Artificial Intelligence In Medicine 150 (2024) 102823

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Research paper

A label information fused medical image report generation framework

Shuifa Sun [a,b], Zhoujunsen Mei [b,c], Xiaolong Li [b,d], Tinglong Tang [b,c], Zhanglin Su [a], Yirong Wu [e,*]

[a] School of Information Science and Technology, Hangzhou Normal University, Hangzhou, 311121, Zhejiang, China
[b] Yichang Key Laboratory of Intelligent Medicine, Yichang Key Laboratory of Intelligent Medicine, Yichang, 443002, Hubei, China
[c] College of Computer and Information Technology, China Three Gorges University, Yichang, 443002, Hubei, China
[d] College of Economics and Management, China Three Gorges University, Yichang, 443002, Hubei, China
[e] Institute of Advanced Studies in Humanities and Social Sciences, Beijing Normal University, Zhuhai, 519087, Guangdong, China

ARTICLE INFO

Keywords:
Medical image
Text generation
Attention mechanism
Multi-modal feature fusion
Feature extraction

ABSTRACT

Medical imaging is an important tool for clinical diagnosis. Nevertheless, it is very time-consuming and error-prone for physicians to prepare imaging diagnosis reports. Therefore, it is necessary to develop some methods to generate medical imaging reports automatically. Currently, the task of medical imaging report generation is challenging in at least two aspects: (1) medical images are very similar to each other. The differences between normal and abnormal images and between different abnormal images are usually trivial; (2) unrelated or incorrect keywords describing abnormal findings in the generated reports lead to mis-communications. In this paper, we propose a medical image report generation framework composed of four modules, including a Transformer encoder, a MIX-MLP multi-label classification network, a co-attention mechanism (CAM) based semantic and visual feature fusion, and a hierarchical LSTM decoder. The Transformer encoder can be used to learn long-range dependencies between images and labels, effectively extract visual and semantic features of images, and establish long-term dependent relationships between visual and semantic information to accurately extract abnormal features from images. The MIX-MLP multi-label classification network, the co-attention mechanism and the hierarchical LSTM network can better identify abnormalities, achieving visual and text alignment fusion and multi-label diagnostic classification to better facilitate report generation. The results of the experiments performed on two widely used radiology report datasets, IU X-RAY and MIMIC-CXR, show that our proposed framework outperforms current report generation models in terms of both natural linguistic generation metrics and clinical efficacy assessment metrics. The code of this work is available online at https://github.com/watersunhznu/LIFMRG.

1. Introduction

The task of automatic medical image report-generation aims to generate 6C (clear, correct, concise, complete, consistent, and coherent) reports from a given medical image [1]. Automatic generation of high-quality medical imaging reports can greatly accelerate workflow automation, reduce physicians' workloads, reduce the probability of incorrect reports, and improve the quality and standardization of medical reports. Therefore, it has become a popular research topic in the field of artificial intelligence and intelligent medicine.

The task of automatic medical image report generation belongs to a subclass of the image caption task. Therefore, a considerable part of the generation models for medical image reports are derived or optimized from the image caption models. These models can be broadly classified into three categories: (1) encoding-decoding based models [2,3]; (2) graph structure based models [4,5]; (3) reinforcement learning based models [6]. The most widely used is the encoding-decoding based model [7], which originates in the field of machine translation. Its encoder uses a convolutional neural network (CNN) to extract image features, and its decoder uses a recurrent neural network (RNN) to generate reports. The report quality can be further improved by replacing the encoder network or decoder network, training additional classifiers to predict disease labels or medical labels [8], and utilizing knowledge graphs [9].

Although the existing models can generate a complete report, it is far from the goal of generating 6C reports. Currently, the challenges in the field are as follows: first, medical images are different from conventional images (e.g., landscape, photos). Medical images are highly similar, with only minor differences at the area of anomalies, leading

* Corresponding author.
E-mail address: yrwu@bnu.edu.cn (Y. Wu).

https://doi.org/10.1016/j.artmed.2024.102823
Received 15 October 2022; Received in revised form 21 February 2024; Accepted 21 February 2024
Available online 22 February 2024

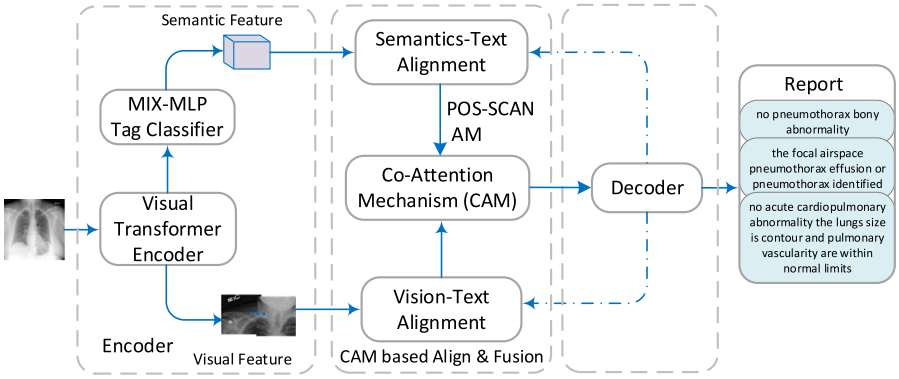

Fig. 1. A diagram of the proposed framework. From left to right in the three dashed boxes: encoder, co-attention mechanism(CAM) based alignment and fusion, and decoder.

the model to generate a large number of similar reports [10]. Second, for multiple anomalies found by visual feature extractors, it is difficult for the existing models to sort out the logical relationships between anomalies and generate the correct anomaly labels. Therefore, it becomes an urgent task to develop models, in which the encoder is used to distinguish the trivial differences between images and extract correct feature vectors, and the decoder is used to generate a high-quality report containing accurate anomalous keywords.

In this paper, leveraging Transformer [11], MIX-MLP [12] and attention mechanism [13], we proposed an encoding-decoding based medical image report generation framework, which is composed of three main parts, i.e., an encoder consisting of a visual and semantic feature extractor based on Vision Transformer (ViT) encoder and a MIX-MLP tag classifier, a co-attention mechanism (CAM) based visual and semantic feature alignment and fusion module [13], and a decoder, as shown in Fig. 1. Overall, the contributions of our work are as follows:

(1) A Transformer based Vision Transformer encoder [14] is used as a visual and semantic feature extractor to obtain both types of features simultaneously. The feature information is extracted from the penultimate layer of ViT encoder, which is separated into visual and semantic features and input to the downstream module separately. As far as we know, this is the first time that ViT encoder is adopted in the encoding-decoding framework for medical image report generation, which is used to effectively extract low-level features of images and establish long-term dependencies between modalities.

(2) The MIX-MLP network [12] is added to the multi-label classification network in order to accurately classify the features extracted by the visual feature extractor. The MIX-MLP network can learn features with different labels in multiple dimensions, improving the classification accuracy while maintaining the time complexity $O(n)$ of the MLP network. The Focal Loss [15] function is also introduced in the multi-label classification network to alleviate the problem of imbalanced and sparse label distribution.

(3) A POS-SCAN based co-attention mechanism for visual and text alignment is proposed to infer the similarity between global images and text, which is achieved by mapping the visual semantic information and multi-label classification information into the same joint semantic space to align with text information, so that images and text can be matched at fine granularity.

Finally, we assessed the performance of the proposed framework on two datasets. The results show that our framework outperforms a variety of current state-of-the-art approaches in image caption and automatic medical image report generation.

2. Related works

Automatic medical image report generation is essentially an image caption task. So, many models applied in the field of image caption can also be applied to the task of automatic medical image report generation, such as the encoder–decoder architecture (CNN-RNN) proposed by Vinyals et al. [7]. Inspired by the RNN-RNN based machine translation architecture, the encoder RNN is replaced by the CNN to extract the rich representation of the input image. The main stream feature extractor in the encoder is usually the CNN based VGG [16] or ResNet [17]. CNN have great advantages in extracting local features of images. However, it cannot handle long-distance relationships well due to the limitations of convolutional operations. Recently, Vision Transformer (ViT) [14], which combines a CNN with a form of self-attention, was used for image classification. A self-attention mechanism can be used to further process the output of CNN to achieve better results, especially on some large-scale datasets [18]. Dai et al. [19] used the Hybrid architecture for medical image classification and proposed the TransMed model, which combines the advantages of CNN and Transformer to effectively extract low-level features of images and establish long-term dependencies between modalities. Inspired by this work, we adopt ViT as the visual feature extractor to obtain the final visual features.

The goals of label classification tasks can be divided into: single label binary classification, single label multiple classification, and multi-label classification. For medical images, one image often corresponds to multiple labels. So the label classification target in the feature extraction process is generally multi-label classification. Traditional label classifiers generally use multi-layer perceptron (MLP) or convolutional neural networks to classify labels. MIX-MLP is a computer vision model that only uses the MLP structure to replace the convolutional operation (Conv) in traditional CNN and the self attention mechanism (Self Attention) in Transformer [12]. For the aforementioned encoding-decoding based medical image report generation framework [8], the label information play a key role in the process. Because of the excellent performance of MIX-MLP, compared with the CNN and Transformer based approach, we select MIX-MLP as the label classifier in the framework to generate the correct anomaly labels in the highly similar medical images.

Ba et al. [20] and Xu et al. [21] found that the attention mechanism can be used to combine the image features extracted by the image encoder with the label information output from the image multi-label classification network to assist in the generation of medical image reports. Since the label information obtained from the multi-label classification network is local information in multiple places, while the image features extracted by the image encoder are global information, the feature information of this image cannot be well represented if the two are simply stitched together [22,23]. Jing et al. [8] used a co-attention mechanism that focuses on both image features and label information, which stitches together the image features and label information to develop a text generation model to form a complete medical image report. Zhou et al. [13] proposed a POS-SCAN model based on the SCAN text-image matching model [24] that can focus on the correct object when generating the report. It significantly improves

the accuracy in unsupervised learning conditions. In this work, POS-SCAN based co-attention mechanism is adopted to align and fuse the visual information with label information on the mapping space.

The decoder can generate a complete report by generating different topics and then generating sentences from the topics. Considering the structural nature of medical image reports, i.e., a sentence or a paragraph to describe an abnormality, and then multiple sentences or paragraphs to form a complete medical image report. The structural characteristics of the hierarchy LSTM network are compatible with those of medical image report, which can be used to solve the problem that hidden information cannot be retained in a long paragraph generation by a LSTM network [14,25]. The hierarchical LSTM network consists of two LSTM networks: a sentence LSTM (Sentence LSTM) and a word LSTM (Word LSTM). The Sentence LSTM generates topics and the Word LSTM generates a sentence for each topic. The hierarchical LSTM enables the models to model long sentences better than the RNN model with multiple LSTMs [3]. Although the fully connected networks work well in the Transformer based GPT series, such as ChatGPT [26], the LSTM network is selected as the final output layer due to its low computing load and memory usage. On the other hand, because the attention of this work is paid on the improvement of the encoder, we select the well validated LSTM network as the decoder. At the same time, large language models (LLMs) may bring new opportunities for the generation based language tasks [26,27], but they are not readily applicable in clinical settings due to a lack of medical knowledge in specialized domains [28].

According to the above analysis, the current mainstream research related to encoder–decoder models mainly focuses on the improvement of the decoder. However, the encoder is responsible for processing the input medical images into feature vectors and its performance will affect the overall model performance [14]. Additionally, the co-attention mechanism is responsible for fusing the visual and semantic information extracted by the encoder in order for the decoder to generate a high-quality report. So, our work focuses on the encoder and fusion parts, and adopts the existing hierarchical LSTM network based decoder to generate the final report. Overall, the research related to medical image report generation is still in its infancy. The research on multi-modal feature extraction, alignment and fusion in the field of medical image report generation is still limited.

3. Method

In this paper, we propose a label information fused medical image report generation framework, using a Transformer [11] encoder as a visual feature extractor, a MIX-MLP network [12] as a multi-label classifier, a co-attention mechanism based on a POS-SCAN image text matching model [13] to align and fuse visual and semantic features [8], and a hierarchical LSTM network [29] text generator to generate reports, as shown in Fig. 2. Among them, the Transformer network as a visual feature extractor can discover fine-grained anomalies in images. The MIX-MLP network can learn the relationship between anomaly labels and generate accurate anomaly labels for multiple anomalies found. The POS-SCAN based co-attention mechanism can align visual information with label information on the mapping space and fuse them. The hierarchical LSTM network can generate sentences containing anomaly labels by the guide of the co-attention mechanism to compose a complete medical report.

3.1. Encoder

3.1.1. Visual and semantic features

For a given medical image $\mathrm{Img}\in\mathbb{R}^{H\times W\times C}$, where H and W denote the image height and width, respectively, and C is the number of channels. Since the Transformer network can only accept one-dimensional embedding sequence input, we divide the image into M image blocks and flatten them into 2-dimensional vector $\mathbf{i}_p\in\mathbb{R}^{M\times(P^2\times C)}$. The resolution of each image block is $(P,\ P)$ and the total number of image blocks is $M = HW/P^2$. The first step is to extract the visual features and the primary semantic features, i.e., Img$\rightarrow$ $\mathbf{f}_v$, $\mathbf{f}_{s'}$. $\mathbf{f}_v\in\mathbb{R}^{M\times D}$ is the visual feature, which is obtained by projecting $\mathbf{i}_p$ to the D dimension space through a fully connected layer. The primary semantic feature $\mathbf{f}_{s'}\in\mathbb{R}^{D}$ is input to the label classifier to obtain the semantic feature $\mathbf{f}_s\in\mathbb{R}^{N\times D_1}$. Similar to the class token in the Bert network [30], we then concatenate a learnable position encoding vector $\mathbf{x}_{class}\in\mathbb{R}^{1\times D}$ and a one-dimensional position embedding vector $\mathbf{E}_{pos}\in\mathbb{R}^{(M+1)\times D}$ that carries position information, and input to the Transformer encoder ($\mathbf{z}_l$). The entire hierarchical LSTM network based encoder consists of L Transformer encoders, each of which contains a Multi-Layer Self-Attention (MSA) and a Multi-Layer Perceptron (MLP) network. LayerNorm (LN) and residual connections are added before MSA and MLP to reduce overfitting and prevent gradient disappearance. The visual feature and primary semantic feature vectors $\mathbf{f}_v$, $\mathbf{f}_{s'}$ are output by Transformer encoder [14] with a vector $\mathbf{z}_e = [\mathbf{x}_{class};\mathbf{x}_1;\mathbf{x}_2;\dots;\mathbf{x}_n]$, where $\mathbf{f}_v = [\mathbf{x}_1;\mathbf{x}_2;\dots;\mathbf{x}_n]$ and $\mathbf{f}_{s'} = [x_{class}]$.

$$\mathbf{z}_0 = [\mathbf{x}_{class};\mathbf{i}_1\mathbf{E};\mathbf{i}_2\mathbf{E};\dots;\mathbf{i}_n\mathbf{E}] + E_{pos} \tag{1}$$

$$\mathbf{z}'_l = MSA(LN(\mathbf{z}_{l-1})) + \mathbf{z}_{l-1}, l = 1,\dots,L \tag{2}$$

$$\mathbf{z}_l = MLP(LN(\mathbf{z}'_l)) + \mathbf{z}'_l, l = 1,\dots,L \tag{3}$$

$$\mathbf{z}_e = LN(\mathbf{z}_l) \tag{4}$$

The $\mathbf{f}_{s'_1}\in\mathbb{R}^{K\times D_1}$ is obtained through a K-dimensional fully-connection layer, where K is the number of tags' type in the dataset, D is the dimension of visual features, and D_1 is the dimension of semantic features.

3.1.2. Tag classification

The label classifier is constructed by connecting Z MIX-Block networks in series. The output of the former MIX-Block network is the input of the latter MIX-Block network. Each MIX-Block consists of two MLP networks, the first MLP (MLP1) network acting on the first dimension of $f_{s'_1}$ and the second MLP (MLP2) network acting on the second dimension of $f_{s'_1}$ as shown in Fig.3. Each MLP network contains two fully connected layers and a GELU activation function. $f_{s'_2}\in\mathbb{R}^{N\times D_1}$ is obtained after $f_{s'_1}$ passes through the full connection layer and the softmax function. The first dimension of $f_{s'_2}$ is the number of tags. The second dimension of $f_{s'_2}$, i.e., the occurrence probability of each tag, is sorted to select the top N tag tensor embeddings to obtain the semantic features f_s, which can be expressed as:

$$U_{*,i} = X_{*,i} + W_2\sigma(W_1 LayerNorm(X)_{*,i}) \tag{5}$$

$$Y_{j,*} = U_{j,*} + W_4\sigma(W_3 LayerNorm(U)_{j,*}) \tag{6}$$

$$f_{s'_2} = softmax(W_{tag}\theta^Z(f_{s'_1})) \tag{7}$$

$$f_s = embedding(\zeta(f_{s'_2})) \tag{8}$$

where W_* is the parameter matrix of the MLP network and σ is the GELU activation function. i, j are the dimensions of the hidden layer of the two MLP networks, which take values independent of the dimensionality of the feature vectors. $\theta^z(x)$indicates that x passes through Z MLP-Block layers, and ζ is the Topk function which selects the top N vectors after sorting $f_{s'_2}$.

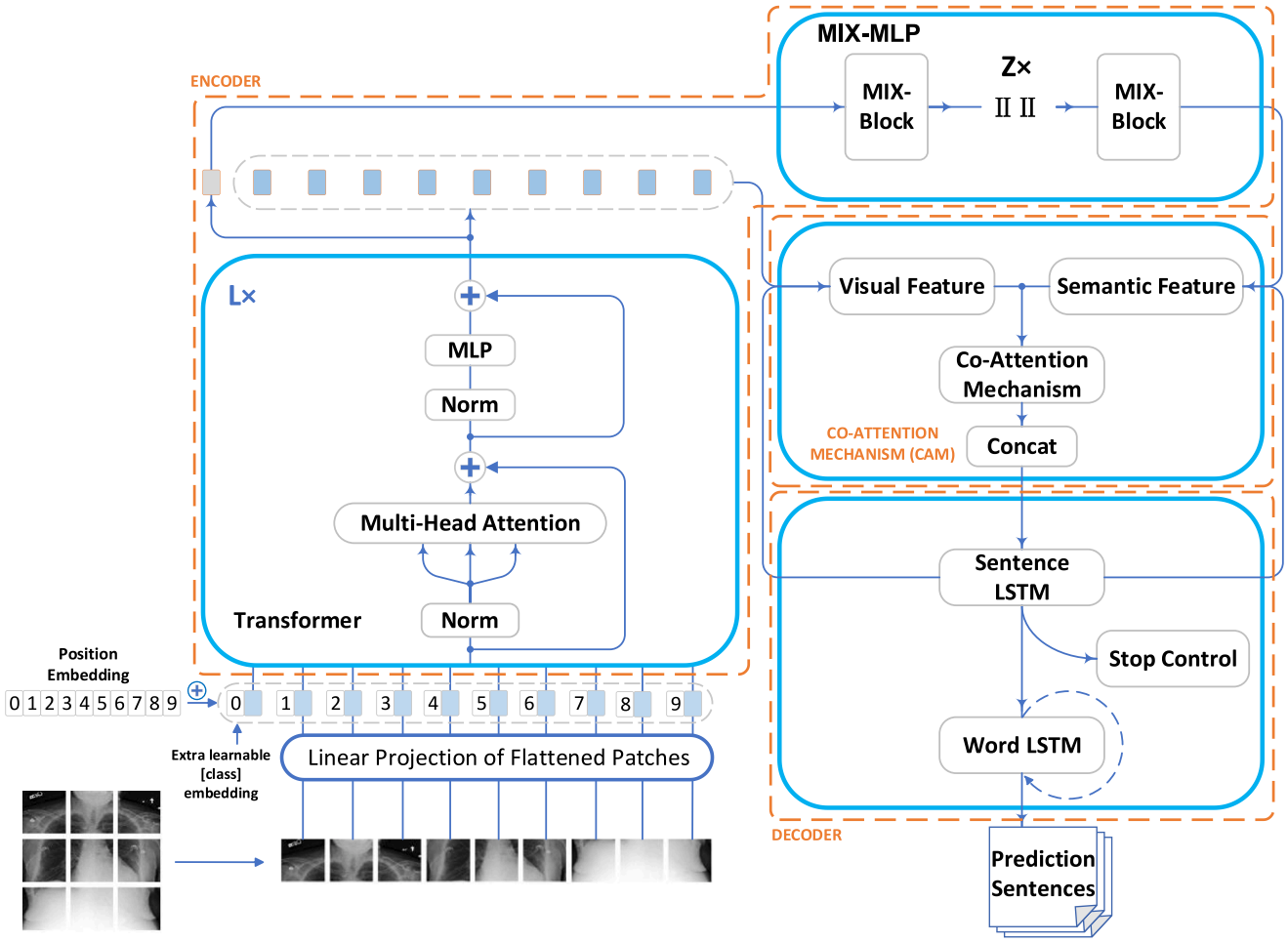

Fig. 2. The detail of the medical report generation model. The visual feature extractor is composed of L Transformer encoders and the multi-label classification network is composed of Z MIX-MLP block stacks.

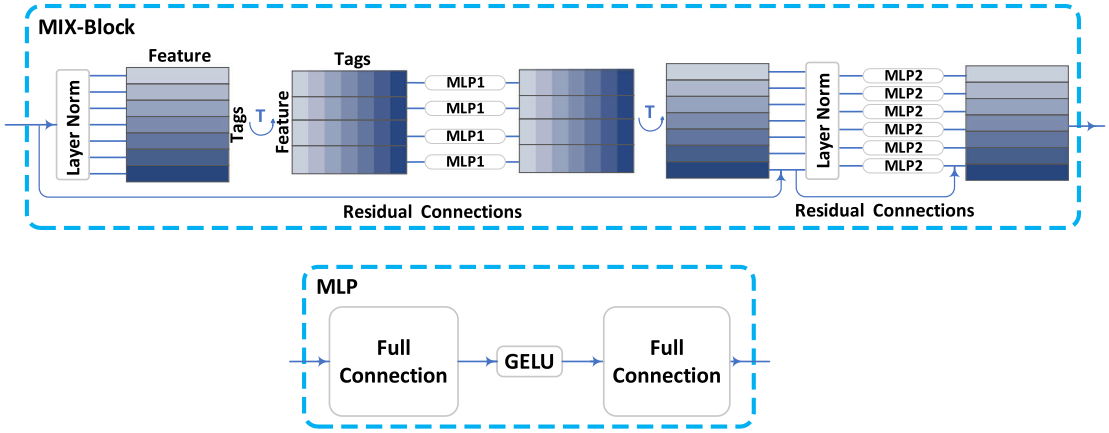

Fig. 3. The Mix-Block structure. It contains two MLPs, and each MLP network consists of two fully connected layers and a GELU activation function.

3.2. Co-attention mechanism

For the $f_v \in \mathbb{R}^{M\times D}$ and $f_s \in \mathbb{R}^{N\times D_1}$, which are output vectors from the encoder, we use an image-text matching mechanism to compute their similarities to the hidden layer state to better achieve visual-semantic feature alignment. Specifically, we calculate the cosine similarity $s_v^{m,t}, s_s^{n,t}$ between f_v, f_s and the hidden layer state $h_{sent}^{(t-1)} \in \mathbb{R}^{T\times D_2}$ of the Sentence LSTM network at the step t−1, respectively, by the following method.

$$s_v^{m,t} = \frac{f'_{v,m}(h'_{v,t})^T}{\|f'_{v,m}\|\|h'_{v,t}\|}, f'_{v,m} = BN(W_v f_{v,m}), \quad h'_{v,t} = BN(W_{v,h} h_{sent,t}^{(t-1)}) \tag{9}$$

$$s_s^{n,t} = \frac{f'_{s,n}(h'_{s,t})^T}{\|f'_{s,n}\|\|h'_{s,t}\|}, f'_{s,n} = BN(W_s f_{s,n}), \quad h'_{s,t} = BN(W_{s,h} h_{sent,t}^{(t-1)}) \tag{10}$$

where $m \in [1, M]$, $n \in [1, N]$, $t \in [1, T]$, D_2 is the dimension of the hidden layer, and BN is the batch normalization layer to control gradient explosion to prevent gradient disappearance and overfitting. $W_v, W_{v,h}$ are the parameter matrices of visual similarity and $W_s, W_{s,h}$ are the parameter matrices of semantic similarity. The normalized visual and semantic similarities are used to calculate the visual soft attention and semantic soft attention feature weights, $a_v^{m,t}, a_s^{n,t}$, respectively. For $a_v^{m,t}, a_s^{n,t}$, their soft attention feature vectors are calculated by the following formula,

$$v_{att}^{(t)} = \sum_{m=1}^{M} a_v^{m,t} f_v^{m,*} \tag{11}$$

$$s_{att}^{(t)} = \sum_{n=1}^{N} a_s^{n,t} f_s^{n,*} \tag{12}$$

After these two vectors are concatenated with a full-connected network function, $cof^{(t)} \in \mathbb{R}^{(D+D_1)}$ is obtained by the following formula,

$$cof^{(t)} = W_{fc}[v_{att}^{(t)}; s_{att}^{(t)}] \tag{13}$$

3.3. Decoder

Leveraging the similarity between the structure of medical image reports and the structure of hierarchical LSTM networks, the long text generation network is developed with a hierarchical LSTM network. The hierarchical LSTM network is composed of a Sentence LSTM and a Word LSTM. Both the Sentence LSTM and the Word LSTM are standard LSTM networks. The Sentence LSTM is a single-layer LSTM network that takes the feature vector $cof \in \mathbb{R}^{T\times(D+D_1)}$ obtained by the co-attention mechanism as the input, and generates the corresponding topic vector $top \in \mathbb{R}^{T\times D_2}$.

The Word LSTM is similar to the Sentence LSTM network, which is a standard LSTM network [14]. The first and second inputs are the topic vector $top^{(t)}$ generated by the word LSTM and a predefined starting token. The subsequent inputs are words of sentence. The hidden layer state is similarly used to predict the distribution of generated words $p(y_t|y_{1:t-1})$. After the Word LSTM generates its word sequence y_r^t, all generated sequences are concatenated to form the final report $y_r^{(1)}; \ldots; y_r^{(t)}$.

3.4. Loss calculation and model training

Considering that there are multiple loss calculations for each training sample, the losses are calculated separately and then summed to get the total loss. Consider each training sample as a tuple(*I*,*G*,*R*), where *I* is an image, *G* is the Ground Truth corresponding to the image *I*, and *R* is the generated report corresponding to the image *I*, consisting of T sentences, each consisting of S_i words. For each training sample (*I*,*G*,*R*), the model first calculates the probability distribution p_{tag} of the tag corresponding to the image *I* over all tags. Considering the sparsity of tag distribution, the focal loss(FL) function is used to calculate the loss of p_{tag} with the real tag. The focal loss function is a loss function to deal with sample classification imbalance,

$$fl(p_{tag}) = \frac{1}{N}\sum_{n=1}^{N} -\alpha(1 - p_{tag,\,n})^{\gamma} log(p_{tag,\,n}) \tag{14}$$

where N is the number of tags, γ is the sample difficulty adjustment factor, and α is the sample weight.

The Sentence LSTM operation is divided into T steps, and the probability distribution p_i' of the i-th sentence at each step at the two states {STOP, CONTINUE} is calculated. Then, the topic vectors are fed into the Word LSTM network to generate words $w_{i,\,j}$. Each generated word sequence is calculated by cross-entropy (CE) losses, the loss $loss_{sent}$ corresponding to the sentence number distribution $p_{stop,\,i}$ and the loss $loss_{word}$ corresponding to the word distribution $p_{i,j}$ for each sentence. Therefore, the total loss is calculated using the following formula,

$$loss(l,\ p,\ w) = \lambda_{tag} fl(p_{tag}) + \lambda_{sent}\sum_{i=1}^{T} loss_{sent}(p_i') + \lambda_{word}\sum_{i=1}^{T}\sum_{j=1}^{S_i} loss_{word}(p_{i,\,j}, w_{i,\,j}) \tag{15}$$

where λ_{tag}, λ_{sent}, λ_{word} are the pre-set weights of each loss. After we conduct several trials, we choose $\gamma = 3$, $\alpha = 0.2$, $\lambda_{tag} = 10$, $\lambda_{sent} = 1$ and $\lambda_{word} = 1$ for the experiments.

Table 1
Details of the IU X-RAY and MIMIC-CXR datasets.

Data set	IU X-RAY			MIMIC-CXR		
	Train	Validation	Test	Train	Validation	Test
Image	5976	747	747	368,960	2991	5159
Report	3165	395	395	222,758	1808	3269
Patient	3165	395	395	64,586	500	293
Average length of report	36.98	36.67	35.43	53.00	53.05	66.40

4. Experimental results and analysis

We evaluated the performance of the models on two datasets. In the following sub-sections, we described the two datasets, the evaluation metrics, the parameter configuration of the experiments, and the analysis of the results.

4.1. Dataset

IU X-RAY (The Indiana University Chest X-ray Collection) [31]: IU X-RAY was collected by Indiana University and used extensively to evaluate the performance of medical report generation models. The dataset contains 7470 image pairs including frontal and lateral X ray images, and 3955 radiology reports in English, each consisting of the following components: impressions, findings, tags, comparisons, and indications.

MIMIC-CXR [32]: MIMIC-CXR is the recently released largest dataset to date for performance evaluation of medical report generation models, including 377, 110 chest X-ray images and 227, 835 radiology reports in English for 64, 588 patients from Beth Israel Deaconess Medical Center. For these two datasets, the IU X-ray dataset uses a ratio of 8:1:1 to obtain the training, validation, and test sets after excluding the samples without text; MIMIX-CXR uses the officially given division sets. The details of the two datasets are shown in Table 1.

4.2. Evaluation metric

In this study, we utilized the evaluation metrics BLEU [33], METEOR [34], and ROUGE-L [35], which are widely used in the existing natural language generation (NLG) domain, where BLEU and METEOR are mainly used to evaluate machine translation performance, and ROUGE-L is used to evaluate the summary generation performance. BLEU-n denotes BLEU score using up to n-grams. These metrics are mainly used to assess the similarity between the generated reports and Ground Truth but they do not reflect anomalies in the generated reports. Therefore, to measure the accuracy of abnormality descriptions, we used the CheXpert tool [36] to tag the reports in the MIMIC-CXR dataset [32], which extracts 14 different categories of labels related to chest disease from each report. Comparisons with model-generated labels were made. Precision, Re-call and F1 scores were calculated as Clinical Efficacy metrics (CE metrics).

Table 2
Experimental parameter settings.

Parameter	Value	Parameter	Value
Visual feature dimensions D	768	Number of image blocks M	49
Semantic feature dimensions D_1	256	Select the number of tags N	10
Sentence LSTM hidden layer dimension D_2	512	Number of subject vectors T	5
Word LSTM hidden layer dimension D_3	512	Tag loss weight λ_{tag}	10
Number of MLP-Block Z	2	Sentence loss weight λ_{sent}	1
MLP-Block hidden layer dimension i	768	Word loss weight λ_{word}	1
MLP-Block hidden layer dimension j	768	Stop to control the input dimension	64
Sample difficulty regulator γ	3	Stop the control threshold	0.5
Sample weight α	0.2	Batch size	16
Image block size P	32	Learning ratio	1e−5

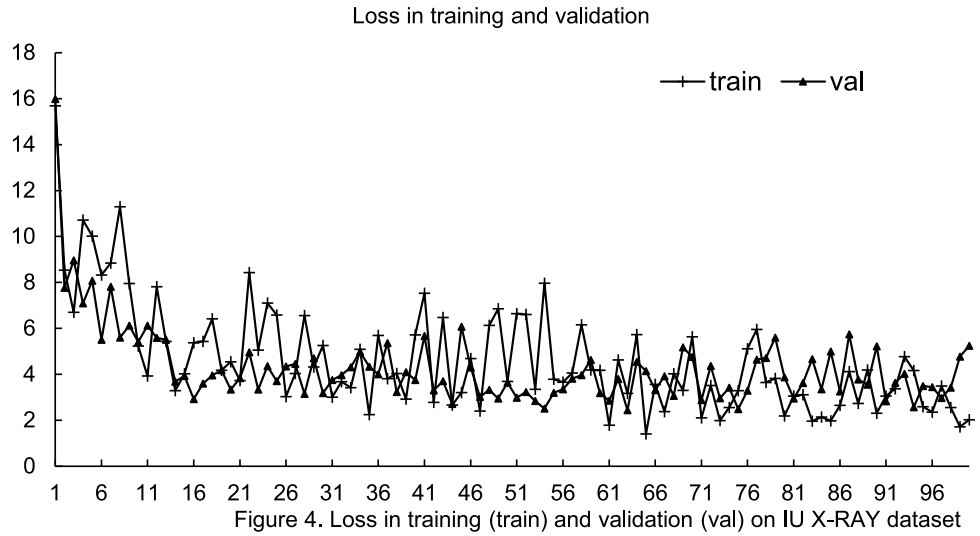

Fig. 4. Loss in training (train) and validation (val) on IU-X-RAY dataset.

Fig. 5. Loss in training (train) and validation (val) on MIMIC-CXR dataset.

4.3. Parameter settings

In the encoder, we used pre-trained Vit [29] on ImageNet as a visual feature extractor to obtain visual and semantic features. The input dimension of the image is 224 × 224, the patch size is set to 32, the visual feature dimension is 768, and the number of image blocks is 49. The number of MLP-Block is set to 2, where the two hidden dimensions of MLP are 768, the weight λ_{tag} is set to 10, the rest of the weights are set to 1, and the number of selected tags N=10, γ = 3, α = 0.2 for focal loss. In the decoder, the dimension of word embedding is set to 512, the stop control threshold of the component is set to 0.5, the input dimension is 64, and the model uses ADAM optimizer to train the model. The Batch size is set to 16 and the learning rate of the encoder and decoder is set to 1e−5. The detailed experimental parameters are set as shown in Table 2.

4.4. Results and analysis

4.4.1. Losses and evaluation scores in training and validation

Our experiments are carried on a desktop PC, containing an Intel Xeon E5-2680 v3 CPU at 2.5 GHz with a 128G of RAM and a Nvidia 3080 Ti GPU with a 10 GB of VRAM. Time cost for each training epoch is 120 s for IU X-ray and 5880 s for MIMIX-CXR, respectively. The losses in training and validation on IU X-RAY and MIMIC-CXR datasets are shown in Figs. 4 and 5, respectively. Because of the size difference between two datasets and the successive computation load of training,

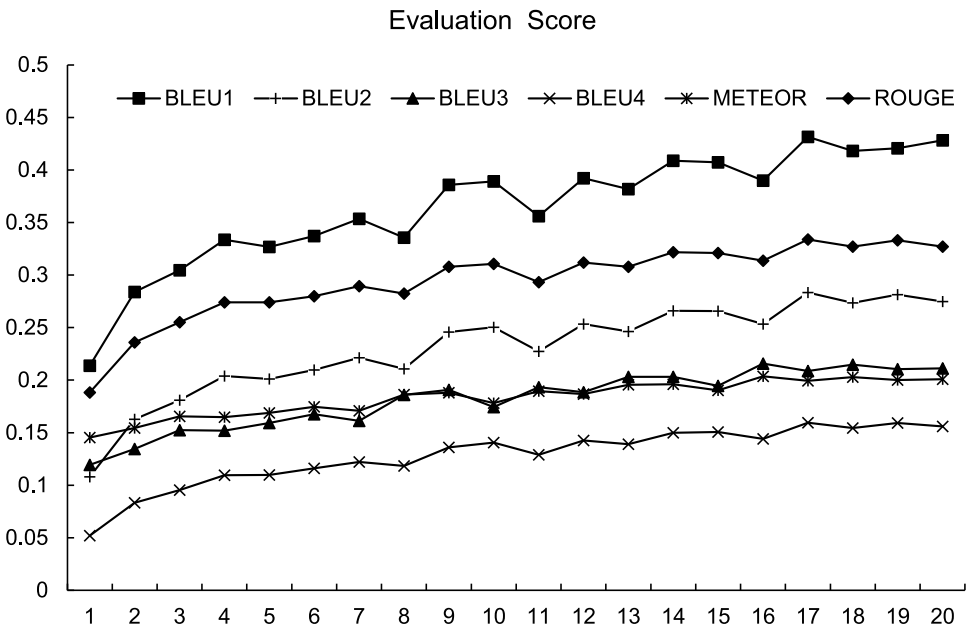

Fig. 6. Evaluation Score on MIMIC-CXR dataset.

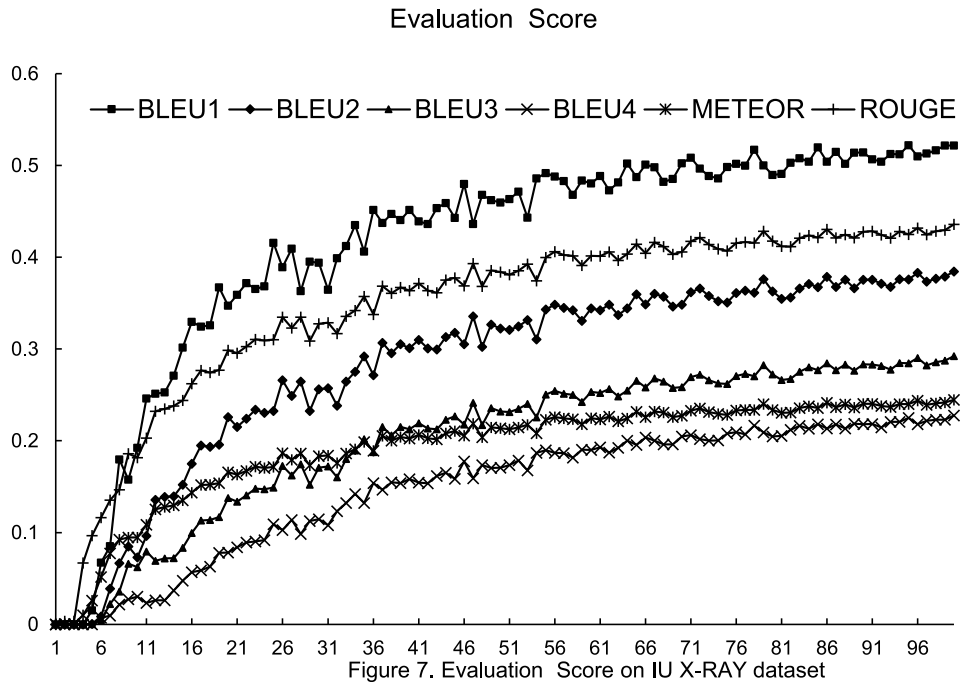

Fig. 7. Evaluation Score on IU X-RAY dataset.

the training epoch of IU X-RAY and MIMIC-CXR dataset is 100 and 20, respectively. As shown in Figs. 4 and 5, both the losses of training and validation decrease with the increase of training epoch. However, after the training epoch reaches around 61 for IU X-RAY dataset, and around 18 for MIMIC-CXR dataset, the decrease of losses is very small, or rather remains oscillatory. That is also the reason that we stop the training at the epoch 20 and 100, respectively.

The evaluation scores on IU X-RAY and MIMIC-CXR datasets are shown in Figs. 6 and 7, respectively. All the six metrics increase with the increase of training epoch. However, after the epoch reaches around 17 for MIMIC-CXR dataset, the increase is very small, or the maximum is achieved. For IU X-ray dataset, the six metrics increase till 100 training epochs, and then the increase in magnitude is smaller and smaller.

4.4.2. Overall experiment results

We compared the performance between our framework and the current mainstream models on these two datasets (Table 1), including Co-Att [8], HGRG [6], R2Gen [37], PPKED [38], M2transformer [39], SBF [40], KGAE [5], and PT [41], as shown in Table 3. The results of the R2Gen model are reproduced under the same experimental conditions as ours based on published source codes. For the rest of the models, we directly quote the experimental results from their papers because the authors did not publish the source codes or the division ratio of the experimental datasets. The models′ performance is lower

Table 3
NLG metrics on both IU X-RAY and MIMIC-CXR datasets.

dataset	model	BLEU 1	BLEU 2	BLEU 3	BLEU 4	METEOR	ROUGE-L
IU-XRAY	Co-Att [8]	0.455	0.288	0.205	0.154	–	0.369
	HGRG [6]	0.438	0.298	0.208	0.151	–	0.322
	R2Gen [33]	0.451	0.293	0.209	0.159	0.185	0.381
	PPKED [35]	0.483	0.315	0.224	0.168	–	0.376
	M2transformer [36]	0.486	0.317	0.232	0.173	0.192	0.390
	SBF [37]	0.487	0.346	0.270	0.208	–	359
	K GAE [5]	0.512	0.327	0.240	0.179	0.195	0.383
	PT [38]	0.496	0.319	0.241	0.175	–	0.377
	Ours	**0.521**	**0.384**	**0.292**	**0.227**	**0.244**	**0.435**
MIMIC-CX R	R2Gen [33]	0.352	0.218	0.145	0.103	0.142	0.277
	PPKED [35]	0.360	0.224	0.149	0.106	0.149	0.284
	M2transformer [36]	0.378	0.232	0.154	0.107	0.145	0.272
	KGAE [5]	0.369	0.231	0.156	0.118	0.153	0.295
	PT [38]	0.351	0.233	0.157	0.118	–	0.287
	Ours	**0.428**	**0.274**	**0.211**	**0.155**	**0.201**	**0.327**

Table 4
CE metrics on MIMIC-CXR datasets.

model	Precision	Re-call	F1
M2transformer [36]	0.240	0.428	0.308
R2GEN [33]	0.333	0.273	0.276
KGAE [5]	0.389	0.362	0.355
Ours (CE)	0.646	0.397	0.308
Ours (FL)	**0.662**	**0.440**	**0.528**

Table 5
Performance of our framework with different components.

	The number of parameters	BLEU4	METEOR	ROUGE
Base	101.9M	0.107	0.164	0.283
Base+ Co	109.4M	0.136	0.182	0.302
Base+ Co+MIX	123.2M	0.152	0.198	0.324
Base+ Co+MIX+FL	123.2M	0.155	0.201	0.327

on MIMIC-CXR dataset than that on IU X-RAY dataset since MIMIC-CXR dataset contains longer reports and more samples. It can be observed that our framework outperforms other models in terms of evaluation metrics on both IU X-ray and MIMIC-CXR datasets.

We also compared the performance in terms of CE evaluation metrics between our framework and other models on MIMIC-CXR dataset, including M2transformer, R2GEN and KGAE models, as shown in Table 4. The experimental results are comparable because they all use the officially divided dataset. It can be seen that our framework achieves a higher performance in terms of CE metrics in multi-label classification, compared with the M2transformer, R2GEN, and KGAE models that use MLP. One of the possible reasons is that the transposition operation of MIX-MLP in our framework makes the network have some spatial learning ability similar to that of CNN networks, which enables our framework to use both the association between labels in different samples and the association between different labels within the same sample.

Chen et al. [37] mentioned that the performance in terms of CE evaluation metrics was low because of the sparse distribution of the labels as well as the distribution imbalance. After we switched to the Focal Loss(FL) function, the performance was improved by 1.6%, 3.3%, and 3.7% in terms of Precision, Re-call, and F1, respectively.

4.4.3. Ablation experiments

We investigated the impacts of different components in our framework on the model performance, including the co-attention mechanism, MIX-MLP network and focal loss function, respectively. A base model (Base) is developed at first, comprised of the Transformer encoder and hierarchical LSTM decoder, in which only visual information, cross-entropy loss function and MLP network are employed. Then, a Base+Co model is developed by leveraging Base model with the label information and the concurrent attention mechanism (Co). Moreover, a Base+Co+MIX model is developed by adding the MIX-MLP network (MIX) onto Base+Co model. Furthermore, a Base+Co+MIX+FL model (Full) is developed using the Focal Loss (FL) function to replace the cross-entropy (CE) loss function. Ablation experiments are performed on the MIMIC-CXR dataset in this study. The results are shown in Table 5. It can be observed that after different components are added, the performance of different models shows an increasing trend in order, which reflects to some extent that the added components have helped to improve the performance of our framework.

4.4.4. Qualitative analysis of some examples

We also conducted a qualitative analysis of some cases, in which two reports are generated by the Base model versus the Full model. Fig. 8 shows two examples of reports generated from chest X-ray images in the MIMIC-CXR and IU-XRAY datasets. Different colors on the text indicate the differences between the generated reports and Ground Truth, where the red sentence indicates that it is incorrectly predicted, blue sentence indicates that it is not fully consistent with ground truth, and green sentence indicates that it is largely consistent with ground truth. Underlined words indicate that anomalous keyword is different between the sentences generated by two models. The difference of anomalous keywords causes the misunderstanding because the meaning of the generated sentences is completely inconsistent with the original meaning of the report.

It can be observed that the generated reports by Base+Co +MIX+FL comply with Ground truth. For examples, these sentences are correctly predicted, including “the patient is status post mitral valve replacement and probably coronary artery bypass graft surgery”, “there is no pneumothorax”, “the heart is mildly enlarged”. etc. For the Base model, these sentences do not match the original meaning of the report due to keyword errors, e.g., “patient” and “heart” are misclassified as “lung”. Additionally, some findings are missed in the generated reports, such as “no pneumothorax” and “no acute bony abnormalities”. Nevertheless, it is observed that there are some reported anomalies that are not reflected in the generated reports for both models, such as “no free air is demonstrated”. We speculate that this anomaly appears just few times in MIMIC-CXR dataset (a total of 22 times in the training set), resulting in the network that do not learn the features of this anomaly well.

Generally, it is difficult to generate reports directly from medical images. By decomposing the task of report generation into feature extraction from medical images and using extracted features as prompts to generate reports, the difficulty of the task can be reduced and the quality of the generated reports can be improved [42]. The superior performance of our framework that integrates label information may be due to multi-dimensional information integration. A model that

	Ground truth	Base	Base+Co+MIX+FL
	the patient is status post mitral valve replacement and probably coronary artery bypass graft surgery. the heart is mildly enlarged. there is patchy basilar opacification suggesting a combination of atelectasis and pleural effusion. streaky left upper lobe opacity suggests minor atelectasis or scarring which is unchanged. there is no pneumothorax. no free air is demonstrated.	the lung is status post median valve replacement median unchanged artery bypass graft surgery. the lung is mildly enlarged. there is no opacity which minor component of atelectasis and pleural effusion. the opacities basilar lobe opacity is minor atelectasis scarring is unchanged. there is no pleural. no pleural air below seen.	the patient is status post median valve replacement and coronary artery bypass graft surgery. the heart is mildly enlarged. there is no opacity of minor combination of atelectasis and effusion. there opacities basilar lobe opacity is minor atelectasis scarring is unchanged. there is no pneumothorax. no pneumothorax air below seen.
	no acute cardiopulmonary findings the heart size and mediastinal contours appear within normal limits. there are low lung volumes with left basilar subsegmental atelectasis. no focal airspace consolidation effusions or pneumothorax. no acute bony abnormalities.	no acute cardiopulmonary abnormality the heart size and mediastinal contours appear within normal limits. there is no lung volumes with broncho vascular basilar atelectasis. no focal consolidation or pneumothorax.	no acute cardiopulmonary abnormality heart size and mediastinal contours appear within normal limits. there is low lung volumes with broncho vascular basilar atelectasis. no focal airspace consolidation or pneumothorax. no pneumothorax bony abnormality.

Fig. 8. Examples of generated reports on MIMIC-CXR and IU-XRAY datasets. The output in red indicates the sentences with wrong predictions, the output in blue indicates the sentences that are not fully consistent with the ground truth, the output in green indicates the sentences that are basically consistent with the ground truth, and the underlined words indicate the keyword differences between our framework and the Base model. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

integrates label information can consider both image information and related label information, and this integration of multidimensional information enables the model to have a more comprehensive understanding of the context and content of the image. The label information provides relevant information about the image content. By integrating the label information, the model can understand specific structures or lesions in the image, thereby generating more accurate and detailed reports. Additionally, label information is used to reduce ambiguity in image interpretation, which helps to improve the performance of our framework. For example, when analyzing breast X-ray images, label information may include breast tissue density, tumor location, etc. This information helps the model to describe images more accurately and reduce possible blurring or uncertainty in medical image reports.

5. Discussion

In this paper, we proposed a framework for medical image report generation, using a Transformer encoder as a visual feature extractor and a MIX-MLP network as a multi-label classifier. In the framework, a co-attention mechanism is also used to align and fuse visual and semantic features. The proposed framework is effectively trained under two criteria (NLG and CE evaluation metrics), ensuring that our framework has good visual and textual feature extraction capability with high multi-label classification performance. This study is one among a few studies addressing the problem of medical image report generation [8,9,25,37–43]. We tried our best effort to make a fair comparison with the state-of-the-art models. Experimental results on two publicly available datasets demonstrate that our framework has a superior performance over other models. We also performed ablation experiments to verify the impact of different components on model performance, showing that our framework is capable of generating long reports with correct medical terminology and meaningful image-to-text attention mapping.

Although our framework has achieved good performance in generating medical reports, it has some limitations. First, the labels related to chest disease are tagged by the CheXpert tool. A better solution is to label the report by doctors, which will improve the performance of the whole report generation system. Second, there are also some missed outliers that are not reflected in the generated reports. In the future, the dataset can be divided into smaller subsets to ensure that the frequency of each anomalous keyword is basically the same. It is also possible to augment the samples with fewer occurrences, and then train the framework sequentially from low to high difficulty [43]. Additionally, for the decoder part, an attempt can be made to generate medical reports using pre-trained language models [44] as word embeddings, which can effectively improve the fluency of the generated reports. Another option for improving the decoder is to adopt LLMs [45], but we need to fine-tune them for a given category medical image. Finally, knowledge graphs [9] can be introduced into our proposed framework to improve the quality of features extracted from images since knowledge graphs can learn the relationships between abnormal keywords, which is very helpful to generate a high-quality medical report.

CRediT authorship contribution statement

Shuifa Sun: Conceptualization, Formal analysis, Methodology, Supervision, Validation, Writing – original draft. **Zhoujunsen Mei:** Conceptualization, Data curation, Formal analysis, Methodology, Resources, Software, Validation, Writing – original draft. **Xiaolong Li:** Data curation, Software, Validation, Visualization. **Tinglong Tang:** Data curation, Investigation, Project administration, Resources, Visualization, Writing – review & editing. **Zhanglin Su:** Data curation, Software, Visualization, Writing – review & editing. **Yirong Wu:** Funding acquisition, Methodology, Project administration, Supervision, Visualization, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known conflict of interests that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was partly supported by the Chinese National Social Science Foundation Project (Grant No. 20BTQ066).

References

[1] European Society of Radiology (ESR). Good practice for radiological reporting. Guidelines Eur Soc Radiol (ESR) Insights Imaging 2011;2(2):93–6.
[2] Shin HC, Roberts K, Lu L, et al. Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. Piscataway, NJ: IEEE; 2016, p. 2497–506.
[3] Yin C, Qian B, Wei J, et al. Automatic generation of medical imaging diagnostic report with hierarchical recurrent neural network. In: 2019 IEEE international conference on data mining. Piscataway, NJ: IEEE; 2019, p. 728–37.
[4] Li CY, Liang X, Hu Z, et al. Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, (1):Menlo Park: AAAI; 2019, p. 6666–73.
[5] Liu F, You C, Wu X, et al. Auto-encoding knowledge graph for unsupervised medical report generation. Adv Neural Inf Process Syst 2011;34(1):16266–79.
[6] Li CY, Liang X, Hu Z, et al. Hybrid retrieval-generation reinforced agent for medical image report generation. In: Proceedings of the 32nd international conference on neural information processing systems. Red Hook, NY: Curran Associates Inc; 2018, p. 1537–47.
[7] Vinyals O, Toshev A, Bengio S, et al. Show and tell: A neural image caption generator. In: Proceedings of the IEEE conference on computer vision and pattern recognition. Piscataway, NJ: IEEE; 2015, p. 3156–64.
[8] Jing B, Xie P, Xing E. On the automatic generation of medical imaging reports. In: Proceedings of the 56th annual meeting of the association for computational linguistics (Volume 1: Long Papers). Melbourne, Australia: Association for Computational Linguistics; 2018, p. 2577–86.
[9] Zhang Y, Wang X, Xu Z, et al. When radiology report generation meets knowledge graph. In: Proceedings of the AAAI conference on artificial intelligence, vol. 34, (1):Menlo Park: AAAI; 2020, p. 12910–7.
[10] Wu J, Chen T, Wu H, et al. Fine-grained image captioning with global-local discriminative objective. IEEE Trans Multimed 2020;23(1):2413–27.
[11] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Proceedings of the 31st international conference on neural information processing systems. Red Hook, NY: Curran Associates Inc; 2017, p. 6000–10.
[12] Tolstikhin IO, Houlsby N, Kolesnikov A, et al. MLP-mixer: An all-MLP architecture for vision. Adv Neural Inf Process Syst 2021;34(1):24261–72.
[13] Zhou Y, Wang M, Liu D, et al. More grounded image captioning by distilling image-text matching model. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Piscataway, NJ: IEEE; 2020, p. 4777–86.
[14] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In: International conference on learning representations. 2020, p. 1–21, openreview.net: ICLR.
[15] Lin TY, Goyal P, Girshick R, et al. Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. Piscataway, NJ: IEEE; 2017, p. 2980–8.
[16] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: International conference on learning representations. San Diego; 2015, p. 1–14, openreview.net: ICLR.
[17] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. Piscataway, NJ: IEEE; 2016, p. 770–8.
[18] Yang K, Qinami K, LiF F, et al. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In: Proceedings of the 2020 conference on fairness, accountability, and transparency. New York, NY: Association for Computing Machinery; 2020, p. 547–58.
[19] Dai Y, Gao Y, Liu F. Transmed: Transformers advance multi-modal medical image classification. Diagnostics 2021;11(8):1384–92.
[20] Ba J, Mnih V, Kavukcuoglu K. Multiple object recognition with visual attention. In: ICLR (Poster). 2015, openreview.net: ICLR.
[21] Xu K, Ba JL, Kiros R, et al. Show, attend and tell: neural image caption generation with visual attention. In: Proceedings of the 32nd international conference on international conference on machine learning, vol. 37, 2015, p. 2048–57, JMLR.org: JMLR.
[22] Liu C, Mao Z, Liu AA, et al. Focus your attention: A bidirectional focal attention network for image-text matching. In: Proceedings of the 27th ACM international conference on multimedia. New York, NY: ACM; 2019, p. 3–11.
[23] Wang H, Zhang Y, Ji Z, et al. Consensus-aware visual-semantic embedding for image-text matching. In: European conference on computer vision. Berlin: Springer, Cham; 2020, p. 18–34.
[24] Lee KH, Chen X, Hua G, et al. Stacked cross attention for image-text matching. In: European conference on computer vision. Berlin: Springer, Cham; 2018, p. 212–28.
[25] Yuan J, Liao H, Luo R, et al. Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. In: International conference on medical image computing and computer-assisted intervention. Berlin: Springer, Cham; 2019, p. 721–9.
[26] Liu Y, Han T, Ma S, et al. Summary of chatGPT/GPT-4 research and perspective towards the future of large language models. Radiology 2023;1(2):100017.
[27] Touvron H, Lavril T, Izacard G, et al. Llama: open and efficient foundation language models. 2023, arXiv preprint arXiv:230213971.
[28] Liu Z, Li Y, Shu P, et al. Radiology-llama2: best-in-class large language model for radiology. 2023, arXiv preprint arXiv:230906419.
[29] Krause J, Johnson J, Krishna R, et al. A hierarchical approach for generating descriptive image paragraphs. In: Proceedings of the IEEE conference on computer vision and pattern recognition. Piscataway, NJ: IEEE; 2017, p. 317–25.
[30] Devlin J, Chang MW, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Human language technologies, Volume 1 (Long and Short Papers). Minnesota: Association for Computational Linguistics; 2019, p. 4171–86.
[31] Demner-Fushman D, Kohli MD, Rosenman MB, et al. Preparing a collection of radiology examinations for distribution and retrieval. J Am Med Inform Assoc 2016;23(2):304–10.
[32] Johnson AEW, Pollard TJ, Berkowitz SJ, et al. MIMIC-CXR, A de-identified publicly available database of chest radiographs with free-text reports. Sci Data 2019;6(1):1–8.
[33] Papineni K, Roukos S, Ward T, et al. Bleu: A method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the association for computational linguistics. United States: Association for Computational Linguistics; 2002, p. 311–8.
[34] Denkowski M, Lavie A. Meteor 1.3: automatic metric for reliable optimization and evaluation of machine translation systems. In: Proceedings of the sixth workshop on statistical machine translation. United States: Association for Computational Linguistics; 2011, p. 85–91.
[35] Lin CY, Hovy E. Manual and automatic evaluation of summaries. In: Proceedings of the ACL-02 workshop on automatic summarization, vol. 4, United States: Association for Computational Linguistics; 2002, p. 45–51.
[36] Irvin J, Rajpurkar P, Ko M, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI conference on artificial intelligence, vol. 33, (1):Menlo Park: AAAI; 2019, p. 590–7.
[37] Chen Z, Song Y, Chang TH, et al. Generating radiology reports via memory-driven transformer. In: Proceedings of the 2020 conference on empirical methods in natural language processing. United States: Association for Computational Linguistics; 2020, p. 1439–49.
[38] Liu F, Wu X, Ge S, et al. Exploring and distilling posterior and prior knowledge for radiology report generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Piscataway, NJ: IEEE; 2021, p. 13753–62.
[39] Nooralahzadeh F, Gonzalez NP, Frauenfelder T, et al. Progressive transformer-based generation of radiology reports. In: Findings of the association for computational linguistics: EMNLP 2021. United States: Association for Computational Linguistics; 2021, p. 2824–32.
[40] Wang Z, Zhou L, Wang L, et al. A self-boosting framework for automated radiographic report generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Piscataway, NJ: IEEE; 2021, p. 2433–42.
[41] Wang Z, Han H, Wang L, et al. Automated radiographic report generation purely on transformer: A multi-criteria supervised approach. IEEE Trans Med Imaging 2022;41(10):2803–13.
[42] Wang X, Peng Y, Lu L, et al. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In: Proceedings of the IEEE conference on computer vision and pattern recognition. Salt Lake City, UT: IEEE; 2018, p. 9049–58.
[43] Liu F, Ge S, Wu X. Competence-based multimodal curriculum learning for medical report generation. In: Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (Volume 1: Long Papers). United States: Association for Computational Linguistics; 2021, p. 3001–12.
[44] Lee J, Yoon W, Kim S, et al. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020;36(4):1234–40.
[45] OpenAI. GPT-4 technical report. 2023, arXiv:2303.08774.