

Research Article

Medical Image Description Based on Multimodal Auxiliary Signals and Transformer

Yun Tan , Chunzhi Li , Jiaohua Qin , Youyuan Xue, and Xuyu Xiang

Central South University of Forestry and Technology, Changsha, China

Correspondence should be addressed to Yun Tan; tantanyun@hotmail.com

Received 29 June 2023; Revised 6 November 2023; Accepted 20 January 2024; Published 13 February 2024

Academic Editor: Costa Gianni

Copyright © 2024 Yun Tan et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Medical image description can be applied to clinical medical diagnosis, but the field still faces serious challenges. There is a serious problem of visual and textual data bias in medical datasets, which are the imbalanced distribution of health and disease data. This can greatly affect the learning performance of data-driven neural networks and finally lead to errors in the generated medical image descriptions. To address this problem, we propose a new medical image description network architecture named multimodal data-assisted knowledge fusion network (MDAKF), which introduces multimodal auxiliary signals to guide the Transformer network to generate more accurate medical reports. In detail, audio auxiliary signals provide clear abnormal visual regions to alleviate the visual data bias problem. However, the audio modality signals with similar pronunciation lack recognizability, which may lead to incorrect mapping of audio labels to medical image regions. Therefore, we further fuse the audio with text features as the auxiliary signal to improve the overall performance of the model. Through the experiments on two medical image description datasets, IU-X-ray and COV-CTR, it is found that the proposed model is superior to the previous models in terms of language generation evaluation indicators.

1. Introduction

In recent years, with the development of interdisciplinary fusion techniques, automatic generation of image descriptions using deep learning has become a popular research in the field of computer vision [1–3]. Image description is a combination of image recognition and text generation from machine learning. Mainstream tasks include news image description generation [4], machine vision quizzing [5], visual inference, remote sensing image description [6], and automatic generation from text to image. These techniques have a great value in many practical application scenarios, for example, in the medical field where it can play the role of a doctor writing medical reports [7]. In early childhood education, it can play the role of a lecturer [8]. It can also be used to help visually impaired people to perceive the visual content of their surroundings.

Currently, as society progresses, people's desire to pursue a healthy life has led to a dramatic increase in the number of medical images, which has also led to an increased workload

for imaging physicians. However, the excessive pressure and the lack of experienced imaging doctors make patients spend more time waiting for medical reports. To solve this problem, we started to study the automatic generation of radiology reports. Generally, a diagnostic report written by a physician is a textual presentation that gives information about the patient's medical image symptoms, as shown in Figure 1. The generated medical report should be accurate, complete, and readable.

In recent years, with the continuous development of deep learning technology, the technique of medical image description tasks has continued to mature. Nowadays, most of the models adopt the encoder-decoder framework [9]. The encoder consists of a convolutional neural network (CNN) that extracts image features, and the decoder is initially composed of a recurrent neural network (RNN) that generates text descriptions. In order to improve the above network architecture, the researchers applied transformer [10] to the task of medical image description. Compared with existing image description methods, the transformer model

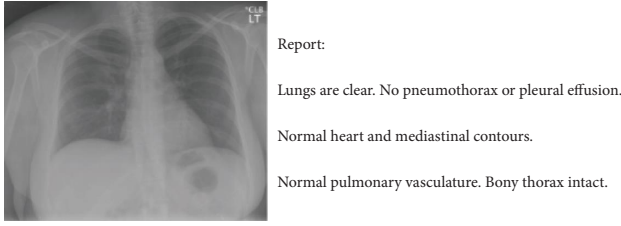


FIGURE 1: Example of a chest X-ray image and its report.

can better interact with multimodal data in the multihead attention module; thus, it can better reason about multimodal data and finally output a more accurate description.

There are two key problems with the current deep learning-based medical image description. On the one hand, despite the emergence of public medical image description datasets, such as IU X-ray and MIMIC-CXR [11], medical image description datasets are still scarce. In addition, these datasets may suffer from data bias problems. In terms of the visual bias problem, the distribution of medical image data is extremely unbalanced, which can lead to the low sensitivity of existing deep learning models to odd disease conditions [12]. In terms of the text bias problem, as shown in Figure 1, the medical report describes all symptoms in the image, so the generated reports are mostly occupied by health descriptions. In addition, as far as some normal regions are concerned, these similar symptom statements are always repeated in the report of the dataset. This unbalanced text distribution will make normal statements too repetitive and make it difficult for the model to generate anomaly descriptions. Some researchers have proposed using the migration learning [13] approach to address the data bias problem. In 2021, Wu et al. [14] explored the role of text labeling auxiliary information on the transformer model to generate medical image descriptions, and the experiments finally showed that multimodal data interaction can improve the accuracy of the model. On the other hand, although the current state-of-the-art radiology report automatic generation models show a great improvement in performance metrics, the generated medical reports still do not meet the market application standards and have the problem of inconsistency with real reports.

To address the above issues, this manuscript started to explore the impact of multimodal auxiliary data on the performance of the transformer model. A multimodal data-assisted knowledge fusion framework (MDAKF) is proposed to mimic the working mode of a radiologist. It will pay more attention to the areas with a high probability of disease like an experienced radiologist diagnosis. Then, it analyzes and views the overall image to label abnormal areas and finally accurately writes the corresponding report. MDAKF introduces two modules, namely, the multimodal data-assisted knowledge module (MDAK) and the multimodal data fusion module (MDF). MDAK can mitigate visual data bias by extracting abnormal regions based on input images. MDF can enhance the information exchange between different modalities [15] so that text and audio multimodal [16] information complement each other. Thus, our MDAKF model can generate more accurate and sensitive medical report descriptions of rare anomalous causes.

In summary, the main contributions of this manuscript are as follows:

- (1) To alleviate the data bias problem, we proposed a multimodal data-assisted knowledge fusion network, including multimodal data-assisted knowledge (MDAK) and multimodal data fusion (MDF).
- (2) To further enhance the audio-aided knowledgeability in the Transformer model, we introduced a variety of multimodal feature fusion methods to combine text and audio features to achieve data complementarity and jointly promote model performance.
- (3) A variety of related experiments were performed on IU-X-ray and COV-CTR datasets, and the data analysis of various performance indicators confirmed the effectiveness of the proposed method and outperformed the previous models.

The rest of this manuscript is organized as follows: Section 2 describes the related work and methods for generating medical image descriptions. Section 3 describes the overall architecture of the MDAKF network. This is followed by experimental results (see Section 4) and our conclusions (see Section 5).

2. Related Works

This manuscript introduces related work in three aspects: transformer-based image description, automatic medical image report generation, and multimodal feature fusion.

2.1. Transformer-Based Image Description. The task of image description generation has received extensive attention from many researchers. Most current approaches for the automatic generation of image text descriptions use an end-to-end encoding-decoding structure. The encoder typically uses a convolutional neural network (CNN) to perform feature extraction on the input image. The decoder uses a recurrent neural network RNN or a variant LSTM for natural language processing to convert the input image features on the encoding side into text descriptions. Some researchers introduced the Transformer encoding and decoding architecture [17–19] instead of the above network architecture; experiments have shown that the Transformer model can improve various performance indicators in medical image description tasks. The grid features extracted in Transformer are also called feature maps, although its advantage lies in its ability to cover the entire image and capture the details of the target [20]. However, the semantic hierarchy of this grid feature is usually relatively low. Ji et al. [21] suggested utilizing a faster R-CNN network to extract features of the target region. They then incorporated global information into transformer encoders to combine the benefits of both top-down and bottom-up image description generation schemes.

M2Transformer [22] utilizes an encoder-decoder architecture to transform images into descriptive text, where decoders are mesh-connected to encoders and the input of each decoder is weighted by the results of all encoders,

facilitating the capture of more detailed features. At the same time, attention slots are used in the attention mechanism of the encoding phase to increase memory and provide a priori knowledge for the subsequent process.

The dual-level [17] Transformer aligns the regions and grid features. It utilizes the comprehensive relation attention (CRA) module to obtain self-attention information for regions and grids by focusing on absolute position information and relative position information. Subsequently, the self-attention information is aligned by utilizing the locality-constrained cross attention (LCCA) module. Finally, the input features undergo encoding and decoding processes in order to generate a textual description.

2.2. Automatic Generation of Medical Image Reports. Current medical image description tasks are still dominated by the encoder-decoder framework, which translates images into individual descriptive sentences. This framework has been very successful in driving the technology forward. However, medical image description differs to a greater extent from natural image description in that it is more specialized and diverse. Describing specific regions in radiological images has greater accuracy requirements. Park et al. [23] proposed generating medical image reports that used the encoder-decoder framework. A convolutional neural network (CNN) was used as encoding to extract the image features. The decoder was composed of a recurrent neural network (RNN) that performs well in language generation tasks [24], such as long short-term memory network [25] (LSTM) and gated recurrent unit [26] (GRU). Later, in order to increase the variability of the generated text, Chen et al. [27] proposed a dual LSTM to generate standard and anomaly reporting information separately. Recently, some researchers have used Transformer at the encoder and decoder [22] as the network architecture for image description generation.

To make the generated radiology reports more accurate and convincing, some researchers started using deep learning to simulate the process that doctors go through when writing medical reports. For example, when radiologists observe medical images, they combine knowledge and work experience in the medical field to write a complete diagnostic report that reflects medical images. Zhang et al. [28] found that auxiliary signals help in the generation of image descriptions. Liu et al. [29] proposed adding a text visual dual attention mechanism to the CNN-LSTM structure to make the generated text information more complete. It was shown that the introduction of auxiliary signals positively affected the generation of various types of image descriptions.

When processing NLP tasks, previous textual information is usually crucial for subsequent studies. LSTM, RNN, GRU, and other deep learning models have the function of recording previous textual information, but the above models have specific storage mechanisms, and the capacity of their storage modules can be limited, which may be the reason for the effectiveness of the decoding function. Song et al. [30, 31]

proposed a new storage component relational memory RM to save feature information to make the model have a larger storage capacity. It can be seen from the experimental results that improving the storage module in the transformer can indeed improve overall performance.

2.3. Multimodal Feature Fusion. Modal is a form of expression of information in the computer field, and multimodal refers to the combination of multiple modalities. Multimodal feature fusion is the focus of current research on multimodal information processing [32]. The expression of different modes is not exactly the same, directly fusing them may appear as information redundancy. A good multimodal feature fusion algorithm can make the feature information richer [33].

Multimodal feature fusion techniques are currently used in various computer vision fields, such as image description generation and image segmentation. In the early stages of the study, the commonly used multimodal fusion methods mainly included element product, element sum, or even simple concatenation between different types of features, which are somewhat simple but lack in-depth analysis [34]. In 2018, Wu and Han [35] proposed a new multimodal fusion method, namely, multimodal cyclic fusion. This feature fusion can take full advantage of the interactions between multimodal feature elements and further improve the performance. Specifically, after reshaping visual or text vectors into cyclic matrices, respectively, they defined two interaction operations between the original feature vectors and the reshaped cyclic matrix. Finally, they used element-by-element sums to obtain a joint representation of these two cross-fused vectors. As each row of the cyclic matrix was shifted by one element, using the newly defined interaction operations, they explored almost all possible interactions between different modal vectors. Recently, Yang et al. [36] introduced a new transformer-based architecture that used “fusion bottlenecks” for multilayer modal fusion through a small amount of bottleneck delay, so that the information between different modalities was processed by the model and the necessary information was shared. The results showed that this algorithm can not only reduce the computational cost but also improve the fusion performance.

3. The Proposed Method

Automatic medical image report generation is a cross-modal task combining text and image vision and is essentially an image-to-text generation task. We take radiological images as input to the model and transform them into the corresponding source sequence $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$, $\mathbf{x}_n \in \mathbf{R}^d$, where d is the medical image visual feature extracted from the visual extractor and d is the size of the feature vector. The corresponding report is the target sequence $\mathbf{Y} = (y_1, y_2, \dots, y_t)$, $y_t \in V$, where y_t is the generated token, t is the length of the generated token, and V is the vocabulary of all possible tokens. An overview of our proposed model is shown in Figure 2, the details of which are illustrated in the following subsections.

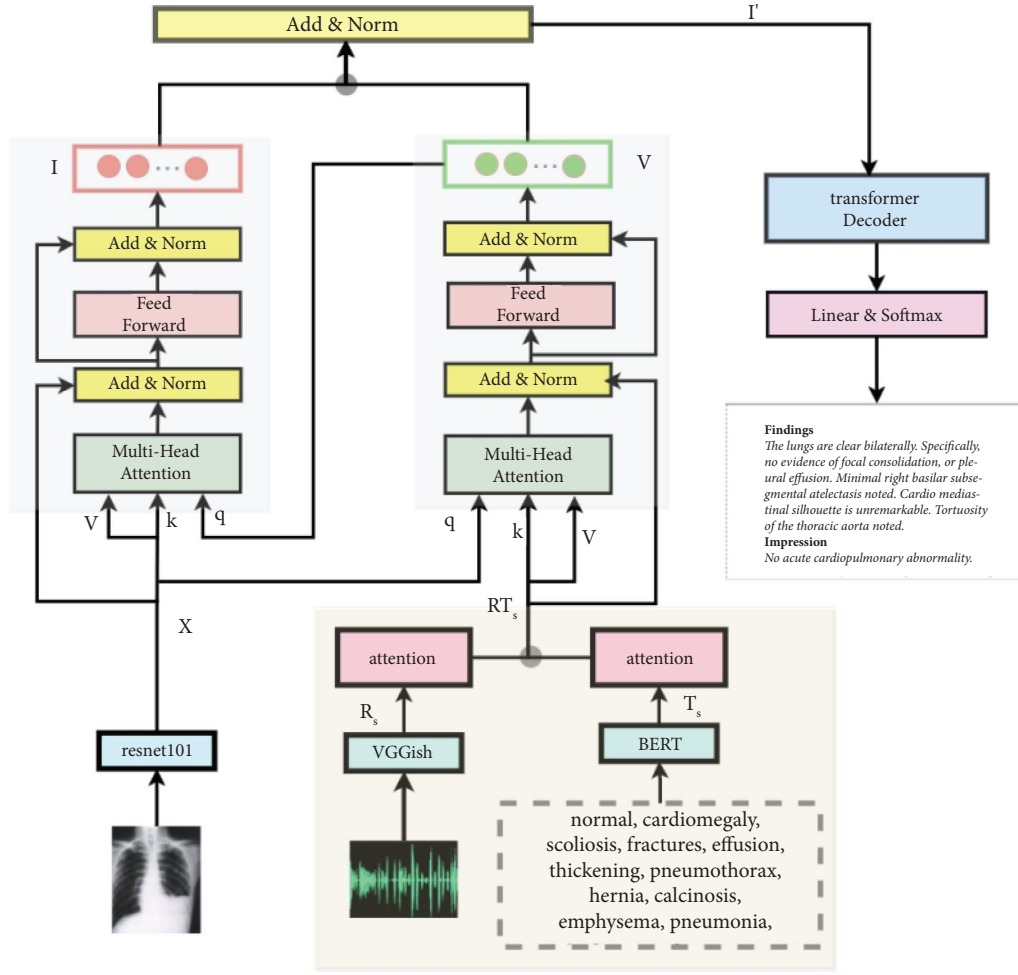


FIGURE 2: Framework of the MDAKF model.

3.1. Model Structure. Our model can be divided into four main parts: the visual feature extraction module, the multimodal-assisted signal feature extraction and fusion module, the Transformer encoder, and the Transformer decoder. Among them, our innovation is mainly in how to implement audio and text features to assist medical images to generate more accurate reports. The four components and the training objectives of the tasks are described in detail below.

3.1.1. Visual Feature Extraction. The visual extraction module passes the input medical image to extract visual feature sequence X_n by a pretrained convolutional neural network (CNN), such as VGG or ResNet, and the encoded result is used as the source sequence for all subsequent modules. The process is formulated as shown in the equation:

$$\begin{aligned} X_n &= \{x_1, x_2, \dots, x_n\} \\ &= f(\text{Image}). \end{aligned} \quad (1)$$

3.1.2. Multimodal Data-Assisted Knowledge Fusion. The multimodal auxiliary signal feature extraction and fusion module consists of two main parts: one part is to perform feature extraction on the two types of auxiliary signals of audio and text through VGGish [37] and BERT [38] models, respectively, and finally obtain R_s and T_s . The other part is to realize the feature alignment and fusion of multimodal features to obtain RT_s :

$$\begin{aligned} R_s &= \{r_1, r_2, \dots, r_s\} \\ &= \text{VGGish}(\text{audio}), \\ T_s &= \{t_1, t_2, \dots, t_s\} \\ &= \text{BERT}(\text{text}), \\ RT_s &= \text{fusion}(R_s, T_s). \end{aligned} \quad (2)$$

3.1.3. Transformer Encoder. We used a two-branch transformer standard encoder for encoding the visually extracted features and the auxiliary signal features, respectively. Here

the output is the hidden states I and V encoded from the input features X_n and RT_s . Finally, since the feature hidden sequences I and V are aligned, they are directly summed to obtain I' , where LayerNorm denotes layer normalization:

$$\begin{aligned} I &= \{\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_n\} \\ &= f(x_1, x_2, \dots, x_n), \\ V &= \{\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_s\} \\ &= f_e(\mathbf{rt}_1, \mathbf{rt}_2, \dots, \mathbf{rt}_s), \\ I' &= \text{LayerNorm}(I + V). \end{aligned} \quad (3)$$

3.1.4. Transformer Decoder. The backbone decoder uses a transformer variant architecture (R2Gen) containing RM [30] storage components, which converts the layer normalization of each decoder to MCLN [30]. The transformer decoder is also stacked by a multihead attention mechanism and feedforward neural network. The hidden state output from the encoding side and the sequence of y_{t-1} output from the previous time slice at the decoding side are fed to the decoding side to finally obtain the target vector sequence y_t :

$$y_t = f_d(I', \text{MCLN}(\text{RM}(y_1, \dots, y_{t-1}))). \quad (4)$$

3.2. Multimodal Auxiliary Signal. On the original model benchmark, they can implement the recording and storage of frequently occurring words and phrases in the RM structure

to drive the model to eventually learn and generate medical reports with more accurate and fluent descriptions. However, this cannot solve the problem of insensitivity of the existing automatic medical report generation techniques to medical abnormal images. We improved the existing model by adding auxiliary signals to enhance the overall performance of the model. Then, we can describe in detail how to use the auxiliary signals to the transformer network architecture.

The use of text tags as auxiliary signals is relatively poor in the actual system, and speech is the mainstream way of human-computer interaction; it will be more convenient. Therefore, we choose audio as the auxiliary signal. We use audio by fusing text features to make it have better semantic information. It can be better aligned with the visual features generated by annotated medical images.

The audio-aided data are selected from the image training corpus with a high frequency of abnormal keywords, such as "emphysema," "pneumonia," "cardiomegaly," "pneumothorax," and "lesion," which contain the attributes of abnormal content categories and abnormal regions in medical images. Finally, the text data are broadcasted by the machine to generate an audio file.

The transformer encoder contains mainly the multi-headed attention mechanism MHA and the feedforward neural network FFN. The MHA consists of n parallel attention strings, as shown in the equation:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (5)$$

$$\text{MHA} = [\text{Attention}_1(Q, K, V); \text{Attention}_2(Q, K, V); \dots \text{Attention}_n(Q, K, V)]. \quad (6)$$

q denotes extracted visual features X_n , and k and v denote audio text fusion features RT_s . Then, these sequences are entered into the encoder of the transformer to obtain the hidden sequence V focusing on visual and audio text fusion

features. Then, we make V as q and visual features as k , v , and the encoder obtain the visual features I with aligned auxiliary information:

$$\text{Attention}(X_n, RT_s) = \text{softmax}\left(\frac{X_n W^Q (RT_s W^K)^T}{\sqrt{d_k}}\right) RT_s W^V, \quad (7)$$

$$V = \text{FFN}(\text{MHA}(X_n, RT_s)),$$

$$I = \text{FFN}(\text{MHA}(V, X_n)).$$

However, the audio modality signals with similar pronunciation lack recognizability, which may lead to incorrect mapping of audio labels to medical image regions. Therefore, we further fuse the audio with text features as the auxiliary signal to improve the overall performance of the model.

3.3. Multimodal Data Feature Fusion. In this manuscript, we fuse audio and text representations by the feature fusion module to obtain a fusion representation of multimodal features instead of the original single audio representation. It is semantically more discernible and crosses the heterogeneous gap.

The four feature fusion schemes are used for the multimodal fusion of two auxiliary information feature vectors: such as Add, concat, and mul product operations and attention selection fusion (ATT). The fusion method is shown in Figure 3.

Add is the point-by-point summation of R_s and T_s , as shown in the equation:

$$\begin{aligned} RT_s &= F(R_s, T_s) \\ &= R_s \oplus T_s. \end{aligned} \quad (8)$$

Concat is the feature splicing of R_s and T_s , as shown in the equation:

$$\begin{aligned} RT_s &= F(R_s, T_s) \\ &= \text{CON}(R_s, T_s). \end{aligned} \quad (9)$$

Mul is the Hadamard product operation of R_s and T_s , as shown in the equation:

$$\begin{aligned} RT_s &= F(R_s, T_s) \\ &= R_s \otimes T_s. \end{aligned} \quad (10)$$

ATT refers to constructing two attention mechanisms for audio and text representations separately and then performing a point-by-point summation operation on the features after the attention mechanism is selected, as shown in the following equation:

$$A_{R_s} = \text{sigmoid}(L1(R_s)), \quad (11)$$

$$A_{T_s} = \text{sigmoid}(L2(T_s)), \quad (12)$$

$$RT_s = F(R_s, T_s) = A_{R_s} \otimes R_s \oplus A_{T_s} \otimes T_s,$$

where $L1$ and $L2$ denote the two fully connected layers and A_{R_s} and A_{T_s} denote the features acquired after passing the attention mechanism.

Then, by interacting with the multimodal auxiliary signal and the multiheaded attention mechanism in the encoder, the trained model can generate medical analysis reports that pay more attention to the regions indicated by the auxiliary signals. This solution addresses the issue of imbalanced data distribution within medical datasets.

4. Experimental Setup

In this section, we describe in detail two public datasets along with some widely used metrics and experimental settings. Then, we evaluated and analyzed the proposed approach.

4.1. Datasets, Indicators, and Parameter Settings

4.1.1. Datasets. We conducted experiments on two public datasets: IU-X-ray [39, 40] and COV-CTR [41, 42].

IU-X-ray is a widely used benchmark dataset for evaluating the performance of radiology report generation methods. It contains 7470 chest X-ray images associated with 3955 radiology reports. We randomly split the dataset

into 7:1:2 training validation test sections. There is no overlap of patients in the training, validation, and test sets.

The COV-CTR dataset contains lung CT images and their corresponding diagnostic Chinese reports, where the lung CT images are collected during the COVID-19 outbreak, and Li et al. [41] provided the corresponding diagnostic reports to construct the COV-CTR dataset. It includes a total of 728 images, of which 349 are COVID-19 and 379 are non-COVID-19. For a fair comparison, we randomly divided the data into the training set, validation set, and test set in the ratio of 8:1:1.

4.1.2. Performance Metrics. It refers to the evaluation of the image description generation model, judging the quality of the description generated by the model. Typically, the experiment can use an automated rule-based evaluation method in medical image description tasks [43]. This methodology entails the prior collection of a predetermined quantity of reference descriptions that have been authored by human beings specifically for the provided image. The evaluation of similarity between the description produced by the model and the reference description is achieved by employing keyword matching. This approach can be utilized as a means of evaluating the efficacy of the model. The mainstream metrics include ROUGE, BLEU, CIDEr, and METEOR.

BLEU calculates the degree of overlap of the N-tuples in the generated report and the target report to measure the similarity between statements. METEOR calculates the similarity between candidate and reference texts based on word-level accuracy and recall, as well as penalties for word order. METEOR has the flexibility to handle word matching and word order problems, so it is more reflective of human evaluation of text quality than BLEU. ROUGE-L is responsible for calculating the longest common subsequence of a sentence. CIDEr calculates the cosine similarity between the real description and the model-generated description to measure the effect of the image description.

4.1.3. Parameter Settings. The datasets are pretrained on ImageNet, and the extracted features are 2048 7×7 shaped feature maps, which are further projected into 512 feature maps. For these two datasets, the same hyperparameters are used for training. Specifically, the learning rates of the visual extractor and other parameters are set to $5e-5$ and $1e-4$, respectively, and the batch size is 4. In addition, the number of heads and dimensions of the multiheaded attention is 8 and 512, respectively.

4.2. Model Performance Comparison. We compare our approach with a series of state-of-the-art radiological report generation models (Transformer [44], M2Transformer [22], CoAtt [7], HGRG-Agent [45], KERP [46], PPKED [14], SAT [47], AdaAtt [48], R2Gen [30], and ASGMD [49]). For the IU X-ray dataset, the R2Gen, PPKED, Transformer, MDAK (our), and MDAKF (our) model indicator data are obtained through our experiments. In contrast, the other model

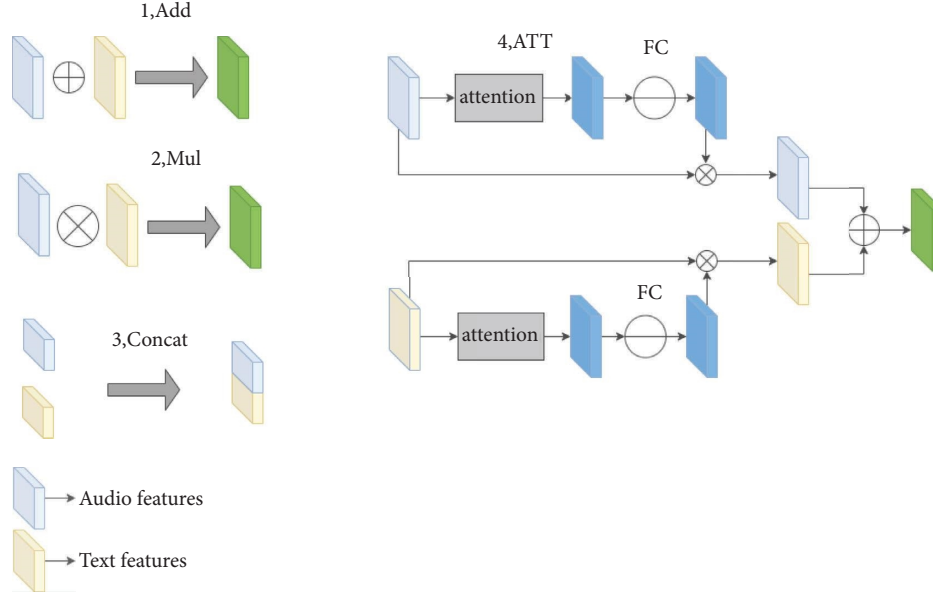


FIGURE 3: Operation of multimodal feature fusion.

indicator data are the results of the original papers. For the COV-CTR dataset, the R2gen, MDAK (our), and MDAKF (our) model indicator data are obtained through experiments, while the remaining experimental data are obtained by referring to the original paper of the ASGK [41] model. As shown in Table 1, MDAKF outperforms state-of-the-art methods in some of the metrics for both the COV-CTR and IU X-ray datasets, which proves the effectiveness and accuracy of incorporating audio as an auxiliary signal in medical image description. Specifically, on the IU X-ray dataset, the MDAK method increases from 0.398 to 0.424 on the evaluation metric CIDEr, which is specifically used to assess the quality of generating reports. On the COV-CTR dataset, the two methods of MDAK and MDAKF reach 1.452 and 1.243 on CIDEr evaluation indicators, respectively.

4.3. Ablation Experiments. In this section, a quantitative analysis is performed to investigate the contribution of each component in MDAKF. The experimental results of adding MDAK and MDAKF are shown in Table 2 below.

From Table 2, it can be seen that the performance of the MDAK model with audio-assisted signals is superior to that of the base (R2gen), which significantly improves the quality of report generation, fully verifying the effectiveness of the MDAK module. The automatic generation indicators of image reports in MDAK have shown some improvement, especially METEOR, and CIDEr. The METEOR score increases from 0.187 to 0.201, and the CIDEr score increases from 0.398 to 0.424. The indicators have also increased on the COV-CTR dataset. For example, the BLEU4 score increases from 0.528 to 0.539, and the ROUGE_L score increases from 0.677 to 0.683. The input of visual features in medical images significantly affects natural language decoders. Therefore, by simulating the working mode of radiologists through audio-assisted signals, the visual encoding process of the model will focus more on the image area aligned with the audio-assisted signal, ultimately

providing richer visual features. These experiments indicate that focusing on the abnormal areas specified by audio tags can improve the quality of medical report generation.

To improve the performance of auxiliary signals, we add a multimodal auxiliary signal fusion module (MDAKF). It can combine the semantic information of text and audio cross-modal data. To verify its effectiveness, we cite four different feature fusion schemes. The symbols in parentheses in MDAKF (add, ATT, cat, and mul) indicate the feature fusion scheme, which can be seen in detail in 3.3. From the experimental results in Table 2, it can be seen that the addition of the MDAKF module can continue to improve the overall performance of the model, and it also has a certain improvement in various indicators. Among them, MDAKF (add) has better overall performance, so we chose it as the final model for MDAKF. In Table 2, the performance indicators of the MDAKF (add) model are improved. In terms of BLEU1 and ROUGE_L indicators, the BLEU1 score increases from 0.470 to 0.494, and the ROUGE_L score increases from 0.371 to 0.389. It can also be seen that multimodal feature fusion has a significant impact on model performance. Therefore, designing a more reasonable multimodal fusion scheme is also a key research direction in the future. Finally, the experimental results indicate that MDAKF can provide better semantic feature information guidance for the model, further improving the performance of the model in generating medical reports.

4.4. Feature Extraction Experiments. Feature extraction has an impact on the MDAKF model; therefore, we also conduct comparative experiments on feature extraction networks. In this part, we explore the affection of different visual feature extraction networks for MDAKF (ATT) on the IU-X-ray dataset and compare it with the baseline. The following networks are selected in the visual feature extraction module: ResNet-101, ResNet-152, and multiple variations of ResNet

TABLE 1: Performance of different methods on the IU X-ray and COV-CTR datasets.

Datasets	Methods	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE_L	CIDEr
IU-X-ray	Transformer	0.422	0.264	0.177	0.120	0.164	0.338	0.421
	CoAtt	0.455	0.288	0.205	0.154	—	0.369	0.277
	HRGR-Agent	0.438	0.298	0.208	0.151	—	0.322	0.343
	PPKED	0.483	0.315	0.224	0.168	0.190	0.376	0.351
	KERP	0.482	0.325	0.226	0.162	0.187	0.339	0.280
	M2Transformer	0.463	0.318	0.214	0.155	0.192	0.335	—
	ASGMD	0.489	0.326	0.232	0.173	0.206	0.397	—
	R2Gen (base)	0.470	0.304	0.219	0.165	0.187	0.371	0.398
	MDAK (our)	0.480	0.328	0.231	0.172	0.201	0.369	0.424
COV-CTR	MDAKF (our)	0.494	0.318	0.229	0.174	0.194	0.389	0.371
	CoAtt	0.709	0.645	0.603	0.552	—	0.748	—
	SAT	0.697	0.621	0.568	0.515	—	0.723	—
	ASGK	0.712	0.659	0.611	0.570	—	0.746	—
	AdaAtt	0.676	0.633	0.596	0.514	—	0.726	—
	R2Gen	0.725	0.641	0.580	0.528	0.399	0.677	1.358
	MDAK (our)	0.723	0.652	0.586	0.545	0.403	0.676	1.452
	MDAKF (our)	0.726	0.651	0.583	0.539	0.401	0.683	1.354

The bold values indicate that the model performance of the algorithm is optimal in a certain type of dataset.

TABLE 2: Ablation experiments of each module.

Dataset	Methods	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE_L	CIDEr
IU-X-ray	R2Gen (base)	0.470	0.304	0.219	0.165	0.187	0.371	0.398
	MDAK	0.480	0.328	0.231	0.172	0.201	0.369	0.424
	MDAKF (add)	0.494	0.318	0.229	0.174	0.194	0.389	0.371
	MDAKF (ATT)	0.505	0.318	0.219	0.159	0.195	0.383	0.344
	MDAKF (cat)	0.484	0.307	0.221	0.167	0.192	0.391	0.334
	MDAKF (mul)	0.457	0.291	0.209	0.159	0.179	0.371	0.372
COV-CTR	R2Gen (base)	0.725	0.641	0.580	0.528	0.399	0.677	1.358
	MDAK	0.723	0.652	0.586	0.545	0.403	0.676	1.452
	MDAKF (add)	0.726	0.651	0.583	0.539	0.401	0.683	1.354
	MDAKF (ATT)	0.727	0.649	0.588	0.537	0.400	0.674	1.243
	MDAKF (cat)	0.722	0.640	0.576	0.524	0.405	0.683	1.305
	MDAKF (mul)	0.718	0.637	0.574	0.521	0.401	0.681	0.302

The bold values indicate that the model performance of the algorithm is optimal in a certain type of dataset.

TABLE 3: Performance of different feature extraction networks.

	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE_L
ResNet-101	0.505	0.318	0.219	0.159	0.195	0.383
ResNet-152	0.489	0.310	0.219	0.157	0.210	0.375
ResNet_101_32 × 8d	0.493	0.306	0.203	0.137	0.198	0.366
wide_ResNet-101_2	0.499	0.309	0.206	0.143	0.198	0.346

The bold values indicate that the model performance of the algorithm is optimal in a certain type of dataset.

network, ResNet_101_32 × 8d, and wide_ResNet – 101_2 network. The experimental results are shown in Table 3. We can see that the optimal performance of the MDAKF model can be achieved by using the ResNet-101 network to extract visual features.

4.5. Visualization Experiments. In order to further verify the validity of our model, we select some medical images from the IU-X-ray and COV-CTR medical image report generation datasets for qualitative analysis. These medical images are used in different models to generate reports. As shown in Figure 4, the first three generated report instances are

selected from IU-X-ray, and the fourth report is selected from COV-CTR. We can observe the medical reports generated by the MDAKF and R2Gen models.

In reports, we can see that the generated reports all abide by a process pattern, reporting first abnormal findings (e.g., “cardiac silhouette” and “lung volume”), followed by underlying disease (e.g., “pleural effusion” and “nontenderness”). In addition, for the necessary medical terms in the basic fact report, MDAKF covers almost all of these terms in its generated reports. We compare the reports generated by R2Gen and MDAKF with the Ground_Truth reports and used red and blue to distinguish the overlap between them and the actual report.





Images				
Ground_Truth	Cardiomediastinal silhouette and pulmonary vasculature are within normal limits . lungs are clear. no pneumothorax or pleural effusion . No acute osseous findings .	The cardiomeastinal silhouette and pulmonary vasculature are within normal limits in size . The lungs are clear of focal airspace disease, pneumothorax or pleural effusion . There are no acute bony findings .	Cardiomediastinal silhouettes are within normal limits . lungs are clear without focal consolidation pneumothorax or pleural effusion . bony thorax is unremarkable .	the thorax was symmetrical, the mediastinal heart shadow was centered , the texture of both lungs was clear and patchy and ground glass shadow was seen in the lower lobe of both lungs with some fuzzy edges . no abnormal density shadow was seen in bilateral thoracic cavities .
R2Gen	The heart pulmonary xxxx and mediastinum are within normal limits . There is no pleural effusion or pneumothorax . There is no focal air space opacity to suggest a pneumonia . There is a calcified granuloma in the left lung base .	Heart size is mildly enlarged . Tortuous aorta . lungs are normally inflated and clear . mild degenerative changes of the spine .	The heart is large . lung volumes are xxxx . . xxxx opacity <unk> in the right midlung . . no focal infiltrates .	the thorax is symmetrical the mediastinal cardiac shadow is in the middle the texture of both lungs is clear. patchy solid shadow is seen in the dorsal part of the lower lobe of the right lung some of the edges are clear and the bronchial air image is seen in the lower lobe of the left lung small shadow
Ours	The cardiomeastinal silhouette and pulmonary vasculature are within normal limits in size . The lungs are clear of focal airspace disease pneumothorax or pleural effusion . lung volumes are low normal . There are no acute bony findings .	The cardiomeastinal silhouette and pulmonary vasculature are within normal limits in size . The lungs are clear of focal airspace disease, pneumothorax or pleural effusion . lung volumes are low normal . There are no acute bony findings .	Low lung volumes . Cardiomeastinal silhouette and pulmonary vasculature are within normal limits . lungs are clear . No pneumothorax or pleural effusion . Calcified bilateral hilar lymph xxxx greater on the left . No acute osseous findings .	the thorax was symmetrical the mediastinal cardiac shadow was centered no enlarged lymph nodes were seen in the mediastinum the texture of both lungs was enhanced a patchy solid shadow with fuzzy margins was seen in the lower lobe of the left lung the bronchi of the lobe were clear and no abnormal density shadow was seen in the bilateral

FIGURE 4: Visualization of medical image description.

As shown in Figure 4, our proposed network model outperforms the baseline model in generated medical reports. The MDAKF module can provide more accurate anomalous visual regions during model training, thereby alleviating the problem of visual data bias. To verify this

result, we extract multiple chest X-rays from the IU-X-ray and COV-CTR datasets and visualize the image and audio attention mapping guided by the MDAKF module. In Figure 5, the first medical image is from the COV-CTR dataset, and the second medical image is from the IU-X-ray


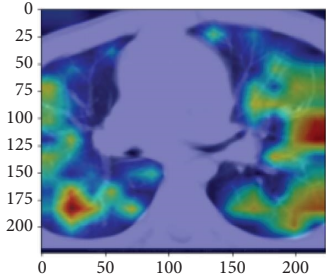

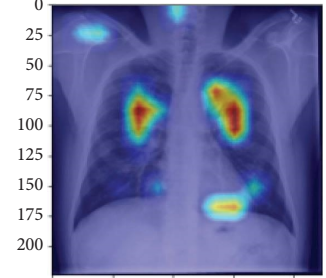
Image	Heatmap	Report
		The thorax was symmetrical, the mediastinal heart shadow was in the middle, no enlarged lymph nodes were seen in the mediastinum, the texture of both lungs was enhanced, the left lung was seen to have limited increased transmission.
		the cardiomedastinal silhouette and pulmonary vasculature are within normal limits in size . the lungs are clear of focal airspace disease pneumothorax or pleural effusion . lung volumes are low normal . there are no acute bony findings.

FIGURE 5: Attentional visualization of MDAKF.

dataset. Through attention heatmap analysis, it is demonstrated that the MDAKF module enables the model to focus more on the abnormal regions.

5. Conclusion

In this manuscript, the multimodal data-assisted knowledge fusion network is proposed to automatically generate medical reports. The network is based on the R2Gen framework and aims to facilitate the generation of diagnostic reports using multimodal auxiliary information features. We study different types of auxiliary signals to achieve the purpose of automatically generated medical reports that focus on disease regions better and alleviate the visual data bias problem. Prominent experiments have demonstrated the effectiveness of our proposed MDAKF network. In the future, we will investigate more efficient multimodal feature fusion methods to enhance the auxiliary signal feature representation.

Data Availability

Two public medical image description datasets were used to support this study and are available at (IU-X-ray: <https://drive.google.com/file/d/1c0BXEuDy8Cmm2jfN0YYGkQxXfZd2ZIoLg/view?usp=sharing>; COV-CTR: <https://github.com/mlil0117/COV-CTR/tree/master/Datasets>). These prior studies (and datasets) are cited at relevant places within the text as references [39–41].

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (No. 62002392 and No. 62372478), in part by the Key Research and Development Plan of Hunan Province (No. 2019SK2022), and in part by the Natural Science Foundation of Hunan Province (No. 2022JJ31019).

References

- [1] N. Yu, X. Hu, B. Song, J. Yang, and J. Zhang, "Topic-oriented image captioning based on order embedding," *IEEE Transactions on Image Processing*, vol. 28, no. 6, pp. 2743–2754, 2019.
- [2] K. Iwamura, J. Y. Louhi Kasahara, A. Moro, A. Yamashita, and H. Asama, "Image captioning using motion-CNN with object detection," *Sensors*, vol. 21, no. 4, p. 1270, 2021.
- [3] P. Liu, Y. Zhou, D. Peng, and D. Wu, "Global-attention-based neural networks for vision language intelligence," *IEEE/CAA Journal of Automatica Sinica*, vol. 8, no. 7, pp. 1243–1252, 2021.
- [4] Y. Jing, X. Zhiwei, and G. Guanglai, "Context-driven image caption with global semantic relations of the named entities," *IEEE Access*, vol. 8, pp. 143584–143594, 2020.
- [5] L. Zhao, X. Lyu, J. Song, and L. Gao, "Guess Which? Visual dialog with attentive memory network," *Pattern Recognition*, vol. 114, Article ID 107823, 2021.

- [6] Q. Yang, Z. Ni, and P. Ren, "Meta captioning: a meta learning based remote sensing image captioning framework," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 186, pp. 190–200, 2022.
- [7] B. Jing, P. Xie, and E. Xing, "On the automatic generation of medical imaging reports," 2020, <https://arxiv.org/abs/1711.08195>.
- [8] N. Naqvi, M. S. Islam, M. Iqbal, S. Kanwal, A. Khan, and Z. Ye, "Deep neural combinational model (DNCM): digital image descriptor for child's independent learning," *Multimedia Tools and Applications*, vol. 81, pp. 29955–29975, 2022.
- [9] J. Yan, Y. Xie, X. Luan, Y. Guo, Q. Gong, and S. Feng, "Caption TLSTMs: combining transformer with LSTMs for image captioning," *International Journal of Multimedia Information Retrieval*, vol. 11, no. 2, pp. 111–121, 2022.
- [10] J. Cao, Y. Jiang, and P. Sun, "Transtrack: multiple object tracking with transformer," 2020, <https://arxiv.org/abs/2012.15460>.
- [11] A. Johnson, M. Lungren, and Z. Lu, "Mimic-cxr-jpg- chest radiographs with structured labels," 2023, <https://arxiv.org/abs/1901.07042>.
- [12] H. Ayesha, S. Iqbal, M. Tariq et al., "Automatic medical image interpretation: state of the art and future directions," *Pattern Recognition*, vol. 114, Article ID 107856, 2021.
- [13] V. Aswiga and P. Shanthi, "A multilevel transfer learning technique and LSTM framework for generating medical captions for limited CT and DBT images," *Journal of Digital Imaging*, vol. 35, no. 3, pp. 564–580, 2022.
- [14] X. Wu, S. Ge, and F. Liu, "Exploring and distilling posterior and prior knowledge for radiology report generation," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, vol. 3, no. 1, pp. 13753–13762, 2021.
- [15] C. Zeng and S. Kwong, "Learning cross-modality features for image caption generation," *International Journal of Machine Learning and Cybernetics*, vol. 13, no. 7, pp. 2059–2070, 2022.
- [16] Y. Huang, J. Chen, W. Ouyang, W. Wan, and Y. Xue, "Image captioning with end-to-end attribute detection and subsequent attributes prediction," *IEEE Transactions on Image Processing*, vol. 29, no. 1, pp. 4013–4026, 2020.
- [17] Y. Luo, J. Ji, X. Sun et al., "Dual-level collaborative transformer for image captioning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 3, pp. 2286–2293, 2021.
- [18] M. Luo, P. Zhou, and W. Yu, "Metaformer is actually what you need for vision," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, vol. 3, no. 1, pp. 10819–10829, 2022.
- [19] Z. Wang, S. Shi, Z. Zhai, Y. Wu, and R. Yang, "ArCo: attention-reinforced transformer with contrastive learning for image captioning," *Image and Vision Computing*, vol. 128, Article ID 104570, 2022.
- [20] P. Bharati and A. Pramanik, "Deep learning techniques-R-CNN to mask R-CNN: a survey," *Computational Intelligence in Pattern Recognition*, vol. 3, no. 1, pp. 657–668, 2020.
- [21] J. Ji, Y. Luo, X. Sun et al., "Improving image captioning by leveraging intra- and inter-layer global representation in transformer network," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 2, pp. 1655–1663, 2021.
- [22] M. Stefanini, L. Baraldi, and M. Cornia, "Meshed-memory transformer for image captioning," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, vol. 3, no. 1, pp. 10578–10587, 2020.
- [23] H. Park, K. Kim, S. Park, and J. Choi, "Medical image captioning model to convey more details: methodological comparison of feature difference generation," *IEEE Access*, vol. 9, no. 1, pp. 150560–150568, 2021.
- [24] N. X. Hanwang, "Multi-level policy and reward-based deep reinforcement learning framework for image captioning," *IEEE Transactions on Multimedia*, vol. 22, no. 5, pp. 1372–1383, 2019.
- [25] Z. Ye, R. Khan, N. Naqvi, and M. S. Islam, "A novel automatic image caption generation using bidirectional long-short term memory framework," *Multimedia Tools and Applications*, vol. 80, no. 17, pp. 25557–25582, 2021.
- [26] C. Xu, M. Yang, X. Ao, Y. Shen, R. Xu, and J. Tian, "Retrieval-enhanced adversarial training with dynamic memory-augmented attention for image paragraph captioning," *Knowledge-Based Systems*, vol. 214, no. 1, Article ID 106730, 2021.
- [27] Y. Chen, F. Chen, and P. Harzig, "Addressing data bias problems for chest x-ray image report generation," 2019, <https://arxiv.org/abs/1908.02123>.
- [28] Y. Zhang, X. Wang, Z. Xu, Q. Yu, A. Yuille, and D. Xu, "When radiology report generation meets knowledge graph," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, pp. 12910–12917, 2020.
- [29] M. Liu, L. Li, H. Hu, W. Guan, and J. Tian, "Image caption generation with dual attention mechanism," *Information Processing & Management*, vol. 57, no. 2, Article ID 102178, 2020.
- [30] Y. Song, T. H. Chang, and Z. Chen, "Generating radiology reports via memory-driven transformer," 2020, <https://arxiv.org/abs/2010.16056>.
- [31] Y. Shen, Y. Song, and Z. Chen, "Cross-modal memory networks for radiology report generation," 2022, <https://arxiv.org/abs/2204.13258>.
- [32] M. A. Azam, K. B. Khan, S. Salahuddin et al., "A review on multimodal medical image fusion: compendious analysis of medical modalities, multimodal databases. fusion techniques and quality metrics," *Computers in Biology and Medicine*, vol. 144, Article ID 105253, 2022.
- [33] K. Chen, R. Wang, and Z. Zhang, *Neural Machine Translation with Universal Visual representation*, International Conference on Learning Representations, Vienna, Austria, 2020.
- [34] M. R. Hassan, S. Huda, M. M. Hassan, J. Abawajy, A. Alsanad, and G. Fortino, "Early detection of cardiovascular autonomic neuropathy: a multi-class classification model based on feature selection and deep learning feature fusion," *Information Fusion*, vol. 77, pp. 70–80, 2022.
- [35] A. Wu and Y. Han, "Multi-modal circulant fusion for video-to-language and backward," *International Joint Conference on Artificial Intelligence*, vol. 3, no. 4, p. 8, 2018.
- [36] S. Yang, A. Arnab, and A. Nagrani, "Attention bottlenecks for multimodal fusion," *Advances in Neural Information Processing Systems*, vol. 34, pp. 14200–14213, 2021.
- [37] P. Seetharaman, K. Kumar, and H. Wu, "Wav2clip: learning robust audio representations from clip," *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 12, no. 1, pp. 4563–4567, 2022.
- [38] L. Dong, S. Piao, and H. Bao, "Beit: bert pre-training of image transformers," 2020, <https://arxiv.org/abs/2106.08254>.
- [39] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman et al., "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.
- [40] drive, "IU-xray dataset," 2023, <https://drive.google.com/file/d/1c0BXEuDy8Cmm2jfn0YYGkQxZFZd2ZIoLg/view?usp=sharing>.
- [41] M. Li, R. Liu, F. Wang, X. Chang, and X. Liang, "Auxiliary signal-guided knowledge encoder-decoder for medical report

- generation,” *World Wide Web*, vol. 26, no. 1, pp. 253–270, 2023.
- [42] github, “COV-CTR dataset,” 2020, <https://github.com/mlil0117/COV-CTR>.
 - [43] X. He, B. Shi, X. Bai, G. S. Xia, Z. Zhang, and W. Dong, “Image caption generation with part of speech guidance,” *Pattern Recognition Letters*, vol. 119, no. 1, pp. 229–237, 2019.
 - [44] N. Shazeer, N. Parmar, and A. Vaswani, “Attention is all you need,” *Advances in Neural Information Processing Systems*, vol. 30, 2017.
 - [45] X. Liang, Z. Hu, and Y. Li, “Hybrid retrieval-generation reinforced agent for medical image report generation,” *Advances in Neural Information Processing Systems*, vol. 31, 2018.
 - [46] C. Y. Li, X. Liang, Z. Hu, and E. P. Xing, “Knowledge-driven encode, retrieve, paraphrase for medical image report generation,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 6666–6673, 2019.
 - [47] K. Xu, J. Ba, and R. Kiros, “Show, attend and tell: neural image caption generation with visual attention,” *International conference on machine learning*, vol. 23, pp. 2048–2057, 2015.
 - [48] J. Lu, C. Xiong, and D. Parikh, “Knowing when to look: adaptive attention via a visual sentinel for image captioning,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 375–383, Honolulu, HI, USA, June 2017.
 - [49] Y. Xue, Y. Tan, L. Tan, J. Qin, and X. Xiang, “Generating radiology reports via auxiliary signal guidance and a memory-driven network,” *Expert Systems with Applications*, vol. 237, Article ID 121260, 2024.