MMTN: Multi-Modal Memory Transformer Network for Image-Report Consistent Medical Report Generation

Yiming Cao[1,2], Lizhen Cui[1,2]*, Lei Zhang[1,2], Fuqiang Yu[1,2], Zhen Li[3], Yonghui Xu[2]*

[1]School of Software, Shandong University, Jinan, China
[2]Joint SDU-NTU Centre for Artificial Intelligence Research (C-FAIR), Shandong University, Jinan, China
[3]Department of Gastroenterology, Qilu Hospital of Shandong University, Jinan, China
{caoyiming, leizh, yfq}@mail.sdu.edu.cn, {clz,qilulizhen}@sdu.edu.cn, xu.yonghui@hotmail.com

Abstract

Automatic medical report generation is an essential task in applying artificial intelligence to the medical domain, which can lighten the workloads of doctors and promote clinical automation. The state-of-the-art approaches employ Transformer-based encoder-decoder architectures to generate reports for medical images. However, they do not fully explore the relationships between multi-modal medical data, and generate inaccurate and inconsistent reports. To address these issues, this paper proposes a Multi-modal Memory Transformer Network (MMTN) to cope with multi-modal medical data for generating image-report consistent medical reports. On the one hand, MMTN reduces the occurrence of image-report inconsistencies by designing a unique encoder to associate and memorize the relationship between medical images and medical terminologies. On the other hand, MMTN utilizes the cross-modal complementarity of the medical vision and language for the word prediction, which further enhances the accuracy of generating medical reports. Extensive experiments on three real datasets show that MMTN achieves significant effectiveness over state-of-the-art approaches on both automatic metrics and human evaluation.

Introduction

Medical image reports utilize free text to describe and explain the medical observations in images, which are mainly written by doctors based on their medical knowledge and experience. To alleviate the heavy workload of doctors, automatic report generation has become a critical task.

The state-of-the-art works in medical report generation task adopt the encoder-decoder architecture (Zhang et al. 2020; Liu et al. 2021a) to automatically generate reports for medical images. Although these works can generate textual narratives for medical images, they are still limited in fully exploiting the information from medical multi-modal data, such as the consistent mapping bewteen medical images and reports and the utilization of important medical terminology knowledge, which is demonstrated in Figure 1. Therefore, there are some issues that need to be further explored:
1) The relationships between multi-modal medical data are not fully explored. Some works (Chen et al. 2020, 2021) only leverage two types of data (*i.e.,* images and text) to generate reports, ignoring essential medical knowledge. Some works introduce medical knowledge (*e.g.,* disease tags (Li et al. 2019) and regions (Liu et al. 2021a)) to guide the report generation, without exploring the correlations between knowledge and images or texts. These works do not fully exploit medical data's multi-modal nature and relationships.
2) The generated reports show a deficiency in both precision and consistency. Most approaches (Yuan et al. 2019; You et al. 2021) directly align image visual features and report linguistic features to generate reports. The limitation of annotated correspondence between images and text results in inaccuracies and inconsistencies in the sentences generated by these methods. In addition, some essential medical terminologies in medical reports cannot be effectively generated.

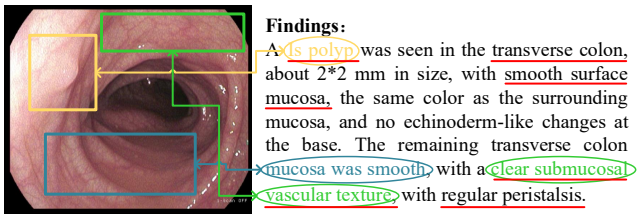

Figure 1: An example of gastroenterology report, where aligned image and report are marked in different colors and medical terminology knowledge are underlined in red.

To tackle the above limitations, in this paper, we propose a **M**ulti-modal **M**emory **T**ransformer **N**etwork (MMTN) to generate semantically coherent and consistent medical image reports. To take full advantage of the multi-modal nature of medical data, MMTN is capable of incorporating and processing multi-modal medical data, *i.e.,* medical image, terminology knowledge, and text report simultaneously, and exploring the interactions between different modalities to improve the quality of medical report generation. To make the report cover important medical terminologies, we designed the MMTN encoder to capture and memorize the relationship between medical images and medical terminologies, which can assist in guiding the transformation from image visual features to report text features with the medium of medical terminologies. Specifically, the grid module and terminology BERT module extract features from medical im-

*Corresponding Authors
Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

ages and terminologies, respectively. The memory augment module is devised to learn the relationship between two features using learnable memory matrices. Furthermore, to exploit the cross-modal complementarity of multi-modal medical features, we apply the multi-modal fusion layer on top of the MMTN decoder to adaptively learn the contribution of multi-modal visual features and linguistic features for word generation. Experimental results on three real-world medical image report datasets illustrate the effectiveness of our MMTN. The contributions are summarized as follows:

- We propose a Multi-modal Memory Transformer Network to process multi-modal medical data including medical image, terminology knowledge, and report text, and design a unique encoder to associate and memorize visual features of medical images and representations of terminologies, which assists in bridging the distance between vision and language.
- We build a multi-modal fusion layer, attached to the top of the MMTN decoder, to weigh the contribution of visual and linguistic features by exploiting the cross-modal complementarity of multi-modal medical features, and to generate an image-report consistent report.
- We experimentally evaluate MMTN using three real-world datasets. The results demonstrate that MMTN outperforms state-of-the-art methods on both automatic metrics and human evaluation, indicating that our MMTN can generate accurate medical reports.

Related Work

The existing works mainly explore the image captioning and report generation for medical domain.

Image Captioning

The task of image captioning has been studied by two main approaches: traditional methods and deep learning based methods. For traditional methods, the retrieval- (Gupta, Verma, and Jawahar 2012) and template-based (Mitchell et al. 2012) models are the most commonly adopted for caption generation. With the development of deep learning (He et al. 2016; Huang et al. 2017), the encoder-decoder structures (Shin et al. 2016) are widely used. The visual captioning models employ attention mechanisms (Rennie et al. 2017; You et al. 2016) to improve performance. In addition, extra information is adopted to assist text generation for Natural Language Processing (NLP) and image caption tasks, such as pre-trained embeddings (Zhang et al. 2019), pre-built knowledge graphs (Li et al. 2019), and pre-trained models (Devlin et al. 2019). The Transformer-based model (Cornia et al. 2020; Zhang et al. 2021) also greatly improves the performance of the task.

However, these methods cannot be directly transferred to medical report generation tasks. Medical reports do not consist of only a sentence of short text but a long paragraph consisting of normal and abnormal descriptions. The image caption methods do not cope effectively with the properties.

Medical Report Generation

Similar to image captioning, most existing works of report generation adopt the encoder-decoder paradigm to generate reports. Works (Yuan et al. 2019; You et al. 2021) fuse the image features with the medical tags or concepts predicted by Convolutional Neural Network (CNN) to generate reports. Some approaches adopted extra information (such as context (Jing, Xie, and Xing 2018) and topic representations (Li et al. 2018)) to assist report generation. Other methods append auxiliary modules to CNN-RNN architecture, such as the recurrent generation model (Xue et al. 2018), and clinical features (Zhou et al. 2021). The graph neural networks (Liang et al. 2018) are derived to the predefined abnormal graphs (Li et al. 2019) and pre-constructed graph embedding modules (Zhang et al. 2020) for report generation. Subsequently, Transformer-based approaches (Chen et al. 2020; Liu et al. 2021a; Cao et al. 2022) are proposed to solve the problem that RNN-based models cannot effectively handle dependencies between distant-location. Works (Chen et al. 2020, 2021) use memory vectors to memorize the interaction between images and reports. Besides, the contrastive model, CA (Liu et al. 2021b), captures and describes abnormal regions, and unsupervised KGAE (Liu et al. 2021c) relaxes the dependency on paired data.

However, these works did not fully explore relationships between multi-modal medical data. Our work differs from these in that we not only associate and memorize the relationship between images and terminologies, but also use the properties of multi-modal data to generate reports.

Multi-Modal Memory Transformer Network

The multi-modal memory Transformer network consists of three core components, namely the MMTN encoder, the MMTN decoder, and the multi-modal fusion layer.

The overall architecture of our MMTN is depicted in Figure 2. The MMTN encoder is in charge of processing input images and medical terminologies into the enriched features, aiming to associate and memorize the relationship between grid features and terminological features. The MMTN decoder receives the output of the encoder and the word embeddings of reports to generate semantic states. The multi-modal fusion layer conducts joint representations of multi-modal features by self-directed learning the contribution of enriched features and semantic states to generate semantically consistent medical reports.

MMTN Encoder

For the generated report to encompass important medical terminologies, the MMTN encoder is devised to associate and memorize the relationship between visual features of medical images and medical terminology representations, which assists in bridging the gap between images and reports. The MMTN encoder consists of a grid module, a terminology BERT, and a memory augment module.

Grid Module Given any medical image I, the grid module is designed to extract grid features $\mathbf{f}^g$ of I. The grid features $\mathbf{f}^g$ are extracted by a pre-trained CNN model (Huang et al. 2017). Specifically, the image I is first divided into several

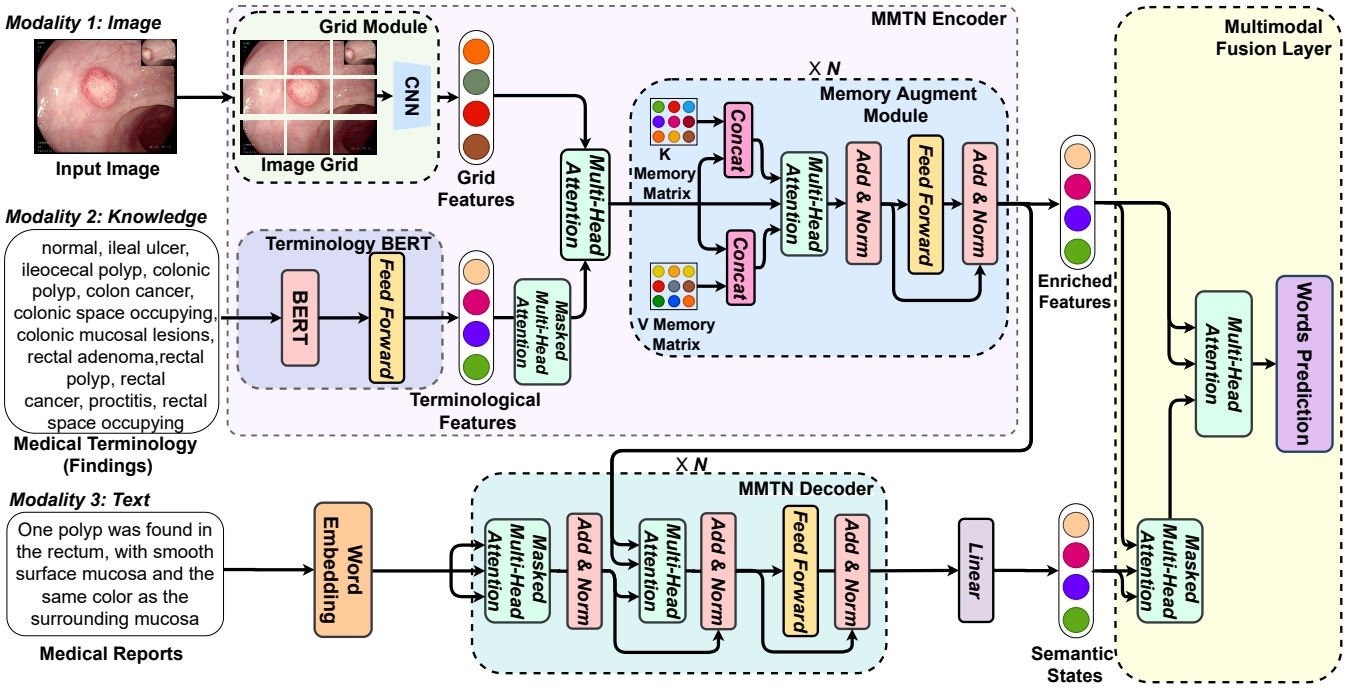

Figure 2: Overview of our proposed MMTN architecture. The input images and medical terminology knowledge are first fed into the MMTN encoder, consisting of the grid module, terminology BERT, and a stack of memory augment modules, to obtain the enriched features. A stack of MMTN decoders is in charge of generating the semantic states. The multi-modal fusion layer measures the contribution of two features to generate a medical report.

equal-sized regions, and then each grid feature $\mathbf{g}_i$ of the region is extracted separately from the last convolutional layer of CNN. Subsequently, the final grid features $\mathbf{f}^g$ are obtained by concatenating each extracted grid feature. The grid module can be expressed as:

$$\mathbf{f}^g = F_{GM}(I) = Concat[\mathbf{g}_1, \mathbf{g}_2 \ldots, \mathbf{g}_R] \quad (1)$$

where $F_{GM}(.)$ denotes the grid module, $Concat$ indicates the concatenation operation, R is the number of regions.

Terminology BERT The terminology BERT is adopted to represent the contextual information of medical terminologies related to medical reports, which helps to improve the contextual relevance of reports.

We first build two corpora of commonly used medical terminologies for gastrointestinal and thoracic diseases. For gastrointestinal diseases, we invite gastroenterologists to provide medical terminologies that often appear in reports, such as "smooth mucosa", "polypoid protrusion", and "surface erosion". In addition, the medical terminologies for thoracic diseases are automatically extracted from the "Findings" part of medical reports with the frequencies no less than three times in the corpus, such as "no pneumothorax", "biapical plural thickening", and "hyperexpanded lung".

Furthermore, we employ a BERT-based module to extract terminological features. The terminology BERT module consists of a pre-trained BERT model (Devlin et al. 2019; Zhang et al. 2021) and a feed-forward network to extract terminological features from the defined terminology corpus. The process can be formalized as:

$$\mathbf{f}^B = BERT(\mathbf{C}) \quad (2)$$

$$\mathbf{f}^t = Att_{mask}\left(FFN\left(\mathbf{f}^B\right)\right) \quad (3)$$

where $\mathbf{f}^B$ is the output of the pre-trained BERT model, $\mathbf{C}$ denotes the word sequence of the terminology corpus, Att_{mask} is the masked multi-head attention, FFN represents the fully connected feed-forward network, and $\mathbf{f}^t$ indicates the terminological features.

Memory Augment Module The memory augment module is proposed to associate and memorize the hidden correlation between medical images and terminologies. For a medical image, there are corresponding medical terminologies in the report to describe it. To exploit the characteristics, we adopt the memory augment module to represent the correlation between visual context and medical terminology features, which is beneficial to guide the report generation.

The input of the memory augment module is the joint features $\mathbf{Q}_m$ generated by grid features $\mathbf{f}^g$ and terminological features $\mathbf{f}^t$ under an attention mechanism. A set of keys and values for self-attention are employed to memorize semantic context information between medical images and terminologies. The keys and values are implemented as two learnable matrices, namely $\mathbf{Mem}_K$ and $\mathbf{Mem}_V$, which can be updated by SGD. The feature interactions in the memory augment module are computed by scaled dot-product attention. Subsequently, the output of multi-head attention is applied

to the feed-forward layer. Finally, the enriched features $\mathbf{f}^e$ are obtained by the residual connection and normalization operation layer. Formally, the process can be defined as:

$$\mathbf{Q}_m = Attention(\mathbf{Q}_j, \mathbf{K}_j, \mathbf{V}_j) \quad (4)$$

$$\mathbf{Q}_j = \mathbf{W}_{Qj} Att_{mask}(\mathbf{f}^t) \quad (5)$$

$$\mathbf{K}_j = \mathbf{W}_{Kj}\mathbf{f}^g, \mathbf{V}_j = \mathbf{W}_{Vj}\mathbf{f}^g \quad (6)$$

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = Softmax(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}})\mathbf{V} \quad (7)$$

$$\mathbf{f}^a = Attention(\mathbf{W}_{Qm}\mathbf{Q}_m, \mathbf{K}_m, \mathbf{V}_m) \quad (8)$$

$$\mathbf{K}_m = Concat[\mathbf{W}_{Km}\mathbf{Q}_m, \mathbf{Mem}_K] \quad (9)$$

$$\mathbf{V}_m = Concat[\mathbf{W}_{Vm}\mathbf{Q}_m, \mathbf{Mem}_V] \quad (10)$$

$$\mathbf{f}^e = AddNorm(FFN(AddNorm(\mathbf{f}^a))) \quad (11)$$

where $\mathbf{Q}_m$ denotes the input of memory augment module, $\mathbf{Q}_x$, $\mathbf{K}_x$ and $\mathbf{V}_x$ ($x \in \{j, m\}$) represent the query, key and value matrix, $\mathbf{W}_{Q_x}$, $\mathbf{W}_{K_x}$ and $\mathbf{W}_{V_x}$ ($x \in \{j, m\}$) are learnable weight matrices, d indicates a scaling factor, $\mathbf{f}^a$ is the output of the multi-head attention layer in this module, and $AddNorm$ is composition of a residual connection and of a normalization layer.

MMTN Decoder

The MMTN decoder is adopted to generate the semantic states based on previously generated words and the enriched features. The text sequence features $\mathbf{f}^w$ of medical reports are extracted by word embedding layer, and then regarded as the input of the first layer of the MMTN decoder. The second layer is a multi-head attention operation with K and V matrices from the enriched features $\mathbf{f}^e$ of MMTN encoder. The MMTN decoder can be formalized as:

$$\mathbf{f}^s = AddNorm(Att_{mask}(\mathbf{f}^w)) \quad (12)$$

$$\mathbf{f}^m = AddNorm(Attention(\mathbf{W}_{Qh}\mathbf{f}^s, \mathbf{W}_{Kh}\mathbf{f}^e, \mathbf{W}_{Vh}\mathbf{f}^e)) \quad (13)$$

$$\mathbf{f}^h = AddNorm(FFN(\mathbf{f}^m)) \quad (14)$$

where $\mathbf{f}^s$ and $\mathbf{f}^m$ denote the intermediate outputs of the decoder, and $\mathbf{f}^h$ is the semantic states.

Multi-Modal Fusion Layer

Two modal features are obtained by modules mentioned above, namely the enriched features $\mathbf{f}^e$ and the semantic states $\mathbf{f}^h$. To obtain a semantically coherent medical report, we designed a multi-modal fusion layer, attached to the upper layer of the MMTN decoder. The module combines the feature information of two modalities to calculate the contribution of visual features and linguistic features to each generated sequence. The multi-modal fusion layer can be defined as follows:

$$\mathbf{Q}_o = \mathbf{W}_{Qo} Att_{mask}(\mathbf{W}_{Qa}\mathbf{f}^e, \mathbf{W}_{Ka}\mathbf{f}^h, \mathbf{W}_{Va}\mathbf{f}^h) \quad (15)$$

$$\mathbf{K}_o = \mathbf{W}_{Ko}\mathbf{f}^e, \mathbf{V}_o = \mathbf{W}_{Vo}\mathbf{f}^e \quad (16)$$

$$Output = Attention(\mathbf{Q}_o, \mathbf{K}_o, \mathbf{V}_o)\mathbf{W}_A \quad (17)$$

where $\mathbf{Q}_o$, $\mathbf{K}_o$ and $\mathbf{V}_o$ are the query, key and value matrix of the multi-head attention, $Output$ denotes the result of multi-head attention for the generated reports, $\mathbf{W}_{Qx}$, $\mathbf{W}_{Kx}$, $\mathbf{W}_{Vx}$ ($x \in \{o, a\}$), and $\mathbf{W}_A$ are learnable weight matrices.

Training

For each training sample $(\mathbf{I}, r)$, where $\mathbf{I}$ is a group of images and r is the corresponding medical report composed of ground truth sequences, the loss $\mathcal{L}$ of report generation is minimized by the cross-entropy loss:

$$\mathcal{L}(\theta) = -\sum_{i=1}^{M} \log(p_\theta(s_i|s_{1:i-1})) \quad (18)$$

where θ is the parameters of our MMTN model, $s_{1:M}$ represents ground truth sequences of the report r.

Experiment

In this section, we first describe the experimental settings. Then, we demonstrate the experimental results, including performance comparisons, case studies, and ablation studies to evaluate the performance of MMTN against state-of-the-art baseline methods.

Experimental Settings

Dataset We conduct experiments on three datasets.

1) ***Gastrointestinal Endoscope image dataset (GE)*** is a private dataset contains white light images and their Chinese reports from the Department of Gastroenterology. The dataset consists of 3,168 patients. Each patient has multiple gastrointestinal endoscope images from different perspectives with their corresponding medical reports. We obtain 15,345 images and 3,069 reports collected from the dataset by selecting patients with 5 images. We collect 126 medical terminologies from gastroenterologists, including 89 abnormal findings and 37 normal findings.

2) ***IU-CX*** *(Demner-Fushman et al. 2016)* is a public chest X-ray dataset. We select 2,896 radiology reports with frontal and lateral view images from the original dataset. We extract 97 medical terminologies from the <Abstract> field, including 80 abnormal and 17 normal findings.

3) ***MIMIC-CXR*** *(Johnson 2019)* is the largest public chest X-ray dataset including 473,057 images and 206,563 reports. We adopt the same criterion with IU-CX to select samples, which results in 142,772 images and 71,386 reports. The medical terminologies are the same as IU-CX.

Parameter Settings The method is implemented in Pytorch 1.7.1 based on Python 3.8.5 and trained on a server with an Intel Core i9-10900K CPU, and an Nvidia RTX 3090 GPU. We randomly split both datasets into 7:1:2 training:validation:testing data to train and evaluate our method. A pre-trained DenseNet-121 is adopted to extract grid features, with 7×7 grid size. The Chinese word segmentation module of Jieba (Jieba 2019) is employed for processing the reports of GE. The number of heads is set to 8, the layer number N of Transformer is 3, and the number of memory vectors is 40 rows. If not specifically specified, the hidden dimension of MMTN is 512. The dropout probability is 0.1. The ADAM optimizer with a batch size of 32 and a learning rate of 1e-5 is employed to minimize the loss function.

Dataset	Architecture	Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	ROUGE-L
GE	CNN-RNN -based	SaT	0.643	0.552	0.506	0.414	0.557	0.613
		AAtt	0.649	0.549	0.491	0.419	0.579	0.617
		CoAtt	0.774	0.654	0.618	0.575	0.674	0.748
		RGKG	0.752	0.676	0.609	0.554	0.684	0.726
	Transformer -based	Transformer	0.689	0.572	0.584	0.521	0.604	0.691
		R2GEN	0.779	0.677	0.619	0.574	0.679	0.736
		PPKED	0.791	0.684	0.624	0.579	0.691	**0.749**
		CMN	0.782	0.679	0.621	0.572	0.686	0.742
		MMTN(ours)	**0.799**	**0.692**	**0.634**	**0.589**	**0.703**	0.748
IU-CX	CNN-RNN -based	SaT	0.216	0.124	0.087	0.066	0.294	0.307
		AAtt	0.220	0.127	0.089	0.068	0.295	0.308
		CoAtt	0.455	0.288	0.205	0.154	0.277	0.369
		HRGRA	0.438	0.298	0.208	0.151	0.343	0.322
		KER	0.455	0.288	0.205	0.154	0.277	0.369
		RGKG	0.441	0.291	0.203	0.147	0.304	0.367
	Transformer -based	Transformer	0.396	0.254	0.179	0.135	-	0.342
		R2GEN	0.470	0.304	0.219	0.165	-	0.371
		PPKED	0.483	0.315	0.224	0.168	0.351	0.376
		CMN	0.475	0.309	0.222	0.170	-	0.375
		AlignTransformer	0.484	0.313	0.225	0.173	-	**0.379**
		MMTN(ours)	**0.486**	**0.321**	**0.232**	**0.175**	**0.361**	0.375
Dataset	**Architecture**	**Methods**	**BLEU-1**	**BLEU-2**	**BLEU-3**	**BLEU-4**	**METEOR**	**ROUGE-L**
MIMIC-CXR	CNN-RNN -based	SaT	0.299	0.184	0.121	0.084	0.124	0.263
		AAtt	0.299	0.185	0.124	0.088	0.118	0.266
	Transformer -based	Transformer	0.314	0.192	0.127	0.090	0.125	0.265
		R2GEN	0.353	0.218	0.145	0.103	0.142	0.277
		PPKED	0.360	0.224	0.149	0.106	0.149	**0.284**
		CMN	0.353	0.218	0.148	0.106	0.142	0.278
		AlignTransformer	0.378	0.235	0.156	0.112	0.158	0.283
		MMTN(ours)	**0.379**	**0.238**	**0.159**	**0.116**	**0.161**	0.283

Table 1: Comparison of baselines and MMTN on automatic metrics on the three datasets.

Baselines We compare our MMTN to the following state-of-the-art approaches. The CNN-RNN-based methods include SaT (Vinyals et al. 2015), AAtt (Lu et al. 2017), CoAtt (Jing, Xie, and Xing 2018), and RGKG (Zhang et al. 2020). The Transformer-based methods are Transformer (Chen et al. 2020), R2GEN (Chen et al. 2020), PPKED (Liu et al. 2021a), CMN (Chen et al. 2021), and AlignTransformer (You et al. 2021). For the IU-CX dataset, we also compare with HRGRA (Li et al. 2018) and KER (Li et al. 2019) that utilize template retrieval method for thoracic diseases, and the templates are not defined in GE and MIMIC-CXR dataset.

Evaluation Metrics We employ both automatic metrics and human evaluation to evaluate the performance for the medical report generation. 1) ***Automatic Metrics***: BLEU (unigram to 4-gram) (Papineni et al. 2002), ROUGE-L (Lin 2004), METEOR (Banerjee and Lavie 2005), and CIDEr (Vedantam, Zitnick, and Parikh 2015). 2) ***Human Evaluation***: For the samples in GE, we randomly select 50 samples and invite gastroenterologists and graduate students who collaborate with us as experts to evaluate the reports generated by baseline methods and our MMTN. Each sample is given the ground-truth report, and experts are asked to select the most consistent report among those generated by the different methods. Evaluation metrics include the report completeness, the correctness of generated abnormality findings, and contextual coherence. We collect results from 10 experts and calculate the ratio of the number of times that each model is selected to the number of total evaluations as the human evaluation score of each model.

Results on Report Generation

Automatic Evaluation We compare MMTN with baseline methods on three datasets for the report generation task, with all performances on automatic metrics shown in Table 1. It is highlighted that the **best** and second best results. Our MMTN is superior to all baseline models on BLEU-n and CIDEr (or METEOR) scores on three datasets, demonstrating the effectiveness and accuracy of MMTN in generating medical reports. MMTN is second only to PPKED and AlignTransformer on ROUGE-L. PPKED incorporates additional semantic information and abnormality graph (*i.e.,* abnormal regions and observation graph) into the generation model, which guides it to learn the most common subsequence of ground truth reports. AlignTransformer introduces additional disease label predictions to guide the generation of abnormality descriptions and therefore achieves the best performance on IU-CX. Our MMTN also achieves a competitive performance on ROUGE-L compared to the above two methods. The results on automatic metrics demonstrate that our MMTN is capable of generat-

Method	SaT	AAtt	CoAtt	RGKG	Transformer	R2GEN	PPKED	CMN	MMTN
Human Evaluation Score	0.018	0.020	0.062	0.122	0.054	0.132	0.192	0.152	**0.248**

Table 2: The results of our MMTN and baselines on human evaluation scores.

Metrics	AlignTrans	MMTN	t	p
BLEU-1	0.378	0.379	-4.950	0.008**
BLEU-2	0.235	0.238	-6.124	0.004**
BLEU-3	0.156	0.159	-3.674	0.021*
BLEU-4	0.112	0.116	-4.899	0.008**
METEOR	0.158	0.161	-3.598	0.024*
ROUGE-L	0.283	0.283	0.000	1.000

Table 3: Results of t-test analysis (*: $p < 0.05$, **: $p < 0.01$)

ing accurate and coherent reports.

In addition, we obtain some observations by comparing methods with different architectures. First, models guided by medical knowledge (*i.e.,* HRGRA, KER, RGKG, PPKED, and our MMTN) obtain higher or equivalent automatic metrics scores. This observation validates that knowledge is essential to guide the transformation from visual features to linguistic features in the medical domain. Second, compared with the vanilla CNN-RNN structure (*i.e.,* SaT), the vanilla Transformer (*i.e.,* Transformer) works slightly better. Consistent with the performance, most Transformer-based models outperform CNN-RNN-based models on automatic evaluation metrics, indicating that self-attention plays a positive role in the transformation of multi-modal features. Third, compared to models using the co-attention mechanism, approaches equipped with memory modules (*i.e.,* R2GEN, CMN, and our MMTN) exhibit better performance. One possible explanation is that using memory modules enables visual and linguistic features to be transformed in a single identical space. Our MMTN outperforms R2GEN and CMN in most metrics, illustrating that associating visual features with medical terminologies representations facilitates report generation. Last, the CoAtt, HRGRA, and PPKED adopt extra semantic information (*e.g.,* medical tags, report templates, and abnormal graphs). The three methods also achieve good outcomes on specific metrics, which shows that additional information is helpful for performance improvement. However, our MMTN still achieves state-of-the-art performance without using such information.

Human Evaluation To evaluate the clinical readability of the generated report, we invite three digestive gastroenterologists and seven graduate students to evaluate the reports generated by MMTN and baseline methods. Given random 50 samples of GE[1], we ask each expert to select one report that is most consistent with the ground truth descriptions for each sample. The human evaluation score for each method is the proportion of times the method is selected by experts out of the total number of evaluations. For example, MMTN is selected 124 times by experts as the report closest to the ground truth, so its human evaluation score is 124 / 500 = 0.248. The human evaluation results are presented in Table 2. The results show that the MMTN is better than baseline methods in clinical practice, demonstrating MMTN's capability of generating accurate and reliable reports.

Significant Tests To verify whether there are significant differences between our MMTN and state-of-the-art models, we conduct a t-test on automatic metrics. Due to the page limitation, only results on MIMIC-CXR with minimal improvement compared to the strongest baseline (*i.e.,* AlignTransformer) are presented. As shown in Table 3, the samples show significant differences on BLEU-1-4 and METEOR, indicating that the improvement of MMTN is significant compared to baseline methods, and the comparison results rule out the possibility that the advantage of our algorithm is the result of sampling difference.

Qualitative Analysis To further investigate the effectiveness of our MMTN, we conduct qualitative analysis on three datasets with their ground-truth and generated reports. We randomly select a sample from each dataset to perform a case study, and visualization results are shown in Figure 3. The first row is the sample from GE (note that gastroenterologists translate the ground-truth and generated reports of GE from Chinese to English), and the middle and last row represent the sample from IU-CX and MIMIC-CXR, respectively. It can be observed that MMTN is capable of generating reports consistent with the ground truth. In GE sample, the generated report accurately reports the locations (*i.e.,* ascending colon) and types (*i.e.,* polyp) of lesions. Similarly, in IU-CX and MIMIC-CXR samples, MMTN also accurately describes most types of lesions, such as opacities, cavitary lesion, and hyperinflation. In addition, MMTN also generates the descriptions for normal regions, such as "smooth mucosa", "No pleural effusion", and "No focal consolidation". Normal descriptions generation facilitates the coherence and completeness of the report. It is worth noting that the reports generated by MMTN cover almost all of common medical terminologies.

To further investigate how the MMTN associates visual information of images and representations of medical terminologies, we visualize image-text attention mappings from the multi-head attention of the decoder. Figure 3 shows intermediate image-text correspondences for several medical terminologies between visual features and word embeddings. It is observed that MMTN correctly aligns regions in images with indicated terminologies. Taking the first case in Figure 3 as an example, our MMTN can correctly identify diseases, *i.e.,* "hemispheric polyp", and can also indicate medical terminologies about the position and trait, such as "ascending colon", "smooth mucosa" and "vascular texture". This observation suggests that our model not only generates

[1]The human evaluation did not evaluate the IU-CX and MIMIC-CXR datasets because we did not have access to results provided by professional radiologists.

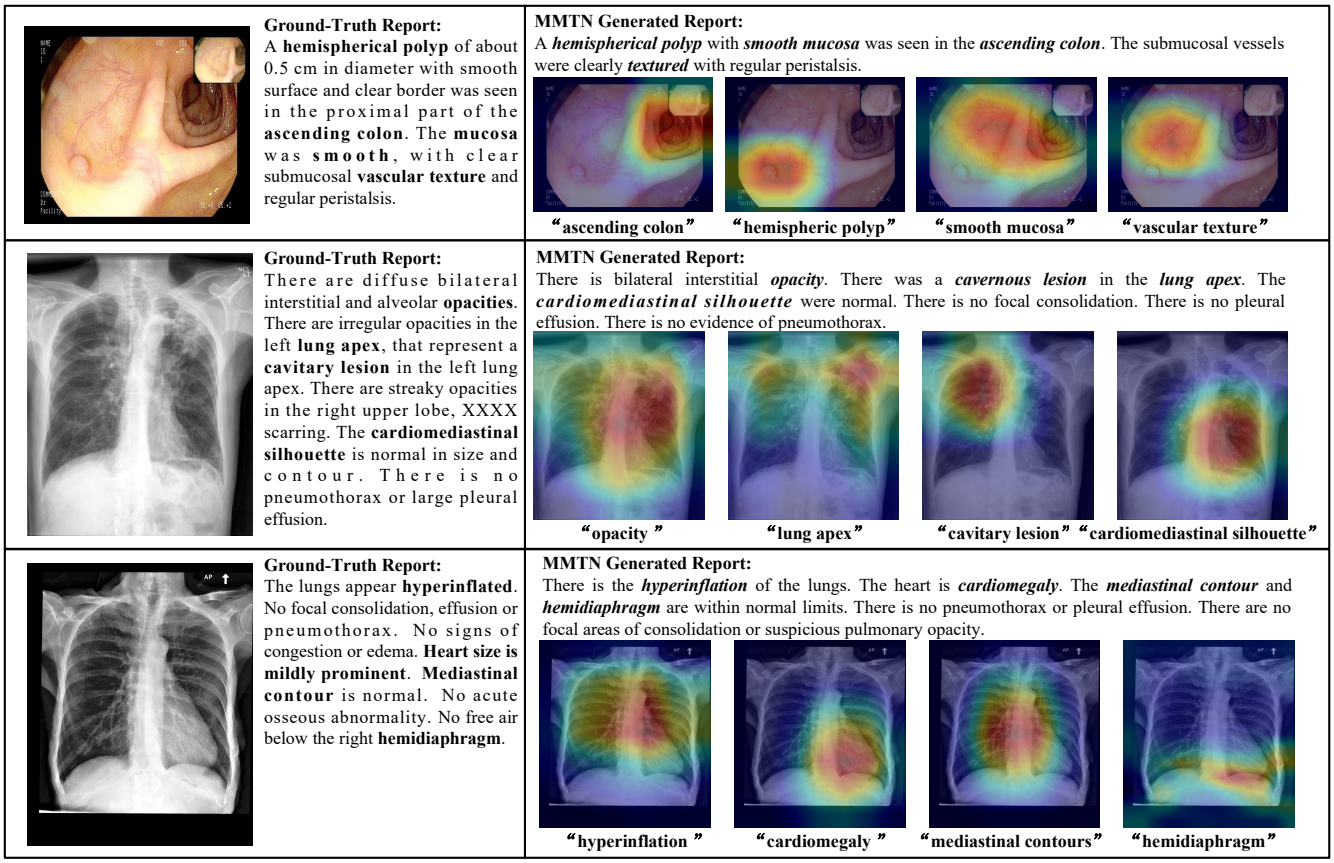

Figure 3: Visualizations of image-text attention mappings on GE (the first row), IU-CX (the middle row), and MIMIC-CXR (the last row). The left part is the image and its ground-truth report, and the right part is the MMTN generated reports and the mappings of image region and medical terminologies. Colors from blue to red represent the weights from low to high.

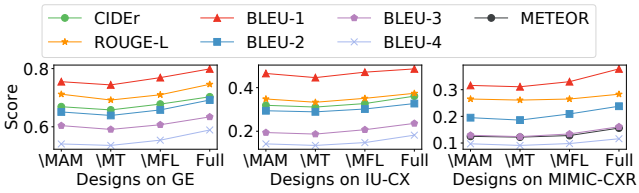

Figure 4: Ablation study for different designs.

coherent medical reports but also enhances the alignment between the images and the generated texts.

Ablation Studies

Effect of components. We conduct ablation studies on the three datasets to investigate the effectiveness of each module of MMTN. Specifically, \MAM excludes the memory augment module from MMTN, \MT does not consider medical terminologies and only utilizes the grid feature as the output of the MMTN encoder, and \MFL drops the multi-modal fusion layer. As shown in Figure 4, MMTN\MT has the worst performance, revealing that introducing medical terminologies can effectively improve report generation accuracy. On the other hand, the performance of MMTN\MAM is poor, which demonstrates that aligning and memorizing the relationship between images and terminologies is indeed helpful to bridging the distance between visual and linguistic features. The performance of MMTN\MFL is similar to that of MMTN\MAM, indicating the multi-modal fusion layer plays a certain role in improving performance. These results suggest that the modules mentioned above are efficient for the report generation task.

Conclusion

In this paper, we propose a multi-modal memory Transformer network to address multi-modal medical data, including image, text report, and terminology knowledge to improve the quality of medical report generation. To cover important medical terminologies in the generated reports, the MMTN encoder is designed to align and memorize the relationship between visual and terminological features. Further, we employ the multi-modal fusion layer to calculate the contribution of vision and language features to the report. Extensive experiments on three real world datasets demonstrate that our proposed MMTN achieves superior performance than mainstream approaches.

Acknowledgments

This work is partially supported by National Key R&D Program of China No.2021YFF0900800; NSFC No.62202279; Shandong Provincial Key Research and Development Program (Major Scientific and Technological Innovation Project) (No. 2021CXGC010506 and NO.2021CXGC010108); the State Scholarship Fund by the China Scholarship Council (CSC).

References

Banerjee, S.; and Lavie, A. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*, 65–72. Association for Computational Linguistics.

Cao, Y.; Cui, L.; Yu, F.; Zhang, L.; Li, Z.; Liu, N.; and Xu, Y. 2022. KdTNet: Medical Image Report Generation via Knowledge-Driven Transformer. In *Database Systems for Advanced Applications - 27th International Conference, DASFAA 2022*, volume 13247 of *Lecture Notes in Computer Science*, 117–132. Springer.

Chen, Z.; Shen, Y.; Song, Y.; and Wan, X. 2021. Cross-modal Memory Networks for Radiology 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, ACL/IJCNLP 2021*, 5904–5914. Association for Computational Linguistics.

Chen, Z.; Song, Y.; Chang, T.; and Wan, X. 2020. Generating Radiology Reports via Memory-driven Transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020*, 1439–1449. Association for Computational Linguistics.

Cornia, M.; Stefanini, M.; Baraldi, L.; and Cucchiara, R. 2020. Meshed-Memory Transformer for Image Captioning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020*, 10575–10584. Computer Vision Foundation / IEEE.

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

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*, 4171–4186. Association for Computational Linguistics.

Gupta, A.; Verma, Y.; and Jawahar, C. V. 2012. Choosing Linguistics over Vision to Describe Images. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*. AAAI Press.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, 770–778. IEEE Computer Society.

Huang, G.; Liu, Z.; van der Maaten, L.; and Weinberger, K. Q. 2017. Densely Connected Convolutional Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2261–2269. IEEE Computer Society.

Jieba. 2019. "Jieba" Chinese text segmentation: built to be the best Python Chinese word segmentation module. https://github.com/fxsjy/jieba. Accessed: 2023-03-22.

Jing, B.; Xie, P.; and Xing, E. P. 2018. On the Automatic Generation of Medical Imaging Reports. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018*, 2577–2586. Association for Computational Linguistics.

Johnson, A. E. W. 2019. MIMIC-CXR: A large publicly available database of labeled chest radiographs. *CoRR*, abs/1901.07042.

Li, C. Y.; Liang, X.; Hu, Z.; and Xing, E. P. 2019. Knowledge-Driven Encode, Retrieve, Paraphrase for Medical Image Report Generation. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019*, 6666–6673. AAAI Press.

Li, Y.; Liang, X.; Hu, Z.; and Xing, E. P. 2018. Hybrid Retrieval-Generation Reinforced Agent for Medical Image Report Generation. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018*, 1537–1547.

Liang, X.; Hu, Z.; Zhang, H.; Lin, L.; and Xing, E. P. 2018. Symbolic Graph Reasoning Meets Convolutions. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018*, 1858–1868.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Liu, F.; Wu, X.; Ge, S.; Fan, W.; and Zou, Y. 2021a. Exploring and Distilling Posterior and Prior Knowledge for Radiology Report Generation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021*, 13753–13762. Computer Vision Foundation / IEEE.

Liu, F.; Yin, C.; Wu, X.; Ge, S.; Zhang, P.; and Sun, X. 2021b. Contrastive Attention for Automatic Chest X-ray Report Generation. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, 269–280. Association for Computational Linguistics.

Liu, F.; You, C.; Wu, X.; Ge, S.; Wang, S.; and Sun, X. 2021c. Auto-Encoding Knowledge Graph for Unsupervised Medical Report Generation. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021*, 16266–16279.

Lu, J.; Xiong, C.; Parikh, D.; and Socher, R. 2017. Knowing When to Look: Adaptive Attention via a Visual Sentinel

for Image Captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 3242–3250. IEEE Computer Society.

Mitchell, M.; Dodge, J.; Goyal, A.; Yamaguchi, K.; Stratos, K.; Han, X.; Mensch, A. C.; Berg, A. C.; Berg, T. L.; and III, H. D. 2012. Midge: Generating Image Descriptions From Computer Vision Detections. In *EACL 2012, 13th Conference of the European Chapter of the Association for Computational Linguistics*, 747–756. The Association for Computer Linguistics.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. ACL.

Rennie, S. J.; Marcheret, E.; Mroueh, Y.; Ross, J.; and Goel, V. 2017. Self-Critical Sequence Training for Image Captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 1179–1195. IEEE Computer Society.

Shin, H.; Roberts, K.; Lu, L.; Demner-Fushman, D.; Yao, J.; and Summers, R. M. 2016. Learning to Read Chest X-Rays: Recurrent Neural Cascade Model for Automated Image Annotation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, 2497–2506. IEEE Computer Society.

Vedantam, R.; Zitnick, C. L.; and Parikh, D. 2015. CIDEr: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*, 4566–4575. IEEE Computer Society.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*, 3156–3164. IEEE Computer Society.

Xue, Y.; Xu, T.; Long, L. R.; Xue, Z.; Antani, S. K.; Thoma, G. R.; and Huang, X. 2018. Multimodal Recurrent Model with Attention for Automated Radiology Report Generation. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2018*, volume 11070 of *Lecture Notes in Computer Science*, 457–466. Springer.

You, D.; Liu, F.; Ge, S.; Xie, X.; Zhang, J.; and Wu, X. 2021. AlignTransformer: Hierarchical Alignment of Visual Regions and Disease Tags for Medical Report Generation. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2021*, volume 12903 of *Lecture Notes in Computer Science*, 72–82.

You, Q.; Jin, H.; Wang, Z.; Fang, C.; and Luo, J. 2016. Image Captioning with Semantic Attention. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 4651–4659. IEEE Computer Society.

Yuan, J.; Liao, H.; Luo, R.; and Luo, J. 2019. Automatic Radiology Report Generation Based on Multi-view Image Fusion and Medical Concept Enrichment. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2019*, volume 11769 of *Lecture Notes in Computer Science*, 721–729.

Zhang, H.; Bai, J.; Song, Y.; Xu, K.; Yu, C.; Song, Y.; Ng, W.; and Yu, D. 2019. Multiplex Word Embeddings for Selectional Preference Acquisition. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, 5246–5255. Association for Computational Linguistics.

Zhang, X.; Sun, X.; Luo, Y.; Ji, J.; Zhou, Y.; Wu, Y.; Huang, F.; and Ji, R. 2021. RSTNet: Captioning With Adaptive Attention on Visual and Non-Visual Words. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021*, 15465–15474. Computer Vision Foundation / IEEE.

Zhang, Y.; Wang, X.; Xu, Z.; Yu, Q.; Yuille, A. L.; and Xu, D. 2020. When Radiology Report Generation Meets Knowledge Graph. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*, 12910–12917.

Zhou, Y.; Huang, L.; Zhou, T.; Fu, H.; and Shao, L. 2021. Visual-Textual Attentive Semantic Consistency for Medical Report Generation. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021*, 3965–3974. IEEE.