# Generating Radiology Reports via Multi-feature Optimization Transformer

**Rui Wang[1], and Rong Hua[1*]**
[1] School of Computer Science and Engineering, Shandong University of Science and Technology
Qingdao, 266590, China
[e-mail: 1158157431@qq.com, huarong@sdust.edu.cn]
*Corresponding author: Rong Hua

## *Abstract*

As an important research direction of the application of computer science in the medical field, the automatic generation technology of radiology report has attracted wide attention in the academic community. Because the proportion of normal regions in radiology images is much larger than that of abnormal regions, words describing diseases are often masked by other words, resulting in significant feature loss during the calculation process, which affects the quality of generated reports. In addition, the huge difference between visual features and semantic features causes traditional multi-modal fusion method to fail to generate long narrative structures consisting of multiple sentences, which are required for medical reports. To address these challenges, we propose a multi-feature optimization Transformer (MFOT) for generating radiology reports. In detail, a multi-dimensional mapping attention (MDMA) module is designed to encode the visual grid features from different dimensions to reduce the loss of primary features in the encoding process; a feature pre-fusion (FP) module is constructed to enhance the interaction ability between multi-modal features, so as to generate a reasonably structured radiology report; a detail enhanced attention (DEA) module is proposed to enhance the extraction and utilization of key features and reduce the loss of key features. In conclusion, we evaluate the performance of our proposed model against prevailing mainstream models by utilizing widely-recognized radiology report datasets, namely IU X-Ray and MIMIC-CXR. The experimental outcomes demonstrate that our model achieves SOTA performance on both datasets, compared with the base model, the average improvement of six key indicators is 19.9% and 18.0% respectively. These findings substantiate the efficacy of our model in the domain of automated radiology report generation.

**Keywords:** Attention mechanism; Feature fusion; Radiology report; Transformer; Image-text generation

# 1. Introduction

**R**adiology examination is a crucial component of clinical diagnosis that helps doctors assess the physical condition of patients. The result of a radiology examination consists of a set of radiology images and a corresponding report, where the report is the initial diagnosis made by the radiologist for the patient based on the radiology images. In clinical work, the attending physician makes the final diagnosis based on the results of the radiology examination, combined with clinical symptoms and other examination results. With the growing population, doctors are exhausted by the increasing task of writing radiology reports. In order to alleviate the clinical workload on doctors, automatic generation of radiology reports using deep learning models have attracted the attention of academia and industry.

Most of the automatic generation models of radiology reports are derived from the image captioning [1], both involve the process of converting visual information into natural language. Therefore, these two tasks share many similarities in implementation process and technology. The image captioning task requires the computer to understand the content of the image and automatically generate the corresponding description. In recent years, with the widespread application of mass storage devices, the amount of information has increased exponentially, data-driven artificial intelligence technology has become increasingly mature, and the research on image captioning has also achieved some good results [2-4]. Compared with the multi-angle and multi-scene images used in image captioning, radiology images usually come from the human body structure with fixed scenes and fixed angles, and the similarity between images is high. The normal region in the image occupies the main part, and the characteristics of abnormal regions and normal regions of the image are extremely unbalanced, the abnormal region features learned by the model during the training process are easy to be masked by the normal region features, resulting in the loss of key and primary features in the model operation process, which affects the model's prediction ability for key words and primary words. Therefore, it is necessary to design additional modules for the model to carry out corresponding feature enhancement and alleviate the negative impact caused by data imbalance. As shown in **Fig. 1**, a standard radiology report contains two parts: report and tags. The words in the report part can be divided into key words, primary words, and common abstract words that do not belong to the two parts. Among them, the key words are mainly composed of medical professional words with summative characteristics, and the primary words are mainly composed of adjectives describing medical words. In addition, the labels extracted from the report cannot summarize the medical words of the entire report, so there are certain limitations in the application of label features in model optimization. In addition, the task of automatic radiology report generation requires the model to have the ability to generate long narratives composed of multiple sentences to describe the comprehensive condition of the patient. This requires the model to be able to accurately establish the mapping relationship between visual features and semantic features, reduce the negative impact caused by feature differences, and generate text that meets the structural requirements of radiology reports.
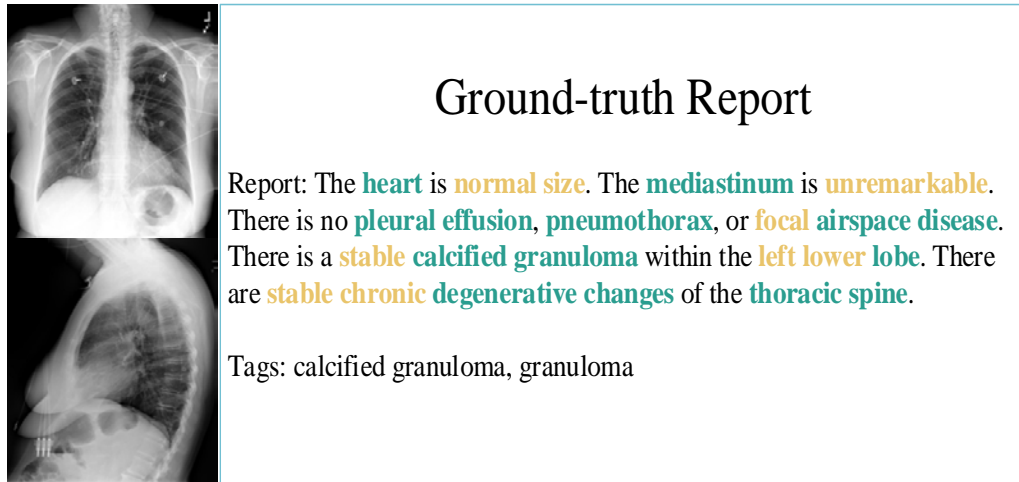
**Ground-truth Report**

Report: The **heart** is **normal size**. The **mediastinum** is **unremarkable**. There is no **pleural effusion**, **pneumothorax**, or **focal airspace disease**. There is a **stable calcified granuloma** within the **left lower lobe**. There are **stable chronic degenerative changes** of the **thoracic spine**.

Tags: calcified granuloma, granuloma

**Fig. 1.** A standard lung radiology medical report from the IU X-Ray dataset, where yellow labeled words indicate key words, green labeled words indicate key words, and Tags represent the corresponding tag words.

In order to alleviate the loss of key and primary features and strengthen the model's ability to understand the relationship between semantic and visual features, we propose to use the MFOT to generate radiology reports. Specifically, the contributions of this paper can be summarized in the following three aspects:

Firstly, we design an MDMA module to perform more comprehensive coding calculation based on the internal structure of visual grid features to reduce the loss of primary features in the encoding process. Secondly, we propose an FP module to fuse semantic features and visual features, improve the model's ability to learn the mapping relationship between different features, reduce the disturbance caused by feature differences, and generate radiology reports with good image alignment. Finally, we propose a DEA module to extract the fine-grained key features, and calculate the detail-enhanced attention according to the extracted key features to reduce the loss of key features, improve the prediction ability of key words, and generate more accurate radiology reports.

## 2. Related Work

### 2.1 Machine Learning in Radiology

Machine learning is the core research direction in the computer field [5], and its good universality directly promotes the cross-research between computer science and other disciplines, and has a wide range of application prospects [6-7]. Radiology is an important subject in the field of medicine. In clinical application, radiology is used to perform radiology diagnosis and radiotherapy. Among them, radiology diagnosis and computer model training can use pictures as data-driven, so there is a natural advantage in the intersection of radiology and computer science. At present, remarkable achievements have been made in the fields of automatic generation of radiology reports [8-17], prediction of relative regional air volume changes in the lungs [18-19], automatic detection of pulmonary nodules [20-21], and radiology image segmentation [22-24]. Among them, the automatic generation model of radiology reports has a particularly significant practical value in clinical applications, which not only

reduces the work pressure of radiologists, but also improves the medical level in developing countries and remote areas, and alleviates the imbalance of global medical resources.

## 2.2 Image Captioning

The automatic generation task of radiology report is derived from the image captioning, which is one of the most active research directions in the field of artificial intelligence. Its purpose is to enable the computer to automatically generate a natural language description sentence according to the given image. Current research has achieved good results [2-4, 25-30]. Among them, Liu et al. [26] designed an attention mechanism for the semantics of coherence attributes and related image regions to deepen the model's understanding of depth image features. Pan et al. [30] designed an attention module that explored the interaction of features within and between modalities, so as to enhance the ability of the model to utilize multi-modal features. Anderson et al. [2] proposed an attention mechanism that coordinates the internal work of the model, so that the model accurately judged the areas that need to be paid attention to and automatically selected the corresponding tensor space when generating different words. However, it neglected the feature differences between different features and did not effectively align the visual images and the abstract words. Huang et al. [28] extended the conventional attention mechanism and built a mapping matrix for the correlation vector and the given attention query vector to strengthen the reasoning ability of the model, thereby reducing the loss of primary features. However, its use of the query vector was too direct, resulting in too many disturbance factors mixed into the attention results. Wang et al. [29] used the method of multi-feature pre-fusion to optimize multi-modal features, but it did not carry out additional fusion measures for the initial semantic features. The model did not adjust the overall structure of the two features, resulting in insufficient interaction between the front and back of the model, and eventually produced feature deviation.

## 2.3 Automatic Radiology Report Generation

Radiology reports include specialized descriptions of multiple areas of the image, which require detailed and accurate wording, and the current models used for image description tasks are not well adapted to this task. In order to generate radiology reports that meet the requirements of clinical application, researchers conducted a series of experiments based on the characteristics of radiology reports and obtained good results [8-17, 31-33]. Most of the initial studies used Recurrent Neural Network (RNN) as decoder to generate reports. Liu et al. [8] first predicted the topics of the report, and then generated corresponding sentences according to these topics to meet the structural requirements of radiology reports. Jing et al. [9] designed a multi-task learning framework that allows models to be trained using label information to improve the quality of generated reports. Zhang et al. [10] built a disease relationship mapping matrix based on medical expertise to reduce misdiagnosis and missed diagnosis of diseases by improving the ability to predict labels. Wang et al. [31] designed a graph convolutional network (GCN) to encode prior medical knowledge and semantic features to optimize the underlying logic of the model. Liu et al. [15] optimized the training data according to the training difficulty, thereby alleviating the problem of data imbalance.

Recently, as Transformers have made significant breakthroughs in many fields, researchers have begun to use transformers to build encoder-decoder frameworks to improve the model's ability to learn long sequences of task dependencies. Among them, Song et al. [11] designed a contrastive learning module to compare abnormal pictures with normal ones, so as to identify abnormal areas. Chen et al. [12] designed a cross-modal cache network to establish the mapping relationship between different features and alleviate the negative impact caused

by feature differences. Chen et al. [13] designed a memory matrix and integrated it into the Transformer model to help the model understand the dependency between texts. You et al. [14] proposed an attention module with hierarchical alignment function to encode multi-modal features, and used multi-granularity features to learn the mapping relationship between visual features and label features to generate radiology reports. However, this method is not optimized for the differences between semantic and visual features, and the generated report content is not comprehensive enough. Wang et al. [32] used the Transformer architecture for both encoding and decoding calculations, allowing the model to establish long dependencies on both image and semantic features. Liu et al. [16] used the anomalous visual features provided by posterior knowledge and the knowledge graph provided by prior knowledge to generate reports to mitigate the negative effects of data bias.

Compared with these studies, our model can both reduce misdiagnosis and misdiagnosis of disease, generate text that conforms to the radiology reporting structure, and reduce the loss of key features without the use of label information to avoid new error perturbations. In addition, according to the composition structure of grid features, we encode them with multi-dimensional mapping attention, which alleviates the loss of primary features. Finally, additional feature fusion is applied to the initial semantic features, which alleviates the confusion of query vector structure caused by feature fusion, promotes the interaction between the two features, and achieves fine-grained feature alignment.

## 3. Method

The essence of automatic generation of radiology reports is the task of generating images from texts. We use the same method as Chen et al. [13] to formalize images and texts into two sequences, which are denoted by $X$ and $Y$ respectively. $X$ and $Y$ can be expressed as follows:

$$X = \left\{ x_1, x_2, ..., x_A \right\}, x_a \in \mathbf{R}^d \tag{1}$$

$$Y = \left\{ y_1, y_2, ..., y_B \right\}, y_b \in S \tag{2}$$

where $x_a$ is the $a$ th grid feature extracted from the visual extractor, $d$ is the size of the mapped feature vector, $y_b$ represents the vector space corresponding to the $b$ th word in the generated sequence, and $S$ represents the set of all words in the generated report. **Fig. 2** represents the overall structure of the model proposed in this study, and the details are given in Subsection 3.1.

### 3.1 Overview

As shown in **Fig. 2**, our model consists of three parts: visual extractor, encoder, and decoder, where MDMA module exists in the encoder part, DEA module and FP module exist in the decoder part.
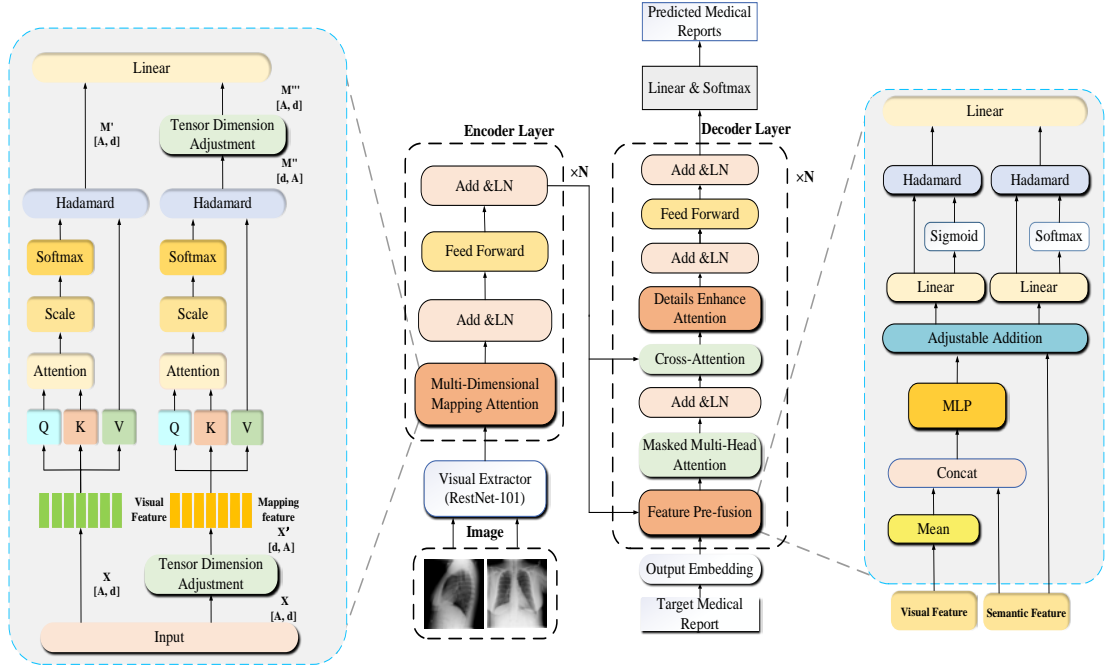
**Fig. 2.** Overall structure of the model

**Visual Extractor:** We use the pre-trained ResNet-101 [34] model to extract visual features. It can extract serialized visual features from radiology images, and the process can be expressed as:

$$\mathrm{F_{ve}}(\boldsymbol{I}) = \boldsymbol{X} \tag{3}$$

where $\mathrm{F_{ve}}$ represents the extraction operation of the visual extractor, and $\boldsymbol{I}$ represents the input radiology images.

**Encoder:** The encoder in this study differs from the vanilla Transformer encoder in that we use the MDMA module to replace the multi-head attention. The encoder module can be mathematically expressed as follows:

$$\boldsymbol{Z} = \mathrm{F_e}(\mathrm{MDMA}(\boldsymbol{X})) \tag{4}$$

where $\mathrm{F_e}$ represents the encoder, $\boldsymbol{Z}$ represents the final output of the encoder, and MDMA represents the operation of the MDMA module. Details about the MDMA module are described in Section 3.2.

**Decoder:** In the design of the decoder, we add the DEA module and FP module to the decoder of the vanilla Transformer. The decoder generates the next word according to the output $\boldsymbol{Z}$ of the encoder and the sequence of words that have been predicted. The specific calculation process can be expressed as:

$$\boldsymbol{y_t} = \mathrm{F_d}(\mathrm{DEA}(\mathrm{CA}(\boldsymbol{Z}, \mathrm{FP}(\boldsymbol{Z}, (\boldsymbol{y_1}, \boldsymbol{y_2}, ..., \boldsymbol{y_{t-1}}))))) \tag{5}$$

where $\boldsymbol{y_t}$ represents the sequence of words predicted in the time step $\boldsymbol{t}$, $\mathrm{F_d}$ represents the decoder, and CA represents the operation of the cross-attention module. Details about the FP module and DEA module are described in Sections 3.3 and 3.4.

**Loss function:** We use the same method as Chen et al. [13] to train our model, using cross-entropy loss function, which is specifically expressed as:

$$L = -\sum_{t=1}^{T} \log(p(\boldsymbol{y_t} \mid \boldsymbol{y_1}, \boldsymbol{y_2}, \ldots, \boldsymbol{y_{t-1}}, \boldsymbol{I}, \boldsymbol{\theta})) \tag{6}$$

where, $\boldsymbol{\theta}$ denotes the parameters of the model.

## 3.2 Multi-dimensional Mapping Attention Module

We use the MDMA module to reduce the loss of primary features during training. The module uses a double-branch structure to encode visual features in two dimensions, so that the model can focus on the primary features within and between grids at the same time, thereby reducing the loss of primary features. In the first branch, MDMA performs multi-head attention calculations directly on input visual features to achieve focus on primary features within the grid, which is expressed as follows:

$$\boldsymbol{M'} = \text{Softmax}(\boldsymbol{W_1}^{\text{T}} \boldsymbol{X} \left(\boldsymbol{W_2}^{\text{T}} \boldsymbol{X}\right)^{\text{T}} / \sqrt{\text{d}_\text{k}})\boldsymbol{W_3}^{\text{T}} \boldsymbol{X} \tag{7}$$

where $\boldsymbol{W_1}^{\text{T}}$, $\boldsymbol{W_2}^{\text{T}}$ and $\boldsymbol{W_3}^{\text{T}}$ are trainable projection matrices, $\text{Softmax}$ represents the Softmax activation function, $\text{d}_\text{k}$ represnts the scaling factor. $\boldsymbol{M'}$ is the output of the first branch. In the second branch, MDMA first makes tensor dimension adjustment on the input visual features, so that the operation of dividing multiple heads directly acts on different grid features, thus realizing multi-head attention calculation between grids. Finally, the result is adjusted again for tensor dimension, and the result is taken as the output of the second branch. The specific calculation can be expressed as follows:

$$\boldsymbol{X'} = \text{Trans}(\boldsymbol{X}) \tag{8}$$

$$\boldsymbol{M''} = \text{Softmax}(\boldsymbol{W_4}^{\text{T}} \boldsymbol{X'} \left(\boldsymbol{W_5}^{\text{T}} \boldsymbol{X'}\right)^{\text{T}} / \sqrt{\text{d}_\text{k}})\boldsymbol{W_6}^{\text{T}} \boldsymbol{X'} \tag{9}$$

$$\boldsymbol{M'''} = \text{Trans}(\boldsymbol{M''}) \tag{10}$$

where $\boldsymbol{W_4}^{\text{T}}$, $\boldsymbol{W_5}^{\text{T}}$ and $\boldsymbol{W_6}^{\text{T}}$ are trainable projection matrices, $\boldsymbol{X'}$ represents the visual feature after adjusting the tensor dimension, $\text{Trans}$ represents the tensor dimension adjustment operation, $\boldsymbol{M''}$ is the hidden state during the computation of the second branch, $\boldsymbol{M'''}$ is the output of the second branch. Finally, the output of the two branches is combined using the full connection layer, and the result is taken as the output of the MDMA module. The specific calculation process can be expressed by (11):

$$\boldsymbol{M} = \boldsymbol{W_7}^{\text{T}}(\text{concat}(\boldsymbol{M'}, \boldsymbol{M'''})) \tag{11}$$

where $\boldsymbol{W_7}^{\text{T}}$ is trainable projection matrices, $\text{concat}$ represents the concatenation operation, $\boldsymbol{M}$ is the output of the MDMA module.

## 3.3 Feature Pre-fusion Module

Due to the different data types represented by visual features and semantic features, there are inevitably some feature differences between them. As shown in the left picture of **Fig. 3**, traditional feature fusion methods directly use the cross-attention module to force alignment between visual features and semantic features, without considering the impact caused by

feature differences. This makes it difficult for the model to learn enough information to generate the long narrative structure consisting of multiple sentences required for radiology reports.
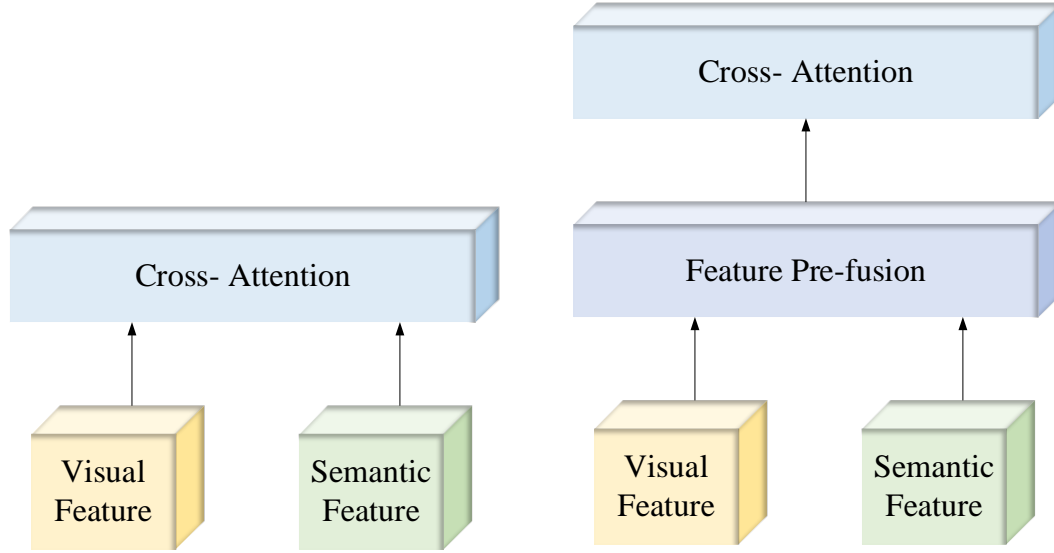


**Fig. 3.** Comparison between the traditional multi-modal fusion method (left) and our proposed multi-modal fusion method (right).

To this end, as shown in the right figure of **Fig. 3**, we design an FP module that can adjust the overall structure of multi-modal features and add it before the cross-attention calculation, thereby creating a buffer space for multi-modal feature fusion, helping the model understand the mapping relationship between visual features and semantic features, and generating a radiology report with appropriate structure. Specifically, the FP module firstly extracts the global features from the visual features, and uses a multilayer perceptron (MLP) to integrate them into the semantic features to obtain the primary multi-modal features. Then, the adjustable addition module is used to integrate the original semantic features into the primary multi-modal features in proportion. Finally, two activation functions, Sigmoid and Softmax, are used for attention calculation at the same time, and the results are output after a fully connected layer reshapes the dimension. The specific calculation formula of the FP module can be expressed as follows:

$$SS = W_8^{\mathrm{T}}(y_1, y_2, \ldots, y_{t-1}) \tag{12}$$

$$\mathrm{Add}(j, k) = j + \mu k, 0 < \mu < 1 \tag{13}$$

$$H_1 = \mathrm{Add}(W_9^{\mathrm{T}}\mathrm{concat}(SS, \mathrm{Mean}(Z))W_{10}^{\mathrm{T}}, SS) \tag{14}$$

$$H_2 = \sigma(W_{11}^{\mathrm{T}}H_1) \odot W_{11}^{\mathrm{T}}H_1 \tag{15}$$

$$H_3 = \mathrm{Softmax}(W_{11}^{\mathrm{T}}H_1) \odot W_{11}^{\mathrm{T}}H_1 \tag{16}$$

$$H_4 = W_{12}^{\mathrm{T}}(\mathrm{concat}(H_2, H_3)) \tag{17}$$

where $SS$ represents the semantic feature, $\mathrm{Add}$ represents the adjustable add operation, $\mu$ represents an adjustable weighting factor, $j$ and $k$ represent the parameters for which the Add operation is required, $\mathrm{Mean}$ represents the averaging operation, $H_1$, $H_2$, $H_3$ represent the hidden state during the execution of the FP module, $W_8^{\mathrm{T}}$, $W_9^{\mathrm{T}}$, $W_{10}^{\mathrm{T}}$, $W_{11}^{\mathrm{T}}$ and $W_{12}^{\mathrm{T}}$ are trainable projection matrices, $\sigma$ represents the Sigmoid activation function, $\odot$ represents the Hadamard product, $H_4$ represents the output of the FP module. Thereafter, $H_4$ is fed to the cross-attention module to generate the current hidden state $H$.

### 3.4 Detail-Enhanced Attention Module

Most of the past studies use the computational results of the cross-attention module to directly predict radiology reports, but it has certain limitations in the radiology report generation task with extremely imbalanced data, and the visual features are difficult to meet the query requirements of key features. To this end, we propose to use the DEA module to strengthen the extracted key features and improve the feature utilization ability of the model. The specific operation process of the DEA module can be expressed by the formula as follows:

$$H' = \mathrm{Mean}(H) \tag{18}$$

$$H'' = \mathrm{Max}(H) \tag{19}$$

$$C_1 = W_{13}^{\mathrm{T}}\mathrm{Conv}(\mathrm{ReLU}(\mathrm{Conv}(H'))) \tag{20}$$

$$C_2 = W_{14}^{T}\mathrm{Conv}(\mathrm{ReLU}(\mathrm{Conv}(H''))) \tag{21}$$

$$C = H \odot (\tanh(C_1 + C_2) \odot \sigma(C_1 + C_2)) + H \tag{22}$$

where $H'$ and $H''$ represent the extracted key features, $\mathrm{Max}$ represents the maximum calculation operation, $C_1$ and $C_2$ represent the context features in the operation process, $W_{13}^{\mathrm{T}}$ and $W_{14}^{T}$ represent the trainable projection matrix, $\mathrm{Conv}$ represents the convolution operation, and $C$ represents the final output context features.

## 4. Experimental Setup and Result Analysis

### 4.1 Dataset and Evaluation Metrics

We adopt the public datasets IU X-Ray [35] and MIMIC-CXR [36], which are widely used in current studies. The IU X-Ray dataset is released by Indiana University in 2015 and includes 7.470 images and 3,995 reports. We use the same data partition as Chen et al. [13] to divide the training, validation, and test sets by the ratio of 7:1:2. The MIMIC-CXR dataset is released by the Beth Israel Deaconess Medical Center, is the largest of its kind, with 473,057 images and 236,563 reports from 63,478 patients. We use the officially provided data partition to increase the credibility of the experiment. In addition, we adopt the same preprocessing to the reports as Chen et al. [13], removing unreported images and converting all letters to lowercase. The top five frequently occurring words in the reports of the two datasets are shown in **Fig. 4**,

and the analysis shows that the abstract words corresponding to no image region have a high frequency.
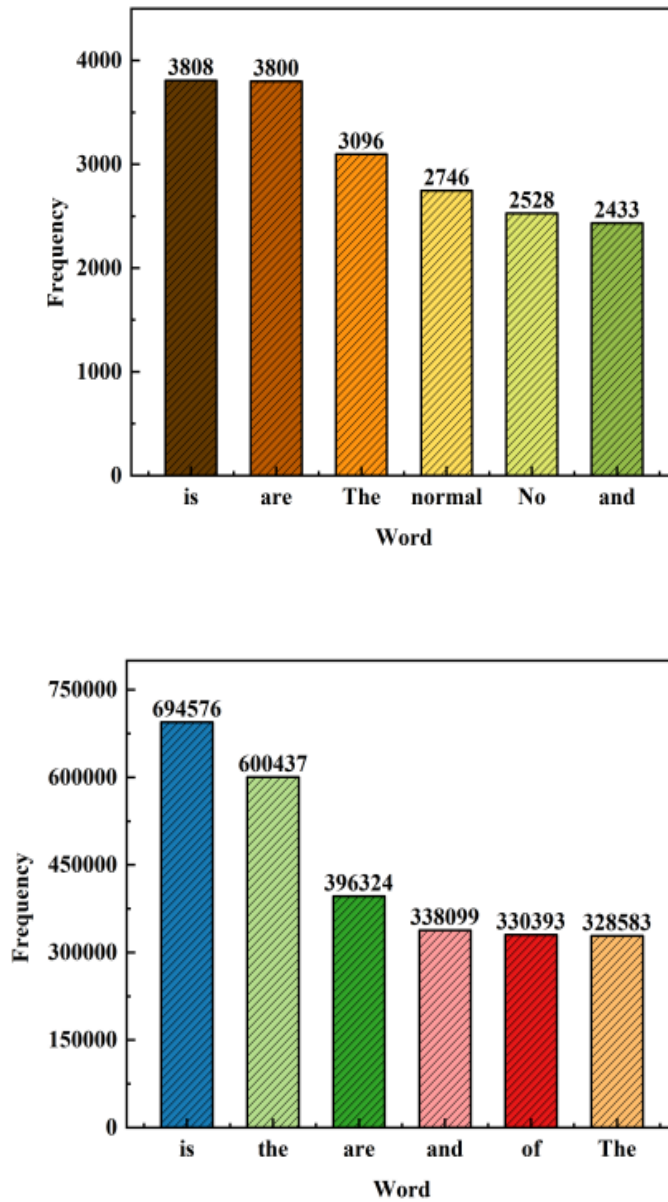
**Fig. 4.** The above figure shows the top five words in the IU X-Ray dataset, and the below figure shows the top five words in the MIMIC-CXR dataset

We evaluate our model using natural language processing (NLP) evaluation metrics BLEU [37], METEOR [38], and ROUGE-L [39]. B1, B2, B3 and B4 are BLEU metrics when N-grams are 1, 2, 3, and 4, respectively, and ROUGE stands for ROUGE-L metric.

## 4.2 Implementation Details

We use a ResNet-101 model that has been pre-trained on ImageNet 1K[1] to extract grid features from images with the dimension of each feature set to 2,048. The number of heads for multi-head attention is set to 8. It should be noted that the IU X-Ray dataset contains frontal and lateral images of one patient, which we combined as input to the visual extractor. In addition, we set the rate decay of each epoch to 0.8, the beam size to 3, the learning rates of the visual extractor and other parameters to 1e-4, 5e-5, and the batch size to 32.

## 4.3 Comparative experiment

As shown in **Table 1**, we used six NLP evaluation indicators to compare the model proposed in this study with the current mainstream models [11-13, 15-19, 40-41]. Among them, AdaAtt [40] could automatically select the features that need to be relied on, M2Transformer [41] used a new connection method to connect the decoder and the encoder, CMCL [15] established a data alignment mechanism based on long short term memory (LSTM), R2Gen [13] designed a memory module to store contextual information and establish a contextual dependency relationship, PPKED [16] detected and assigned disease labels according to the doctor's working style, CA [11] could learn the difference between normal and abnormal samples, GSKET [17] could learn both general and special medical knowledge.

**Table 1.** Comparison of NLG index effect between the proposed model and the existing models

| Dataset | Model | B1 | B2 | B3 | B4 | METEOR | ROUGE |
|---|---|---|---|---|---|---|---|
| IU X-Ray | AdaAtt | 43.6 | 28.8 | 20.3 | 15.0 | - | 35.4 |
| | M2Transformer | 46.3 | 31.8 | 21.4 | 15.5 | - | 33.5 |
| | CMCL | 47.3 | 30.5 | 21.7 | 16.2 | 18.6 | 37.8 |
| | R2Gen | 47.0 | 30.4 | 21.9 | 16.5 | 18.7 | 37.1 |
| | CMN | 47.5 | 30.9 | 22.2 | 17.0 | 19.1 | 37.5 |
| | PPKED | 48.3 | 31.5 | 22.4 | 16.8 | - | 37.6 |
| | CA | 49.2 | 31.4 | 22.2 | 16.9 | 19.3 | 38.1 |
| | GSKET | 49.6 | 32.7 | 23.8 | 17.8 | - | 38.1 |
| | Ours | **51.7** | **35.6** | **25.9** | **19.1** | **21.4** | **39.0** |
| MIMIC-CXR | AdaAtt | 29.9 | 18.5 | 12.4 | 8.8 | 11.8 | 26.6 |
| | M2Transformer | 21.2 | 12.8 | 8.3 | 5.8 | - | 24.0 |
| | CMCL | 34.4 | 21.7 | 14.0 | 9.7 | 13.3 | 28.1 |
| | R2Gen | 35.3 | 21.8 | 14.5 | 10.3 | 14.2 | 27.7 |
| | CMN | 35.3 | 21.8 | 14.8 | 10.6 | 14.2 | 27.8 |
| | PPKED | 36.0 | 22.4 | 14.9 | 10.6 | 14.9 | 28.4 |
| | CA | 35.0 | 21.9 | 15.2 | 10.9 | 15.1 | 28.7 |
| | GSKET | 36.3 | 22.8 | 15.6 | 11.5 | - | 28.4 |
| | Ours | **38.5** | **23.7** | **16.5** | **12.1** | **15.6** | **28.6** |

---

[1] https://www.imagenet.org/challenges/LSVRC/index.php

Further analysis shows that our model has a significant performance improvement compared with the image captioning models AdaAtt [40] and M2Transformer [41], which indicates that the image captioning model cannot meet the needs of the radiology report generation task. In addition, by comparing with the specialized models for radiology reports [11-13, 15-17], the scores of our model on the two datasets are significantly better than those of the current state-of-the-art models. Specifically, BLEU-1 improves from 49.6 to 51.7 on IU X-Ray dataset and from 36.3 to 38.5 on MIMIC-CXR dataset.

## 4.4 Ablation experiment

As shown in **Table 2**, we conducted complete ablation experiments on two datasets for the three newly proposed modules in the model. Base represents the base model without adding our proposed module, MDMA, DEA, and FP represent the multi-dimensional mapping attention module, detail-enhanced attention module, and pre-fusion module respectively, Base+MDMA+DEA+FP represents the final model of this study. Compared to the base model, each module of our model improves. Among them, the MDMA module has an average increase of 4.1% compared with the base model in six evaluation indicators on the IU X-Ray dataset and 5.1% on the MIMIC-CXR dataset. The DEA module improves by 6.0% on IU X-Ray dataset and 8.0% on MIMIC-CXR dataset. The FP module achieves 2.2% improvement on IU X-Ray dataset and 2.7% improvement on MIMIC-CXR dataset. The experimental results prove that the three modules we proposed have strong robustness and achieve good results without using label data.

**Table 2.** Effect comparison between the basic model and the model in this study

| Dataset | Model | B1 | B2 | B3 | B4 | METEOR | ROUGE |
|---------|-------|-----|-----|-----|-----|--------|-------|
| IU X-Ray | Base | 44.3 | 28.4 | 20.6 | 15.9 | 18.0 | 34.7 |
| | Base+MDMA | 45.2 | 29.2 | 21.7 | 16.9 | 18.8 | 36.0 |
| | Base+DEA | 46.6 | 29.8 | 21.4 | 16.1 | 19.7 | 38.6 |
| | Base+FP | 44.6 | 28.8 | 21.0 | 16.2 | 18.5 | 36.3 |
| | Base+MDMA+FP | 45.5 | 29.9 | 22.3 | 17.6 | 19.0 | 37.2 |
| | Base+MDMA+DEA | 50.1 | 32.5 | 24.5 | 18.2 | 20.1 | **39.1** |
| | Base+DEA+FP | 49.3 | 32.0 | 23.2 | 17.6 | 20.4 | 38.8 |
| | Base+MDMA+DEA+FP | **51.7** | **35.6** | **25.9** | **19.1** | **21.4** | 39.0 |
| MIMIC-CXR | Base | 32.9 | 20.1 | 13.4 | 9.7 | 13.1 | 26.9 |
| | Base+MDMA | 34.6 | 20.9 | 14.2 | 10.6 | 13.6 | 27.5 |
| | Base+DEA | 35.6 | 21.6 | 14.9 | 10.7 | 14.3 | 27.3 |
| | Base+FP | 33.5 | 20.4 | 13.9 | 10.2 | 13.5 | 27.2 |
| | Base+MDMA+FP | 35.4 | 21.6 | 14.7 | 11.2 | 14.6 | 28.5 |
| | Base+MDMA+DEA | 37.5 | 23.0 | 15.5 | 11.4 | 14.9 | **27.6** |
| | Base+DEA+FP | 37.0 | 22.8 | 15.2 | 10.9 | 14.4 | 27.5 |
| | Base+MDMA+DEA+FP | **38.5** | **23.7** | **16.5** | **12.1** | **15.6** | **28.6** |

We use the Beam Search algorithm to generate the final report. The number of words retained in each time step of the algorithm is denoted as the hyperparameter Beam, and the size of Beam will directly affect the final performance of the model. As shown in **Table 3**,

when Beam is set to 3, our model achieves the best performance in both datasets. In addition, in IU X-Ray dataset, when the Beam is 5, ROUGE-L index score is the highest.

Table 3. Hyperparameter ablation study based on IU X-Ray dataset and MIMIC-CXR dataset

| Dataset | Beam | B1 | B2 | B3 | B4 | METEOR | ROUGE |
|---------|------|------|------|------|------|--------|-------|
| IU X-Ray | 1 | 46.8 | 30.4 | 21.1 | 15.6 | 19.8 | 37.2 |
| | 2 | 49.4 | 31.6 | 22.4 | 16.3 | 20.7 | 37.5 |
| | 3 | **51.7** | **35.6** | **25.9** | **19.1** | **21.4** | 39.0 |
| | 4 | 50.4 | 32.7 | 22.8 | 18.8 | 20.5 | 38.1 |
| | 5 | 51.3 | 34.3 | 25.3 | 19.0 | 20.2 | **39.3** |
| MIMIC-CXR | 1 | 34.2 | 20.8 | 12.5 | 9.5 | 12.5 | 24.6 |
| | 2 | 36.6 | 22.2 | 15.0 | 10.9 | 14.2 | 26.8 |
| | 3 | **38.5** | **23.7** | **16.5** | **12.1** | **15.6** | **28.6** |
| | 4 | 37.8 | 22.9 | 15.8 | 11.2 | 14.3 | 27.3 |
| | 5 | 37.2 | 22.4 | 15.6 | 10.8 | 14.0 | 27.0 |

In the FP module, we introduce the adjustable addition module, which contains a manually adjusted weighting factor to adjust the proportion of introduced original semantic features. As shown in **Table 4**, we evaluated the performance of the model with different $\mu$ selected. When $\mu$ is 0.6, the comprehensive index achieves the best effect.

Table 4. Ablation study of the weighting factor $\mu$ in the IU X-Ray dataset and MIMIC dataset

| Dataset | $\mu$ | B1 | B2 | B3 | B4 | METEOR | ROUGE |
|---------|------|------|------|------|------|--------|-------|
| IU X-Ray | 0.0 | 47.3 | 31.0 | 21.9 | 16.3 | 20.1 | 37.3 |
| | 0.2 | 47.5 | 31.1 | 22.1 | 16.6 | 20.5 | 37.2 |
| | 0.4 | 49.2 | 33.6 | 23.6 | 18.7 | 20.8 | 38.2 |
| | 0.6 | **51.7** | **35.6** | **25.9** | **19.1** | **21.4** | **39.0** |
| | 0.8 | 50.5 | 34.2 | 25.0 | 18.2 | 21.1 | 38.3 |
| MIMIC-CXR | 0.0 | 36.3 | 22.9 | 14.7 | 10.9 | 14.0 | 26.6 |
| | 0.2 | 36.6 | 23.1 | 15.2 | 11.2 | 14.3 | 27.5 |
| | 0.4 | 37.5 | 23.0 | 15.7 | 11.6 | 14.8 | 27.9 |
| | 0.6 | **38.5** | **23.7** | **16.5** | **12.1** | **15.6** | **28.6** |
| | 0.8 | 38.0 | 23.4 | 15.9 | 11.8 | 15.2 | 28.2 |

## 4.5 Complexity analysis

As shown in **Table 5**, we conducted a complexity analysis of this model based on the IU X-Ray dataset, and compared with the R2Gen [13] model, our model achieved better results while using fewer parameters.

**Table 5.** The Results of Complexity Analysis

| Dataset | Model | Parameters | B1 | B2 | B3 | B4 | METEOR | ROUGE |
|---------|-------|-----------|------|------|------|------|--------|-------|
| IU X-Ray | Base | 56.95M | 44.3 | 28.4 | 20.6 | 15.9 | 18.0 | 34.7 |
| | R2Gen | 78.47M | 47.0 | 30.4 | 21.9 | 16.5 | 18.7 | 37.1 |
| | Ours | 77.21M | **51.7** | **35.6** | **25.9** | **19.1** | **21.4** | **39.0** |

## 4.6 Generalization analysis

We verify the generalization of the model with the help of the current SOAT model: S2-Transformer [42]. We integrated the proposed three modules into the S2-Transformer model, trained using cross-entropy loss on two datasets, and the final results are shown in **Table 6**, Our model has a certain optimization effect on S2-Transformer, improving 8.2% and 10.1% on IU X-Ray dataset and MIMIC-CXR dataset, respectively, which shows that our model has good generalization.

**Table 6.** The Results of Generalization Analysis

| Dataset | Model | B1 | B2 | B3 | B4 | METEOR | ROUGE |
|---------|-------|------|------|------|------|--------|-------|
| IU X-Ray | $S^2$-Transformer | 43.5 | 28.0 | 20.7 | 15.8 | 17.7 | 35.6 |
| | $S^2$-Transformer+Ours | **46.7** | **31.1** | **22.7** | **17.2** | **18.9** | **37.5** |
| MIMIC-CXR | $S^2$-Transformer | 33.0 | 19.8 | 12.7 | 9.2 | 12.2 | 27.1 |
| | $S^2$-Transformer+Ours | **35.2** | **21.9** | **14.5** | **10.4** | **13.7** | **28.2** |

## 4.7 Analysis of results

In order to more intuitively show the optimization effect of our proposed model and its three new modules, we used five models including base model to test and analyze a group of medical cases in IU X-Ray dataset containing images of front and side chest, and the results are shown in **Fig. 5**. Among them, Ground-truth is the report manually written by doctors, Base is the report generated by the base model, Base+MDMA is the report generated by adding the MDMA module to the base model, Base+FP is the report generated by adding the FP module to the base model, Base+DEA is the report generated by adding the DEA module to the base model, Ours is the report generated by the complete MFOT model, words marked in yellow are primary words and words marked in green are key words.

**Fig. 5.** Comparison of reports generated from Ground-truth, Base, Base+MDMA, Base+FP, Base+DEA, and Base+MDMA+DEA+FP model.

Further analysis reveals that the base model did not generate appropriate language structures, and the second sentence was completely contrary to the human-written report. The Base+MDMA model generated more primary words than the base model and successfully recognized the aortic morphology, which shows that the MDMA module has a good optimization effect on the primary features. The Base+FP model generated more long narrative sentences than the base model, which is more in line with the needs of radiology reports and alleviates the feature gap directly between multi-modal features. The Base+DEA model generated more key words than the base model, and the description of the normal area was basically consistent with the manual report, so as to successfully identify "lung granuloma", but the primary features were not strengthened, so the appropriate primary words were not generated. The MFOT model not only generated a report with appropriate structure, but also had the expected number of primary words and key words, and accurately identified the disease and the location of the disease.
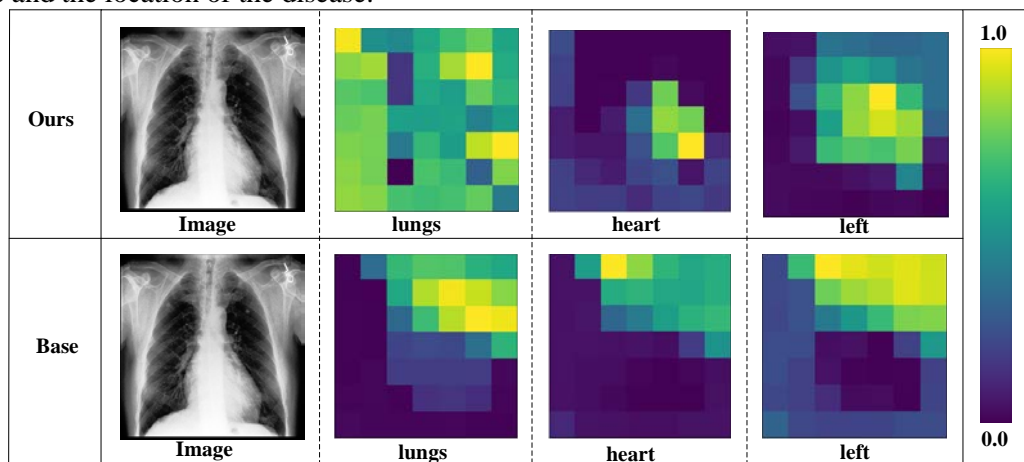


**Fig. 6.** The visual image of the image text attention mapping generated by Base and Base+SA+DEA+FP model, with colors from blue to green representing weights from low to high.

As shown in **Fig. 6**, we performed visual operations on the cross-attention module in the base model and the multi-feature optimized Transformer based on the MIMIC-CXR dataset. Through the visualization of words such as "heart", it can be found that our model can more accurately establish the dependency relationship between the image and the generated word, and successfully accomplish the feature alignment task without relying on label features.
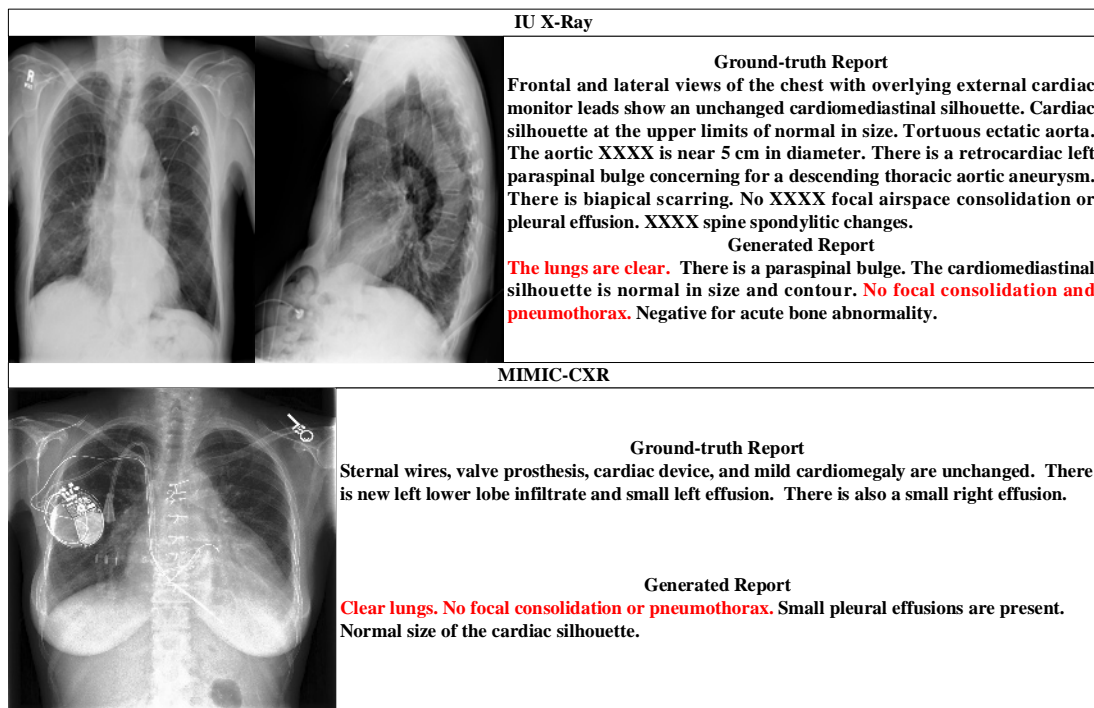


**Fig. 7.** Failure cases of MFOT model on IU X-Ray dataset and MIMIC-CXR dataset. Sentences marked in red indicate high-frequency sentences in radiology reports.

As shown in **Fig. 7**, we present the failure cases generated by the MFOT model. Analysis shows that the MFOT model successfully generates some key words and primary words in both cases, such as "pleural effusion", "paraspinal bulge", etc., but there are still some key words and primary words missed. We believe that this is because multiple key words and primary words correspond to the same image region, and the ability of the proposed model to align multiple semantic information with the same image region cannot meet the actual clinical needs, resulting in the quality of generated reports. In addition, a variety of diseases have certain associations with each other, such as "paraspinal bulge" and "thoracic aortic aneurysm", and our model does not learn the relationship between diseases like clinical practice, resulting in insufficiently comprehensive reports generated. In summary, this task faces two problems. First, the model cannot generate appropriate reports when multiple diseases correspond to the same regional image, and second, the model also needs to learn the relationship between diseases, both of which are very valuable research directions.

## 5. Conclusion

This paper proposes to use the MFOT to generate radiology reports, which is the first deep learning model that is specially optimized for the generation of medical professional words without using label data. Specifically, we propose three new modules: MDMA, DEA, and FP. Among them, the MDMA module and DEA module can respectively reduce the loss of primary features and key features in the process of model calculation, and the FP module can improve the interaction ability between multi-modal features. The three modules cooperate with each other, so that the problem of misreporting and underreporting of diseases can be preliminatively solved, and the report structure generated basically meets the requirement of long narrative composed of multiple sentences. However, the model proposed in this study still has some limitations. First of all, it is difficult for this model to cope with the special cases where multiple key words and primary words correspond to the same visual region. Therefore, a memory module will be designed for the model in the future to strengthen the model's understanding of the special region. Secondly, this model does not learn the relationship between diseases, which makes it difficult for the model to establish a logical relationship between diseases. In the future, we will try to establish a knowledge map of disease relationships for the model to strengthen the model's learning of basic medical knowledge.

## Acknowledgement

## References

[1]   X. Chen, H. Fang, T.Lin, R. Vedantam, S. Gupta, P. Dollar, and C. L. Zitnick, "Microsoft coco captions: Data collection and evaluation server," *arXiv preprint arXiv:1504.00325*, 2015. Article (CrossRef Link)

[2]   P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6077-6086, 2018. Article (CrossRef Link)

[3]   O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3156-3164, 2015. Article (CrossRef Link)

[4]   K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. of the 32nd International Conference on Machine Learning (PMLR)*, vol. 37, pp. 2048-2057, 2015. Article (CrossRef Link)

[5]   L. Zhou, S. Pan, J. Wang, and A. V. Vasilakos, "Machine learning on big data: Opportunities and challenges," *Neurocomputing*, vol. 237, pp. 350-361, 2017. Article (CrossRef Link)

[6]   S. Siddique and J. C. Chow, "Machine learning in healthcare communication," *Encyclopedia*, vol. 1, no.1, pp. 220-239, 2021. Article (CrossRef Link)

[7]   K. J. W. Tang, C. K. E. Ang, T. Constantinides, V. Rajinikanth, U. R. Acharya, and K. H. Cheong, "Artificial intelligence and machine learning in emergency medicine," *Biocybernetics and Biomedical Engineering*, vol. 41, no. 1, pp. 156-172, 2021. Article (CrossRef Link)

[8]   G. Liu, T. M. H. Hsu, M. McDermott, W. Boag, W. H. Weng, P. Szolovits, and M. Ghassemi, "Clinically accurate chest x-ray report generation," in *Proc. of the 4th Machine Learning for Healthcare Conference (PMLR)*, vol. 106, pp. 249-269, 2019. Article (CrossRef Link)

[9]   B. Jing, P. Xie, and E. Xing, "On the Automatic Generation of Medical Imaging Reports," in *Proc. of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp.2577-2586, 2018. Article (CrossRef Link)

[10] Y. Zhang, X. Wang, Z. Xu, Q. Yu, A. Yuille, and D. Xu, "When radiology report generation meets knowledge graph," in *Proc. of the AAAI Conference on Artificial Intelligence (AAAI)*, pp. 12910-12917, 2020. Article (CrossRef Link)

[11] X. Song, X. Zhang, J. Ji, Y. Liu, and P. Wei, "Cross-modal Contrastive Attention Model for Medical Report Generation," in *Proc. of the 29th International Conference on Computational Linguistics (COLING)*, pp. 2388-2397, 2022. Article (CrossRef Link)

[12] Z. Chen, Y. Shen, Y. Song, and X. Wan, "Cross-modal memory networks for radiology report generation," in *Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL/IJCNLP)*, pp. 5904-5914, 2021. Article (CrossRef Link)

[13] Z. Chen, Y. Song, T. H. Chang, and X. Wan, "Generating radiology reports via memory-driven transformer," in *Proc. of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1439-1449, 2020. Article (CrossRef Link)

[14] D. You, F. Liu, S. Ge, X. Xie, J. Zhang, and X. Wu, "Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation," in *Proc. of the 24th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 72-82, 2021. Article (CrossRef Link)

[15] F. Liu, S. Ge, and X. Wu, "Competence-based Multimodal Curriculum Learning for Medical Report Generation," in *Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL/IJCNLP)*, pp. 3001-3012, 2021. Article (CrossRef Link)

[16] F. Liu, X. Wu, S. Ge, W. Fan, and Y. Zou, "Exploring and Distilling Posterior and Prior Knowledge for Radiology Report Generation," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13753-13762, 2021. Article (CrossRef Link)

[17] S. Yang, X. Wu, S. Ge, S. K. Zhou, and L. Xiao, "Knowledge matters: Chest radiology report generation with general and specific knowledge," *Medical Image Analysis*, vol. 80, p.102510, 2022. Article (CrossRef Link)

[18] K. J. Chae, J. Choi, G. Y. Jin, E. A. Hoffman, A. T. Laroia, M. Park, and C. H, Lee, "Relative regional air volume change maps at the Acinar scale reflect variable ventilation in low lung attenuation of COPD patients," *Academic Radiology*, vol. 27, no. 11, pp. 1540-1548, 2020. Article (CrossRef Link)

[19] E. Kim, Y. Lee, J. Choi, B. Yoo, K. J. Chae and C. H. Lee, "Machine Learning-based Prediction of Relative Regional Air Volume Change from Healthy Human Lung CTs," *KSII Transactions on Internet and Information Systems*, vol. 17, no. 2, pp. 576-590, 2023. Article(CrossRef Link)

[20] C. Liu, S. C. Hu, C. Wang, K. Lafata, and F. F. Yin, "Automatic detection of pulmonary nodules on CT images with YOLOv3: development and evaluation using simulated and patient data," *Quantitative Imaging in Medicine and Surgery*, vol. 10, no. 10, pp. 1917-1929, 2020. Article (CrossRef Link)

[21] S. Mei, H. Jiang, and L. Ma, "YOLO-lung: A Practical Detector Based on Improved YOLOv4 for Pulmonary Nodule Detection," in *Proc. of 2021 14th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics*, pp. 1-6, Oct. 2021. Article (CrossRef Link)

[22] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L.Lu, A. L. Yuille, and Y. Zhou, "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation," *arXiv preprint arXiv:2102.04306*, 2021. Article (CrossRef Link)

[23] Y. Zhang, H. Liu, and Q. Hu. "Transfuse: Fusing transformers and cnns for medical image segmentation," in *Proc. of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 14-24, 2021. Article (CrossRef Link)

[24] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," in *Proc. of the European conference on computer vision (ECCV)*, pp. 205-218, 2022. Article (CrossRef Link)

[25] F. Liu, Y. Liu, X. Ren, X. He, and X. Sun, "Aligning visual regions and textual concepts for semantic-grounded image representations," in *Proc. of the 32nd International Conference on Neural Information Processing Systems (NIPS)*, pp. 6847-6857, 2019. Article (CrossRef Link)

[26] F. Liu, X. Ren, Y. Liu, K. Lei, and X. Sun, "Exploring and Distilling Cross-Modal Information for Image Captioning," in *Proc. of the International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 5095-5101, 2019. Article (CrossRef Link)

[27] F. Liu, X. Ren, Y. Liu, H. Wang, and X. Sun, "simNet: Stepwise Image-Topic Merging Network for Generating Detailed and Comprehensive Image Captions," in *Proc. of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 137-149, 2018.
Article (CrossRef Link)

[28] L. Huang, W. Wang, J. Chen, and X. Wei, "Attention on attention for image captioning," in *Proc. of the IEEE/CVF international conference on computer vision* (*ICCV)*, pp. 4634-4643, 2019.
Article (CrossRef Link)

[29] Y. Wang, J. Xu, and Y. Sun, "Based Model fo End-to-End Transformer r Image Captioning," in *Proc. of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 36(3), pp. 2585-2594, 2022.
Article (CrossRef Link)

[30] Y. Pan, T. Yao, Y. Li, and T. Mei, "X-linear attention networks for image captioning," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10971-10980, 2020. Article (CrossRef Link)

[31] S. Wang, L. Tang, M. Lin, G. Shih, Y. Ding, and Y. Peng, "Prior Knowledge Enhances Radiology Report Generation," in *Proc. of AMIA Annual Symposium Proceedings,* pp. 486-495, 2022.
Article (CrossRef Link)

[32] Z. Wang, H. Han, L. Wang, X. Li, L. Zhou, "Automated Radiographic Report Generation Purely on Transformer: A Multicriteria Supervised Approach," *IEEE Transactions on Medical Imaging*, vol. 41, no. 10, pp.2803-2813, 2022.  Article (CrossRef Link)

[33] Z. Wang, M. Tang, L. Wang, X. Li, and L. Zhou, "A medical semantic-assisted transformer for radiographic report generation," in *Proc. of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI),* pp. 655-664, 2022.
Article (CrossRef Link)

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.770-778, 2016.
Article (CrossRef Link)

[35] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. Antani, G. R. Thoma, and C. J. McDonald, "Preparing a collection of radiology examinations for distribution and retrieval," *Journal of the American Medical Informatics Association*, vol. 23, no. 2, pp. 304-310, 2016. Article (CrossRef Link)

[36] A. E. Johnson, T. J. Pollard, N. R. Greenbaum, M. P. Lungren, C. Y. Deng, Y. Peng, and S. Horng, "MIMIC-CXR: A large publicly available database of labeled chest radiographs," *arXiv preprint arXiv:1901.07042*, 2019.  Article (CrossRef Link)

[37] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proc. of the 40th annual meeting of the Association for Computational Linguistics (ACL),* pp. 311-318, 2002. Article (CrossRef Link)

[38] M. Denkowski and A. Lavie, "Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems," in *Proc. of the sixth workshop on statistical machine translation*, pp. 85-91, 2011. Article (CrossRef Link)

[39] C. Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Proc. of the workshop on text summarization branches out*, pp. 74-81, 2004. Article (CrossRef Link)

[40] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3242-3250, 2017. Article (CrossRef Link)

[41] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-Memory Transformer for Image Captioning," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.10575-10584, 2020. Article (CrossRef Link)

[42] P. Zeng, H. Zhang, J. Song, and L. Gao, "S2 Transformer for Image Captioning," in *Proc. of the International Joint Conference on Artificial Intelligence (IJCAI),* pp. 1608-1614, 2022. [Article (CrossRef Link)](#)

**RUI WANG** received his bachelor's degree in computer science and technology from the School of Information Science and Engineering, University of Jinan. He is now a master's candidate in Shandong University of Science and Technology, and his main research fields are artificial intelligence, high performance computing, etc.

**RONG HUA** received his Ph.D. degree in computer Software and Theory from Shandong University of Science and Technology. He is currently an associate professor at Shandong University of Science and Technology, and his main research fields are artificial intelligence, high performance computing, etc.