Badamdorj, Yujie Liu, Yingying Zhu, Jason Truong, and Li Cheng, 'Automated generation of accurate\& fluent medical x-ray reports', *EMNLP*, (2021).

[26] OpenAI, 'Optimizing language models for dialogue', (2023).

[27] Han Qin and Yan Song, 'Reinforced cross-modal alignment for radiology report generation', in *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 448–458, (2022).

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, and et al. Ramesh, 'Learning transferable visual models from natural language supervision', in *International conference on machine learning*, pp. 8748–8763. PMLR, (2021).

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, 'Attention is all you need', *Advances in neural information processing systems*, **30**, (2017).

[30] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan, 'Show and tell: A neural image caption generator', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3156–3164, (2015).

[31] Yen Nhi Truong Vu, Richard Wang, Niranjan Balachandar, Can Liu, Andrew Y Ng, and Pranav Rajpurkar, 'Medaug: Contrastive learning leveraging patient metadata improves representations for chest x-ray interpretation', in *Machine Learning for Healthcare Conference*, pp. 755–769. PMLR, (2021).

[32] Jing Wang, Yingwei Pan, Ting Yao, Jinhui Tang, and Tao Mei, 'Convolutional auto-encoding of sentence topics for image paragraph generation', *arXiv preprint arXiv:1908.00249*, (2019).

[33] Sheng Wang, Zihao Zhao, Xi Ouyang, Qian Wang, and Dinggang Shen, 'Chatcad: Interactive computer-aided diagnosis on medical image using large language models', *arXiv preprint arXiv:2302.07257*, (2023).

[34] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun, 'Medclip: Contrastive learning from unpaired medical images and text', *arXiv preprint arXiv:2210.10163*, (2022).

[35] Tingyi Wanyan, Hossein Honarvar, Suraj K Jaladanki, and Chengxi et al. Zang, 'Contrastive learning improves critical event prediction in covid-19 patients', *Patterns*, **2**(12), 100389, (2021).

[36] Xiancheng Xie, Yun Xiong, Philip S Yu, Kangan Li, Suhua Zhang, and Yangyong Zhu, 'Attention-based abnormal-aware fusion network for radiology report generation', in *International Conference on Database Systems for Advanced Applications*, pp. 448–452. Springer, (2019).

[37] 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 *International conference on machine learning*, pp. 2048–2057. PMLR, (2015).

[38] 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', 457–466, (2018).

[39] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai, 'Auto-encoding scene graphs for image captioning', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10685–10694, (2019).

[40] Changchang Yin, Buyue Qian, Jishang Wei, Xiaoyu Li, and Xianli Zhang, 'Automatic generation of medical imaging diagnostic report with hierarchical recurrent neural network', 728–737, (2019).

[41] Di You, Fenglin Liu, Shen Ge, Xiaoxia Xie, Jing Zhang, and Xian Wu, 'Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation', 72–82, (2021).

[42] Quanzeng You, Hailin Jin, Zhaowen Wang, and Chen Fang, 'Image captioning with semantic attention', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4651–4659, (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', 721–729, (2019).

[44] Zheng-Jun Zha, Daqing Liu, and Hanwang et al. Zhang, 'Context-aware visual policy network for fine-grained image captioning', *IEEE transactions on pattern analysis and machine intelligence*, (2019).

[45] Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan Yuille, and Daguang Xu, 'When radiology report generation meets knowledge graph', **34**(07), 12910–12917, (2020).