2023 5th International Conference on Intelligent Medicine and Image Processing (IMIP) | 979-8-3503-3739-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/IMIP57114.2023.00022

Non-symmetrical Sibling-Stream Network with Adaptive Positional Encoding for Automatic Medical Report Generation

Jiayu Si
Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology,
Anhui University,
Hefei, Anhui, China,
1023209111@qq.com

Haifeng Zhao
Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology,
Anhui University,
Hefei, Anhui, China,
senith@163.com

Lili Huang*
Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology,
Anhui University,
Hefei, Anhui, China,
hill_ahu@ahu.edu.cn

Zhengzheng Tu
Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology,
Anhui University,
Hefei, Anhui, China,
zhengzhengahu@163.com

***Abstract*—Automatic medical report generation technology aims to alleviate the workload of radiologists in writing reports and avoid possible misdiagnosis and missed diagnosis. Recent solutions in medical report generation are mainly based on two mainstream architectures: convolutional neural networks (CNNs) and transformer. However, these two architectures are insufficient for the medical report generation task due to the following three reasons. First, CNN lacks the global context information due to the constrained receptive field of convolution. Second, the complete image is cut into small patches before being fed to the transformer, which leads to the damage to the local details. Third, position encoding is critical to compensate the lost position information caused by serialized input, but the conventional position encoding (e.g., absolute positional encoding) inadequately encodes position information. Therefore, we propose a novel Non-Symmetrical Sibling-Stream Network with Adaptive Positional Encoding which includes three modules: sibling-stream encoder (SSE) module, feature interaction (FI) module, and memory-driven decoder (MDD) module. In detail, in order to cope with respective drawbacks of both architectures, we first propose sibling-stream encoder (SSE) module which combines the global context by transformer and local details (e.g., texture and edge) by CNN. Concurrently, we design a new encoding method named adaptive positional encoding (APE), which can effectively encode the absolute position information within the patch and the relative position information between patches, so as to improve the accuracy of medical report. Then, the feature interaction (FI) module is designed to enhance the image feature representation by interacting the information of sibling-stream. Finally, we introduce a memory-driven decoder to implicitly model and memorize similar patterns in different medical reports during the generation process. Extensive experiments on two public datasets, i.e., IU X-RAY and MIMIC-CXR demonstrate the effectiveness of our proposed model.**

Keywords—medical report generation, positional encoding, transformer, vision and language

I. Introduction

Medical images, such as radiology and pathology images, are widely used for the diagnosis and treatment of many diseases. Professional radiologists write radiology reports based on medical images and clinical information. Considering that this task is time-consuming and laborious, researchers use the method of automatic medical report generation to simulate the radiologists' working patterns, and generate accurate and semantically coherent radiology reports, thereby freeing the radiologists from the arduous tasks.

In recent years, automatic medical report generation has attracted extensive research interest [1, 2, 3]. Most existing studies follow the image captioning approaches and apply similar encoder-decoder structure. Wang et al. [4] employs pure convolutional neural network (CNN) to extract visual feature and adopts long short-term memory (LSTM) as decoder to generate medical report. Chen et al. [5] and Yang et al. [6] occupy CNN and transformer encoder framework in a serial way as visual feature extractor and employ text decoder to generate more clinically accurate and standardized reports. However, the generated report frequently contains some repeated statements, and are unable to describe some rare but important medical terms, such as abnormal and normal.

979-8-3503-3739-6/23/$31.00 ©2023 IEEE
DOI 10.1109/IMIP57114.2023.00022

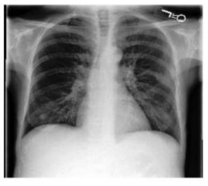
We propose a non-symmetrical sibling-stream network with adaptive positional encoding, which imitates the radiologists' working patterns to generate accurate and semantically coherent medical reports. The pipeline of the proposed method contains three modules: sibling-stream encoder (SSE) module, feature interaction (FI) module, and memory-driven decoder (MDD) module. First, transformer can model global context information, but has insufficient local information, and CNN can extract local details well, but lacks global context due to the constrained receptive field of convolution. Therefore, we develop sibling-stream encoder module to retain the advantages of the transformer and CNN. Concurrently, we design a new coding method: adaptive positional encoding (APE), which is seamlessly embedded into the transformer, and guides the model to pay more attention to the absolute position information within the patch and the relative position information between patches. Second, we propose a feature interaction (FI) module to enhance the image feature representation by interacting the global context and local details of sibling-stream. Finally, we introduce a memory-driven decoder to implicitly model and memorize similar patterns in different medical reports during the generation process. In general, the main contributions of this paper are as follows:

1) We propose a non-symmetrical sibling-stream network to combine the global modeling advantage of transformer and the local modeling advantage of CNN to achieve excellent performance.

2) The adaptive positional encoding is proposed to make up for the defect that the self-attention mechanism can not capture order of sequence information, and model the relative position between patches and the absolute position within the patch.

3) We design a feature interaction module to efficiently interact the sibling-stream feature, which can select favorable information from CNN as a supplement to transformer, as well as select favorable information from transformer as a supplement to CNN.

4) Extensive experiments on two public datasets, i.e., IU X-RAY and MIMIC-CXR demonstrate the superiority of our proposed model.



Finding

There is no focal consolidation, pleural effusion or pneumothorax. Bilateral nodular opacities that most likely represent nipple shadows. The cardiomediastinal silhouette is normal. Clips project over the left lung, potentially within the breast. The imaged upper abdomen is unremarkable. Chronic deformity of the posterior left sixth and seventh ribs are noted.

Fig. 1. An example of MIMIC-CXR image and its corresponding finding.

II. Related Work

A. Vision Transformers

Convolutional neural networks (CNN) [7] have dominated the computer vision field for many years with great success. Recently, the pioneering work ViT [8] demonstrates that pure transformer-based architectures can also achieve very competitive results. Built on the success of ViT, many efforts have been committed to designing better transformer-based architectures [9, 10, 11] for various vision tasks, including image classification [8, 10], object detection [12], image captioning [13, 14] and report generation [15, 16, 17]. In recent year, there has also been a lot of interest in combining convolutional neural network (CNN) with transformer, e.g., by using transformer to re-encode features extracted by CNN for downstream tasks. TransFuse [36] combines transformer and CNN in a parallel style, where both global dependency and low-level spatial details can be extracted in a much shallower manner. CCTNet [37] combines the local details (e.g., texture and edge) by the CNN and global context by transformer to utilize the respective advantages of the CNN and transformer.

B. Positional Encodings

Vision transformers have attracted great attention in the field of computer vision due to their competitive performance and superior ability to model long-range dependencies. The core of the transformer is the self-attention mechanism, which lacks the ability to perceive the order of tokens. Therefore, positional encodings are commonly employed to incorporate the order of sequences. In order to make up for the lack of the ability of the transformer to obtain the sequence order, the method of embedding positional encoding is usually adopted. The positional encoding mainly includes two methods, one is absolute, while the other is relative. These encoding methods are described in detail below.

Absolute Positional Encoding. The absolute positional encoding [18, 19] is proposed to solve the problem that the transformer cannot capture the timing information of the sequence. The encoding is generated with the sinusoidal functions of different frequencies, which allows the model to easily learn to focus on absolute position. However, the absolute positional encoding can make the model lose translation invariance when processing image information and prevent the model from processing longer sequences during test than during train.

Relative Positional Encoding. The relative positional encoding [20, 21] can explicitly model the positional relationship between the tokens in the input sequence of transformer, which improves the representation ability of the model. Compared to the absolute one, the relative positional encoding can keep translation invariant and naturally process longer sequences during test than during train.

III. Proposed Method

As shown in Fig. 2, we propose the non-symmetrical sibling-stream network with adaptive positional encoding for this automatic medical report generation task, where we develop sibling-stream encoder module to utilize advantage of sibling-stream network to extract deep spatial and semantic information, devise feature interaction module to enhance the feature representation from sibling-stream network and lessen noise and background information, and introduce memory-driven decoder module to generate long reports with necessary medical terms.

A. Sibling-Stream Encoder Module

We design a sibling-stream encoder (SSE) architecture. Specifically, it contains two branches: C-Stream and T-Stream. C-Stream extracts image features by sharing convolution kernels,

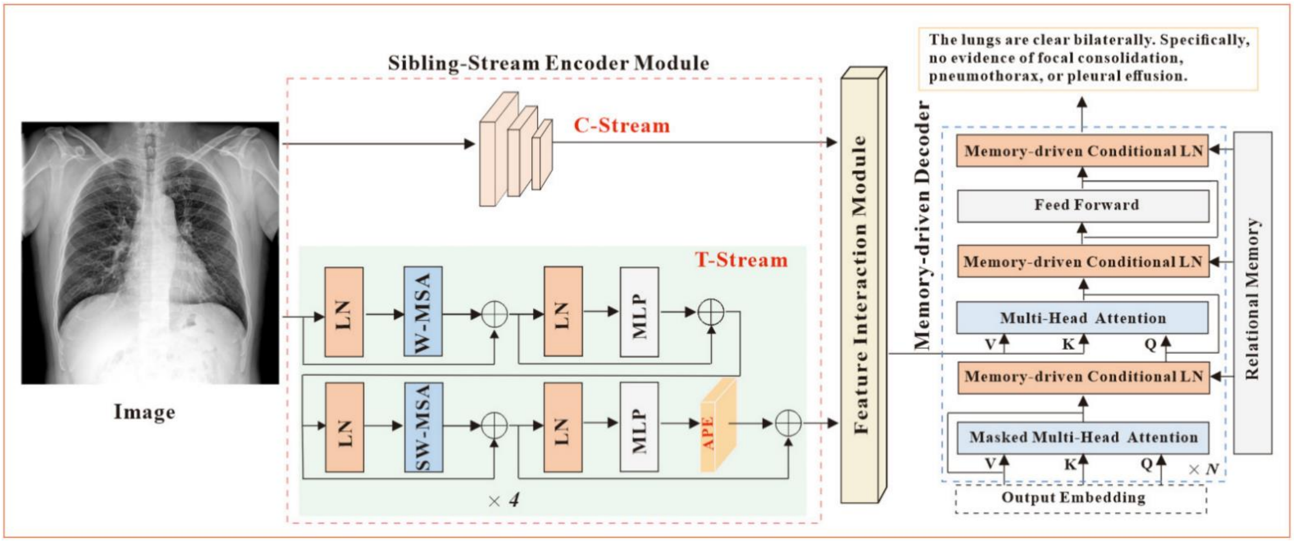

Sibling-Stream Encoder Module
C-Stream
T-Stream
Image
LN
W-MSA
LN
MLP
LN
SW-MSA
LN
MLP
APE
× 4
Feature Interaction Module
Memory-driven Decoder
The lungs are clear bilaterally. Specifically, no evidence of focal consolidation, pneumothorax, or pleural effusion.
Memory-driven Conditional LN
Feed Forward
Memory-driven Conditional LN
Multi-Head Attention
V K Q
Memory-driven Conditional LN
Masked Multi-Head Attention
V K Q
× N
Output Embedding
Relational Memory

Fig. 2. Details of the proposed model for medical report generation.

which is more inclined to obtain the local feature of the image. T-Stream pays more attention to the global context information of the image.

C-Stream: Given a set of radiology images (*I*), the visual backbone extracts the visual features X and results in the source sequence {x1, x2, ..., xs} for the subsequent visual language model. The visual backbone can be formulated based on pretrained Convolutional Neural Networks (CNN) [23, 7]. We discover Resnet101 to be more valid in our generation task and therefore employ it as our C-Stream feature extractor. The process is formulated as:

$$\{x_1, x_2, \ldots, x_s\} = f_v(I) \tag{1}$$

T-Stream: We use Swin [22] as our T-Stream backbone, it is divided into four stages, each of which has essentially the same structure, a Swin Transformer block consists of a shifted window based MSA module, followed by a 2-layer MLP with GELU nonlinearity. Before each MLP and MSA module, a LayerNorm (LN) layer is employed, and a residual link is employed after each module. Swin Transformer block is computed as:

$$\begin{aligned}
\hat{\mathbf{X}}^{l} &= \text{W-MSA}\left(\text{LN}\left(\mathbf{X}^{l-1}\right)\right) + \mathbf{X}^{l-1} \\
\mathbf{X}^{l} &= \text{MLP}\left(\text{LN}\left(\hat{\mathbf{X}}^{l}\right)\right) + \hat{\mathbf{X}}^{l} \\
\hat{\mathbf{X}}^{l+1} &= \text{SW-MSA}\left(\text{LN}\left(\mathbf{X}^{l}\right)\right) + \mathbf{X}^{l} \\
\mathbf{X}^{l+1} &= \text{MLP}\left(\text{LN}\left(\hat{\mathbf{X}}^{l+1}\right)\right) + \hat{\mathbf{X}}^{l+1}
\end{aligned} \tag{2}$$

Positional Encoding: Transformer uses positional encoding to make up for the defect that the self-attention mechanism can not capture order of sequence information. The positional encoding methods mainly include absolute positional encoding [18, 19] and relative positional encoding [20, 21]. There are two main problems with absolute positional encoding. First, it prevents the model from processing longer sequences during test than during train. Second, it causes the model lose the translation invariance because a unique positional encoding vector is added to each patch. The translation invariance plays a significant role in Generation task because we hope the networks give the same response wherever the object is in the image. Swin uses the relative positional encoding to cope with both the aforementioned issues. However, the relative positional encoding cannot provide any absolute position information, the absolute position information is also significant to the Generation task. For the field of automatic medical report generation, the absolute position information within patch and the relative position information between patches in the image are the key factors for evaluating the quality of generated reports.

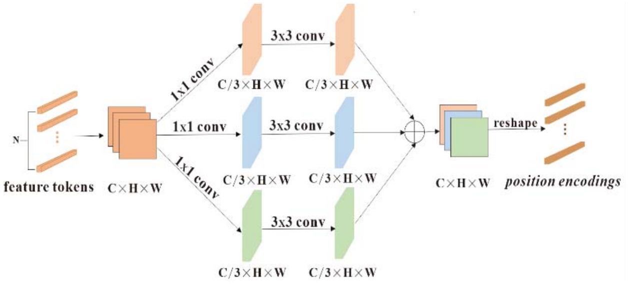


Fig. 3. Detailed illustration of adaptive positional encoding (APE). N is the number of feature tokens.

To solve the problems in above positional encoding methods, we design a new coding method named adaptive positional encoding (APE) inspired by [24], which can be seamlessly inserted into the transformer architecture. In details, APE is shown in the Fig. 3, the input of APE is the output of the last layer of transformer block in each stage. We first reshape the flattened input sequence $X \in R^{B \times N \times C}$ of last layer of transfomer block back to $X' \in R^{B \times H \times W \times C}$. Then, three 1×1 conv and 3×3 conv is repeatedly applied to the local patch in X' to produce the adaptive positional encoding E_i. Finally, the final positional encoding $X \in R^{B \times H \times W \times C}$ can be obtained by concencating E_i on the channel. APE can be efficiently implemented with a 2-D convolution with kernel k(k >= 3) and $\frac{k-1}{2}$ zero paddings. Specially, the zero paddings here are significant to make the model conscious of the absolute position, $i \in \{1,2,3\}$. The process is formulated as:

$$E = cat\left(reshape\left(Conv2d\left(reshape(X)\right)\right)\right) \quad (3)$$

where E represents the output of finally positional encoding.

B. Feature Interaction Module

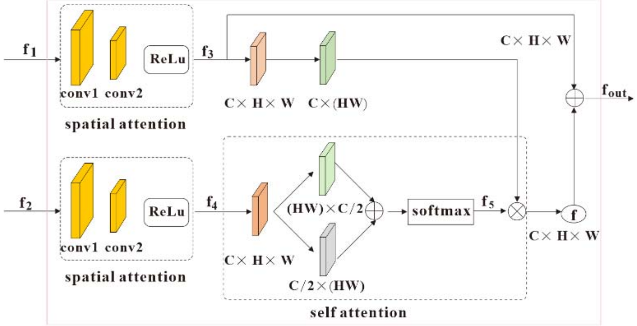


Fig. 4. The overall architecture of Feature Pruning Module. ⊗ is dot product operation, ⊕ represents concatenation operation.

In order to make full use of the complementary information between C-Stream and T-Stream, while reducing the interference of noise and background information caused by the two streams, we devise a feature interaction (FI) module according to the attention mechanism proposed by [2] instead of the direct concatenation operation. Fig. 4 shows the architecture of FI. The f1 indicates rc (C-Stream) or rt (T-Stream), and f2 represents the opposite one. We firstly feed f1 and f2 into a spatial attention to highlight the feature response, and use the ReLu activation function to obtain the outputs f3 and f4, respectively. Then we sent f4 into a self-attention model to capture a spatial weight for f3. This sub-module can make C-Stream and T-Stream information features provide useful information to each other. These feature maps f1 and f2 are fed into two attention modules, which can generate a spatial weight for the other feature map and provide complementary information. The output of FI can be defined as:

$$f_{0ut} = SP(f_i) + f \quad (4)$$

where SP means spatial attention operation. $f_i$represents rc or rt in turn. rc represents C-Stream output. rt represents T-Stream output.

C. Memory-Driven Decoder Module

In order to generate accurate and semantically coherent medical reports, we introduce a memory-driven decoder (MDD) module [5]. Different from the original decoder in the transformer, we use Memory-driven Conditional Layer Normalization and Relational Memory components. On the one hand, for any related images, they may share similar patterns in their reports and they can serve as good references to each other to aid in the generation process. As shown in Fig. 1, patterns such as “the lungs are clear bilaterally” and “no evidence of focal consolidation, or pleural effusion” always appear in the reports of similar images and are displayed concurrently. To take advantage of these characteristics, we use an additional component, i.e., relational memory, to enhance transformer to revise the patterns and facilitate the interactions between computational patterns and generation processes. On the other hand, considering that text generation is a dynamic process that is largely affected by the output of each decoding step, memory-driven decoders can generate more reliable reports. Therefore, we use a new Memory-driven Conditional Layer Normalization and combine it with Relational Memory to enhance transformer decoding. The process is formulated as:

$$y_t = f_d\left(f_{out}, MCLN\left(RM(y_1, \dots, y_{t-1})\right)\right) \quad (5)$$

where $f_d(\cdot)$ refers to the decoder.

Given the aforementioned structure, the entire generation process can be formalized as:

$$p(Y \mid I) = \prod_{t=1}^{T} p\left(y_t \mid y_1, \dots, y_{t-1}, I\right) \quad (6)$$

where $Y = \{y_1, y_2, \dots, y_t\}$ is the target text sequence.

IV. Experiments

In this section, we conduct experiments on two medical image report datasets: IU X-RAY, MIMIC-CXR.

A. Datasets

IU X-RAY [26] provides 7,470 chest X-rays with 3,955 radiology reports. In our experiments, we only utilize samples with both frontal and lateral views, and with complete finding sections in the reports.

MIMIC-CXR [27] includes 473,057 chest X-ray images and 206,563 reports from 63,478 patients.

For both datasets, we follow work [28] to exclude the samples without reports. Then we apply their conventional splits. Specifically, both IU X-RAY and MIMIC-CXR are partitioned into train/validation/test set by 7:1:2 of the entire dataset.

B. Implementation Details

In our model, the size of input images is 224×224. For C-Stream, we adopt resnet101 pretrained on Imagenet as backbone. For T-Stream, we adopt transformer encoder architecture, which embed adaptive positional encoding manner. The input images are divided into a number of patches without overlapping and the resolution of each patch is set as 7 × 7. The output dimension of each stream is 2048. For relational memory, its dimension and the number of heads of multi-head attention are set to 512 and 8, respectively, and the default memory slots is set to 3. In our experiments, we apply cross entropy loss with ADAM optimizer to train whole network, the learning rate of C-Stream and T-Stream is 5e-5, the learning rate of other parameters is set to 1e-4. We apply random cropping and flipping to the medical images to avoid overfitting problems. Note that the above hyper-parameters are obtained by evaluating the models on the validation sets of the two datasets.

C. Baseline and Evaluation Metrics

To compare with our proposed model, the following models are used as the primary baselines:

- **BASE:** This combines convolutional neural network (CNN) and vanilla Transformer, on this basis, increasing

Relational Memory and Memory-driven Conditional Layer Normalization module.

- **T-Stream:** This is a simple alternative of our proposed model where the T-stream is introduced as only branch. At the same time, adaptive positional encoding method is designed, which can be seamlessly embedded in the transformer. This baseline aims to demonstrate the effect of using T-Stream as an extra component.
- **BASE+T-Stream:** The goal of this baseline is to demonstrate the advantage of non-symmetrical sibling-stream network.

The evaluation of the models is preformed using general NLG metrics including BLUE [29], METEOR [30] and ROUGE-L [31]. BLEU and METEOR are originally proposed for machine translation evaluation. ROUGE-L is designed for measuring the quality of summaries.

D. Ablation Study

Effectiveness of Non-Symmetrical Sibling-Stream Network. In order to verify the influence of different modules in the proposed framework, we conduct ablation studies on the IU X-RAY and MIMIC-CXR dataset. Our method adopt a sibling-stream network to model global dependency and low-level spatial details, then, accurate and coherent medical reports can be generated by memory-driven decoder. Therefore, we first conduct experiments to verify the specific impact of sibling-stream on medical report generation through quantitative analysis. We conduct three sets of comparative experiments:

BASE, T-Stream, BASE + T-Stream. On NLG metrics, BASE + T-Stream outperforms the BASE and T-Stream on both datasets, Experiment result (see Table II) demonstrates the effectiveness of our sibling-stream network.

Effectiveness of Adaptive Positional Encoding. We compare the effects of three different positional encoding methods by our proposed approach, and the experimental results (see Table III) show that adaptive positional encoding (APE) outperforms absolute and relative one. We speculate the reason is that adaptive positional encoding can help transformer model global context information and avoid the lost position information caused by serialized input. Then, in the process of generating medical reports, the medical terms can be described more accurately.

Effectiveness of Feature Interaction Module. As can be seen from Table IV, increasing feature interaction (FI) module leads to performance increment by 2% on average. This demonstrates that FI outperforms direct stitching. We analyze FI facilitates information interaction of the sibling-stream network by introducing attention mechanism, which makes them complement each other and avoids noise and background information.

TABLE I. COMPARISONS OF OUR COMPLETE MODEL WITH PREVIOUS STUDIES ON THE TEST SETS OF IU X-RAY AND MIMIC-CXR ABOUT NLG. HIGHER VALUE INDICATES BETTER PERFORMANCE FOR MOST COLUMNS, WHICH PROVES THE VALIDITY OF OUR APPROACH.

DATA	MODEL	YEAR	NLG METRICS					
			BL-1	BL-2	BL-3	BL-4	MTR	RG-L
IU X-RAY	ATT2IN	2017	0.224	0.129	0.089	0.068	-	0.308
	HRGR	2018	0.438	0.298	0.208	0.151	-	0.322
	KERP	2019	0.482	**0.325**	0.226	0.162	-	0.339
	R2Gen	2020	0.470	0.304	0.219	0.165	-	0.371
	CMN	2021	0.475	0.309	0.222	0.170	-	0.375
	PPKED	2021	0.483	0.315	0.224	0.168	-	0.376
	ATformer	2021	0.484	0.313	0.225	0.173		0.379
	DeltaNet	2022	0.485	0.324	**0.238**	**0.184**	-	0.379
	OURS	2023	**0.490**	0.312	0.227	0.174	-	**0.388**
MIMIC -CXR	ATT2IN	2017	0.325	0.203	0.136	0.096	0.134	0.276
	R2Gen	2020	0.353	0.218	0.145	0.103	0.142	0.277
	CMN	2021	0.353	0.218	0.148	0.106	0.142	0.278
	PPKED	2021	0.360	0.224	0.149	0.106	0.149	0.284
	ATformer	2021	**0.378**	0.235	0.156	0.112		0.283
	CMCA	2022	0.360	0.227	0.156	0.117	0.148	0.287
	DeltaNet	2022	0.361	0.225	0.154	0.114	**0.281**	0.277
	KE	2022	0.363	0.228	0.156	0.115	**-**	0.284
	OURS	2023	0.375	**0.247**	**0.174**	**0.130**	0.148	0.315

TABLE II. THE PERFORMANCE OF ALL BASELINES AND OUR FULL MODEL ON THE TEST SETS OF IU X-RAY AND MIMIC-CXR DATASETS WITH RESPECT TO NLG METRICS. BL-N INDICATES BLEU SCORE USING UP TO N-GRAMS.

DATA	MODEL	NLG METRICS					
		BL-1	BL-2	BL-3	BL-4	MTR	RG-L
IU X-RAY	BASE	0.470	0.304	0.219	0.165	0.187	0.371
	T-Stream	0.480	**0.315**	0.225	0.167	0.191	0.384
	BASE+T-Stream	**0.490**	0.312	**0.227**	**0.174**	**0.202**	**0.388**
MIMIC-CXR	BASE	0.353	0.218	0.145	0.103	0.142	0.277
	T-Stream	0.361	0.227	0.148	0.113	0.147	0.302
	BASE+T-Stream	**0.375**	**0.247**	**0.174**	**0.130**	**0.148**	**0.315**

TABLE III. COMPARISON OF VARIOUS POSITIONAL ENCODING (PE) STRATEGIES TESTED ON IU-X RAY AD MIMIC-CXR VALIDATION SET IN TERMS OF THE BL-1 VALUE. THE ABSOLUTE POSITIONAL ENCODING HAS INFERIOR PERFORMANCE TO THE RELATIVE ONE. APE, RPE AND SIN-COS REPRESENT THE ADAPTIVE POSITIONAL ENCODING RELATIVE POSITIONAL ENCODING AND ABSOLUTE POSITIONAL ENCODING RESPECTIVELY.

DATA	MODEL	BL-1
IU X-RAY	sin-cos	0.453
	RPE	0.480
	APE	**0.490**
MIMIC -CXR	sin-cos	0.343
	RPE	0.361
	RPE	**0.375**

TABLE IV. THE EFFECT OF FEATURE INTERACTION MODULE ON BOTH IU X-RAY AND MIMIC-CXR DATASETS. OURS REPRESENTS DIRECTLY CONNECTING SIBLING-STREAM INFORMATION BY CONCATENATION OPERATION, OURS + FP EMPLOYES FEATURE INTERACTION MODULE.

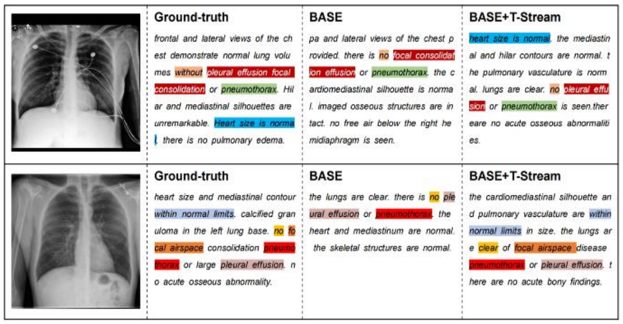
DATA	MODEL	BL-1
IU X-RAY	OURS	0.469
	OURS + FP	**0.490**
MIMIC -CXR	OURS	0.356
	OURS + FP	**0.375**

E. Comparisons with State-of-the-Art Methods

On both datasets, we compare our proposed method with several state-of-the-art image captioning models [32, 33, 28, 34, 5, 35]. Jing et al. [33] introduces a co-attention mechanism and generate the corresponding descriptions. Li et al. [34] and Li et al. [28] focus on generating standardized reports by memorizing templates from a predefined retrieval database. Chen et al. [5] proposes a simple generation-based method to generate medical reports via memory-driven transformer and show that their present approach is superior to previous models with respect to both language generation indicators and clinical evaluation. Liu et al. [35] proposes the Posterior-and-Prior Knowledge Exploring and-Distilling approach to write a corresponding report. By observing the experimental results of Tables I, II, III and 4 from different aspects, we draw some conclusions. First of all, we notice that some studies require external medical knowledge for this task, e.g., HRGR, PPKED, and our complete model outperforms them without these requirements. Second, by observing Table III, it is found that our proposed model achieves the best effect on BLUE-1. We speculate that it is because the network embedded with adaptive positional encoding has stronger abilities to capture detailed information in normal and abnormal regions. The experiments and analyses on the MIMIC-CXR and IU-X RAY datasets demonstrate the effectiveness of our method. In particular, our method not only generates semantically coherent and accurate medical reports supported with accurate medical terms, but also outperforms previous state-of-the-art models on two public datasets.

F. Visualization

Qualitative analysis of some cases with their ground-truth and generated reports from different models. Fig. 5 shows two examples of front and lateral chest X-ray images from MIMIC-CXR and such reports, where different colors on the texts distinguish different medical terms. In these cases, it is observed that our proposed approach is able to generate reports that are much closer to the ground-truth. For example, the generated report correctly describes "heart size is normal" and "no pleural effusion or pneumothorax". In addition, for the necessary medical terms in the real reports, our proposed approach covers most of them in its generated reports. This also demonstrates our method's outstanding ability to characterize image feature and generate accurate medical reports.



Ground-truth	BASE	BASE+T-Stream
frontal and lateral views of the chest demonstrate normal lung volumes without pleural effusion focal consolidation or pneumothorax. Hilar and mediastinal silhouettes are unremarkable. Heart size is normal. there is no pulmonary edema.	pa and lateral views of the chest provided. there is no focal consolidation effusion or pneumothorax. the cardiomediastinal silhouette is normal. imaged osseous structures are intact. no free air below the right hemidiaphragm is seen.	heart size is normal. the mediastinal and hilar contours are normal. the pulmonary vasculature is normal. lungs are clear. no pleural effusion or pneumothorax is seen.thereare no acute osseous abnormalities.

Ground-truth	BASE	BASE+T-Stream
heart size and mediastinal contour within normal limits. calcified granuloma in the left lung base. no focal airspace consolidation pneumothorax or large pleural effusion. no acute osseous abnormality.	the lungs are clear. there is no pleural effusion or pneumothorax. the heart and mediastinum are normal. the skeletal structures are normal.	the cardiomediastinal 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.

Fig. 5. Illustrations of reports from ground-truth, BASE and BASE+T-Stream models for IU X-RAY and MIMIC-CXR. To better distinguish the content in the reports, different colors highlight different medical terms.

V. CONCLUSION

In this paper, we propose a non-symmetrical sibling-stream network based on adaptive positional encoding to generate accurate and semantically coherent medical reports. The experiments and analyses on the IU-X RAY and MIMIC-CXR datasets verify our arguments and demonstrate the effectiveness of our method. Our proposed approach provides a new tack for medical report generation and also expand to other aspects of clinical research. In particular, our approach does not use any medical knowledge as a priori, but also outperforms previous state-of-the-art models on the two public datasets.

ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China (No.62106006), in part by the Natural Science Foundation of Anhui Higher Education Institution of China (No.KJ2020A0040), and in part by the University Synergy Innovation Program of Anhui Province, China(No.GXXT-2020-015，No.GXXT-2021-065).

REFERENCES

[1] Zhang, Y., Wang, X., Xu, Z., Yu, Q., Yuille, A., Xu, D.: When radiology report generation meets knowledge graph. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 12910–12917 (2020)

[2] Wang, Y., Lin, Z., Tian, J., Shi, Z., Zhang, Y., Fan, J., He, Z.: Confidence-guided radiology report generation. arXiv preprint arXiv:2106.10887 (2021)

[3] Li, M., Wang, F., Chang, X., Liang, X.: Auxiliary signal-guided knowledge encoder-decoder for medical report generation. arXiv preprint arXiv:2006.03744 (2020)

[4] Wang, X., Peng, Y., Lu, L., Lu, Z., Summers, R.M.: Tienet: Text-image embedding network for common thorax disease classification and reporting in chest ding network for common thorax disease classification and reporting in chest x-rays. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, pp. 9049–9058. Computer Vision Foundation / IEEE Computer Society

[5] Chen, Z., Song, Y., Chang, T.-H., Wan, X.: Generating radiology reports via memory-driven transformer. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pp. 1439–1449 (2020)

[6] Yang, X., Ye, M., You, Q., Ma, F.: Writing by memorizing: Hierarchical retrieval-based medical report generation. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (2021)

[7] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)

[8] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: 9th International Conference on Learning Representations (2021)

[9] Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 568–578 (2021)

[10] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., J´egou, H.: Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning, pp. 10347–10357 (2021)

[11] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conference on Computer Vision, pp. 213–229 (2020)

[12] Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: deformable transformers for end-to-end object detection. In: 9th International Conference on Learning Representations (2021)

[13] Cornia, M., Stefanini, M., Baraldi, L., Cucchiara, R.: Meshed-memory transformer for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10578–10587 (2020)

[14] Lu, J., Xiong, C., Parikh, D., Socher, R.: 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 (2017)

[15] Xue, Y., Xu, T., Long, L.R., Xue, Z., Antani, S., Thoma, G.R., Huang, X.: Multimodal recurrent model with attention for automated radiology report generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 457–466 (2018). Springer

[16] Yuan, J., Liao, H., Luo, R., Luo, J.: Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 721–729 (2019). Springer

[17] Nooralahzadeh, F., Perez Gonzalez, N., Frauenfelder, T., Fujimoto, K., Krauthammer, M.: Progressive transformer-based generation of radiology reports. In: Findings of the Association for Computational Linguistics: EMNLP 2021, pp. 2824–2832 (2021)

[18] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, , Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)

[19] Gehring, J., Auli, M., Grangier, D., Yarats, D., Dauphin, Y.N.: Convolutional sequence to sequence learning. In: Proceedings of the 34th International Conference on Machine Learning, pp. 1243–1252 (2017)

[20] Shaw, P., Uszkoreit, J., Vaswani, A.: Self-attention with relative position representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 464–468 (2018)

[21] Dai, Z., Yang, Z., Yang, Y., Carbonell, J.G., Le, Q.V., Salakhutdinov, R.: Transformer-xl: Attentive language models beyond a fixed-length context. In: Proceedings of the 57th Conference of the Association for Computational Linguistics, pp. 2978–2988 (2019)

[22] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. arXiv preprint arXiv:2103.14030 (2021)

[23] Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)

[24] Chu, X., Tian, Z., Zhang, B., Wang, X., Wei, X., Xia, H., Shen, C.: Conditional positional encodings for vision transformers. arXiv preprint arXiv:2102.10882 (2021)

[25] Chen, Z., Cong, R., Xu, Q., Huang, Q.: Dpanet: Depth potentiality-aware gated attention network for rgb-d salient object detection. IEEE Transactions on Image Processing (2020)

[26] Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association 23(2), 304–310 (2016)

[27] Johnson, A.E., Pollard, T.J., Greenbaum, N.R., Lungren, M.P., Deng, C.-y., Peng, Y., Lu, Z., Mark, R.G., Berkowitz, S.J., Horng, S.: Mimic-cxr, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042 (2019)

[28] Li, C.Y., Liang, X., Hu, Z., Xing, E.P.: Hybrid retrieval-generation reinforced agent for medical image report generation. In: Advances in Neural Information Processing Systems, pp. 1537–1547 (2018)

[29] Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318 (2002)

[30] Denkowski, M., Lavie, A.: Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In: Proceedings of the Sixth Workshop on Statistical Machine Translation, pp. 85–91 (2011)

[31] Lin, C.-Y.: Rouge: A package for automatic evaluation of summaries. In: Text Summarization Branches Out, pp. 74–81 (2004)

[32] Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V.: Self-critical sequence training for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7008–7024 (2017)

[33] Jing, B., Xie, P., Xing, E.: On the automatic generation of medical imaging reports. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, pp. 2577–2586 (2018)

[34] Li, C.Y., Liang, X., Hu, Z., Xing, E.P.: Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 6666–6673 (2019)

[35] Liu, F., Wu, X., Ge, S., Fan, W., Zou, Y.: Exploring and distilling posterior and prior knowledge for radiology report generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13753–13762 (2021)

[36] Yundong Zhang, Huiye Liu, Qiang Hu.: TransFuse: Fusing Transformers and CNNs for Medical Image Segmentation. MICCAI (1) 2021: 14-24

[37] Wang, H., Chen, X., Zhang, T., Xu, Z., & Li, J. (2022). CCTNet: Coupled CNN and transformer network for crop segmentation of remote sensing images. Remote Sensing, 14(9), 1956.