Automatic Radiology Reports Generation via Memory Alignment Network

Hongyu Shen[1], **Mingtao Pei**[1], **Juncai Liu**[2], **Zhaoxing Tian**[3]
[1] Beijing Institute of Technology
[2] Shandong University of Science and Technology
[3] Beijing Jishuitan Hospital
{hyshen,peimt}@bit.edu.cn, 202282080086@sdust.edu.cn, tjtzx@126.com

Abstract

The automatic generation of radiology reports is of great significance, which can reduce the workload of doctors and improve the accuracy and reliability of medical diagnosis and treatment, and has attracted wide attention in recent years. Cross-modal mapping between images and text, a key component of generating high-quality reports, is challenging due to the lack of corresponding annotations. Despite its importance, previous studies have often overlooked it or lacked adequate designs for this crucial component. In this paper, we propose a method with memory alignment embedding to assist the model in aligning visual and textual features to generate a coherent and informative report. Specifically, we first get the memory alignment embedding by querying the memory matrix, where the query is derived from a combination of the visual features and their corresponding positional embeddings. Then the alignment between the visual and textual features can be guided by the memory alignment embedding during the generation process. The comparison experiments with other alignment methods show that the proposed alignment method is less costly and more effective. The proposed approach achieves better performance than state-of-the-art approaches on two public datasets IU X-Ray and MIMIC-CXR, which further demonstrates the effectiveness of the proposed alignment method.

Introduction

In medical practice, analyzing radiographic images and writing corresponding radiology reports is important. Automatically generating radiology report based on medical images can not only alleviate the burden on doctors, but also help those who do not have much experience, even playing a role in future home medical care. Up to now, there have been many methods for generating reports in this field (Jing, Xie, and Xing 2018; Chen et al. 2020, 2021; Yan and Pei 2022; Yang et al. 2023). Cross-modal mapping between images and texts plays an important role in generating high-quality reports, but due to lack of annotations correspondence between images and texts, it is hard to directly align them. As an alternative, some methods choose to align the image features with the predicted disease labels[1] (Jing, Xie, and Xing 2018; You et al. 2021). Although these direct alignment methods are intuitive, effectiveness is limited not only by the classification accuracy, but also by the limited information that disease labels can provide. Later, Chen et al. propose to use the a Cross-modal Memory Matrix (CMM) as an alignment medium to align the image and text features. As shown in Figure 1 (a), they first extract visual features from images and map words to word embeddings, then align textual and visual features by re-mapping them to a same vector space constructed by the Memory Matrix (M). The M is designed to store cross-modal information to assist the alignment. However they couple the representation and alignment process together, which inevitably brings extra work to M. As a result, the alignment module is unable to take advantage of the knowledge learned by the generator and it is also more difficult to adapt to the generator, which may reduce the alignment effect.

To address the above-mentioned problems, we propose the Memory Alignment (MA) module, a simpler alignment method. As shown in Figure 1 (b), first we embed the visual features into the same vector space of the word embedding to get the visual embeddings, then we bring the visual and word embedding closer by the memory alignment embedding got from the MA module. In our approach, M only needs to focus on how to align the visual features with the textual features, without to learn how to represent them. Another advantage is that the memory alignment embedding only needs to be calculated once during the whole inference process, while the CMN needs to remap each newly generated word. Notably, we include the positional embeddings of the visual features as part of the MA module input to provide some cross-modal information. These inputs may act like some common sense for the alignment module, such as what the visual feature here usually to represent. We argue this may improve the robustness of the alignment model, and can make our alignment more effective and work better with the report generator.

Specially, we first extract visual features by a pretrained CNN, then embed them into the visual embeddings. Afterward, we combine the visual embeddings with their respective positional embeddings as a query, and utilize a memory matrix storing cross-modal knowledge as key and value, to get the memory alignment embeddings through the MA module. In the end, we combine these input embeddings

[1]These disease labels were predicted by the CheXpert (Irvin et al. 2019), which is designed for the MIMIC-CXR.

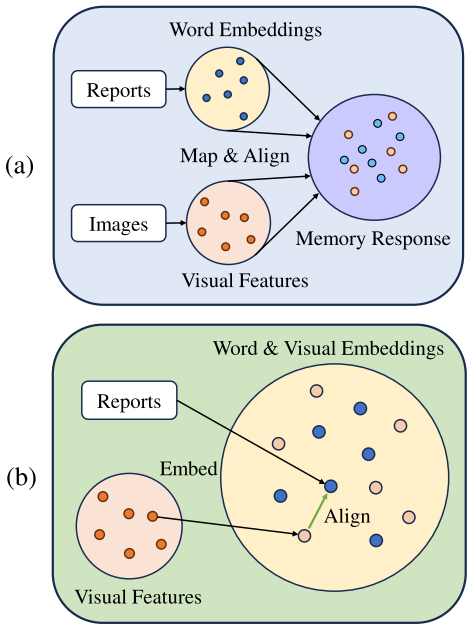

Figure 1: Comparison of two alignment methods. (a) Cross-Modal Memory. (b) Our Memory Alignment method, where the green arrow "Align" represents the memory alignment embedding.

and feed them into the report generator, which consist of a Bert-based encoder and a generation head, to generate the entire article through a chain loop. We adopt both natural language generation metrics and clinical efficacy metrics to evaluate our model on two public datasets, IU X-Ray (Demner-Fushman et al. 2015) and MIMIC-CXR (Johnson et al. 2019), and achieve better performance than state-of-the-art methods, which shows the effectiveness of the proposed memory alignment embedding. We also conducted comparison experiments with other alignment methods, and the results demonstrate that our method is less costly and more effective. The contributions of this paper can be summarized as follows:

- We propose a memory-based alignment method, which can use the cross-modal knowledge learned by itself to assist in generating a coherent and informative reports.
- Comparison experiments with other alignment methods demonstrate that our alignment method achieves superior effectiveness at a lower cost.
- Our model achieve better performance than state-of-the-art radiology reports generation methods, which further proves the effectiveness of the proposed Memory Alignment module.

Related Work

Image Captioning

Image captioning is to automatically generate descriptive text based on the content of the image (Vinyals et al. 2015; Xu et al. 2015; Aneja, Deshpande, and Schwing 2018; Tran, Mathews, and Xie 2020; Chen, Deng, and Wu 2022), which is similar to the Radiology Report Generation. However, one major difference between them is that medical reports usually require many sentences to analyze the radiology image and therefore the length of the medical report is often much longer than the image caption.

Radiology Report Generation

There have been many works on Radiology report generation in recent years, and they can be broadly classified into two main categories: generation-based methods and retrieval-based methods.

Retrieval-based Methods. Retrieval-based methods learn to obtain candidate templates from radiographs and refine the final reports with enriched details. (Li et al. 2018) establish a template database based on language as a prior, and train policy agents to determine whether to retrieve from the template database. (Li et al. 2019) employ a Graph Transformer that dynamically transforms high-level semantics between graph-structured data of multiple domains, which reconciles traditional knowledge- and retrieval-based methods with modern learning-based methods for accurate and robust medical report generation. (Liu et al. 2021b) propose a new knowledge driven encoding, retrieval, and interpretation (KERP) method, convert the visual features of the image into anomaly maps, then retrieve the template and further expand and rewrite it. These methods explore the focus more on how to add additional prior knowledge into the model to better generate, while our article focuses more on how to align the image and text features to improve the quality of the generated radiological reports.

Generation-based Methods. Generation-based methods generally adopt a sequence to sequence generation model. (Jing, Xie, and Xing 2018) propose a multi-task learning framework with a co-attention mechanism. (Xue et al. 2018) employ a multimodal model combines images and encoding of the previous generated sentence to construct attentional inputs to guide the generation of the next sentence and henceforth maintain coherence between the generated sentences (Xue et al. 2018). (Chen et al. 2020) use the Transformer architecture to generate better reports by adding a memory module that allows the model to focus on the similarities of each report . (You et al. 2021) propose to reduce the data bias by aligning visual features and disease labels at multi granularity level, and to generate reports using a Multi-grained Transformer. (Liu et al. 2021a) explore prior knowledge in reports and posterior knowledge in radiography, improving the performance of report generation. (Yang et al. 2023) propose a knowledge base updating mechanism to learn and store medical knowledge automatically, and utilize textual embedding to guide the learning of the visual feature space.

Many of the above mentioned methods contain algorithms or modules for cross-modal alignment. The CoATT (Jing, Xie, and Xing 2018) and the AlignTrans (You et al. 2021) choose to align the visual features with the predicted disease labels. The M2KT (Yang et al. 2023) designs a mutimodal alignment module to align the pooled visual feature

with the pooled textual feature from a Bert and the predicted disease labels. However, such methods rely on label annotations, and the cross-modal information content in labels or pooled textual features is relatively low. Differ from them, we use a memory matrix as an alignment medium, which does not require additional annotations and can provide rich cross-modal knowledge. The CMN (Chen et al. 2021) also employed a memory matrix to store cross-modal knowledge. But they align the visual and textual features by re-mapping features of two modals to a same vector space constructed by the memory matrix. Differ from it, our method is more efficient, and allows the alignment module to utilize the cross-modal knowledge learnt from the generator.

Method

Overview

In this paper, we propose to generate reports based on the Memory Alignment (MA) module. The architecture of our model is shown in Figure 2, which consists of three parts: visual extractor, memory alignment module and report generator. First, we extract patch features from the radiology image by a pretrained CNN, then embed the patch features to the visual embeddings I. Next step, we add the visual embeddings with its positional embedding P as the memory query, to get the memory alignment embeddings R through the MA. Afterward, we get the input embeddings by combining these embeddings as depicted in the figure, then feed them into the Bert encoder to obtain the encoded features E. In the end, the generation head will generate the entire article according to the E through a chain loop.

Visual Extractor

Our model enhire a pretrained CNN, such as Densenet, to extract the visual features of radiology images. Similar to other approaches, images are reshaped to the same size, and the visual features are extracted from the last convolutional layer of CNN.

$$\mathbf{F} = f(Img) \tag{1}$$

where $f(\cdot)$ refers to the CNN visual extractor, $\mathbf{F}$ for the visual features. Then the visual features will be embedded into the visual embeddings $\mathbf{I}$.

$$\mathbf{I}\{v_1, v_2, v_3, ..., v_S\} = \mathrm{Embed}(\mathbf{F}) \tag{2}$$

Where S refers to the number of the visual embeddings.

Memory Alignment Module

Aligning the image and text features can facilitate the report generation process, but due to the lack of corresponding annotations, it is hard to align image and text features directly. Compromised with this, some choose to align the image features with the disease tags, but the effectiveness is limited by the accuracy of the disease classification and the limited information contained in the tags. Different from these methods, Cross-Modal Memory Networks (CMN) enhire a learnable memory matrix (M) to provide cross-modal knowledge (Chen et al. 2021). We strongly agree with this approach, but we also believe that its performance is limited by the implementation method. They combined representation and alignment, so M in CMN not only needs to learn cross-modal knowledge, but also needs to learn the distribution of features, which brings additional parameters and calculations.

While our proposed Memory Alignment (MA) module utilizes M in a more efficient way. As shown in Figure 3, it takes visual embeddings and corresponding positional embeddings as input and a memory alignment embeddings to align the visual and text features. Positional embeddings in the input can provide additional cross-modal knowledge from the generator, and the alignment method dictates that the M does not need to store additional information.

Specifically, we denote the memory matrix as $M = \{m_1, m_2, ..., m_\eta\}$, η for the number of memory vectors, $m_i \in \mathbb{R}^d$, d for the hidden dimension of our model. During the query process, we adopt a muti-head querying, where each head operates in the same way as shown below. For each head, the query $\mathbf{q}$ is formed by summing the visual embeddings $\mathbf{I}$ and the corresponding positional embeddings $\mathbf{P} = \{p_1, p_2, p_3, ..., p_S\}$, while the matrix M serves as both the key $\mathbf{k}$ and value $\mathbf{v}$.

$$\mathbf{q}_i = (v_i + p_i) \cdot W_q \tag{3}$$

$$\mathbf{k}_j = m_j \cdot W_k \tag{4}$$

$$\mathbf{v}_k = m_k \cdot W_v \tag{5}$$

W_q, W_k, W_v in the above formulas are trainable weights. Then we calculate the distance $D_i = \{D_{i1}, D_{i2}, ..., D_{i\eta}\}$ between $\mathbf{q}_i$ and k by

$$D_{ij} = \frac{\mathbf{q}_i \cdot \mathbf{k}_j^\intercal}{\sqrt{d}} \tag{6}$$

Through the distance D we can get the correlation weight w between $\mathbf{q}_i$ and $\mathbf{k}$ by a softmax

$$\mathrm{w}_i = \mathrm{softmax}(D_i) \tag{7}$$

Then we can get the response $\mathbf{r}_i$ of the query $\mathbf{q}_i$ by a weighted sum as follows:

$$\mathbf{r}_i = \sum_{j=1}^{\eta} \mathrm{w}_{ij} \cdot \mathbf{v}_j \tag{8}$$

Finally we concatenate the responses $\mathbf{r}$ from all heads and get the memory alignment embeddings R by a linear layer.

Report Generator

The report generator consists of a Bert-based encoder and a generation head. The generation process consists of two steps: first we process the input and feed it into the encoder, and then input the encoded features into the generation head to generate the next word. After this we will append the newly generated word to the input of encoder and repeat this process until the entire report is generated. Details are as follows.

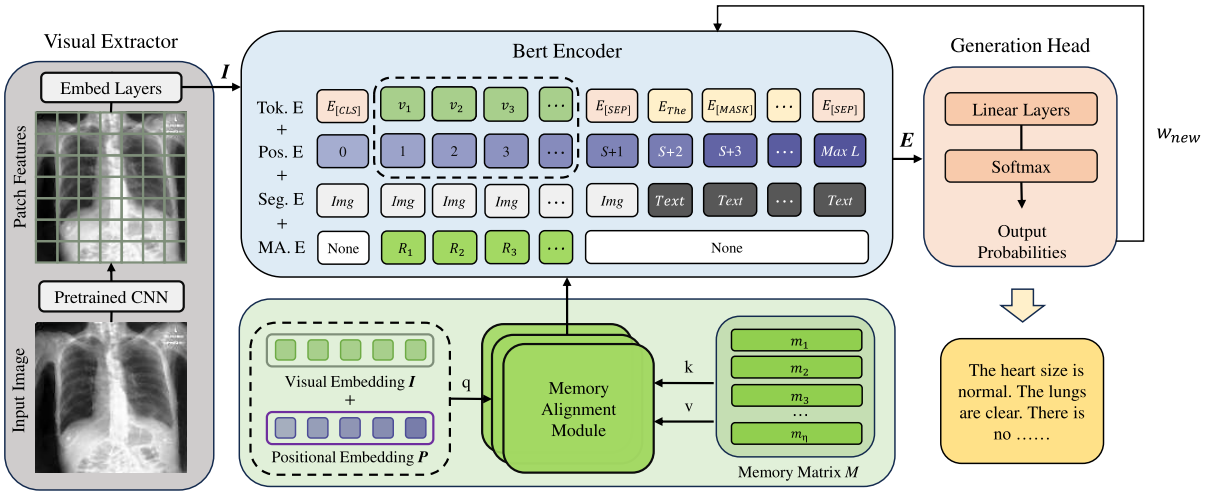

Figure 2: The architecture of our model. In the Bert Encoder, *Max L* represents the max input length, the Tok.E , Pos.E , Seg.E and MA.E refer to the token, positional, segment and memory alignment embeddings, respectively. Vectors I and P in two dashed boxes are the same.

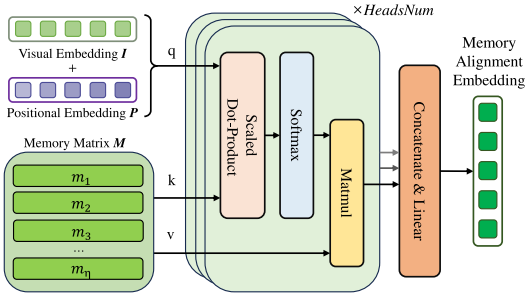

Figure 3: The architecture of Memory Alignment Module

Bert-based Encoder. We enhire a Bert-based encoder to encode the features of different modalities. The input of the encoder consists of the visual embeddings, the generated words and the memory alignment embeddings. First, we embed the input words into the word embeddings, then concatenate the visual embeddings and the word embeddings to obtain the token embeddings. Then, we sum the token embeddings with positional embeddings, segment embeddings and the memory alignment embeddings to get the input embeddings. Specially, the position sequence for the positional embeddings is an increasing sequence from 0 to the max length of the input sequence *Max L*, and the segment embeddings consists of vision (*Img*) and language (*Text*) tags. Finally we feed the input embeddings into the BERT encoders to obtain the encoded features E.

Generation Head. We enhire a generation head to predict the words probabilities from E, which consists of linear layers and a softmax. The newly generated word is fed back into the input of encoder, looping until the entire report is generated.

Experiment Settings

Datasets. We evaluate our method on two public reports generation datasets, MIMIC-CXR and IU X-Ray. All protected health information are de-identified. The MIMIC-CXR is currently the largest dataset for the radiograph reports generation task, which contains 337,110 chest X-ray images and 227,835 reports. For a fair comparison with previous methods, we use the official splitting for training, validation and testing. The IU X-Ray consists of 7,470 frontal and lateral-view chest X-ray images and 3,955 corresponding reports. The findings and impression sections are concatenated as the ground-truth reports. Following the common approach, we randomly split it into training, validation, and testing sets with the ratio of 7:1:2.

Metrics. We choose the Natural Language Generation (NLG) metrics that include BLEU (Papineni et al. 2002), METEOR (Denkowski and Lavie 2011) and ROUGE-L (Lin 2004) scores to evaluate the performance. And the CheXpert (Irvin et al. 2019) labeler is employed to annotate the reports with 14 categories. According to these annotations, the Precision, Recall and F1 scores can be calculated over the generated reports and the ground truth reports as the Clinical Efficacy (CE) metrics. For all metrics, a higher number means better performance.

Settings. We adopt the DenseNet121 (Huang et al. 2017) pre-trained on the ImageNet (Russakovsky et al. 2015) as the visual features extractor. All radiographs are resized and cropped into 224×224, the visual features extracted from one radiograph is 2048×49. The frequency threshold of the tokenizer is set to 3, obtain 7,861 and 764 tokens [2] from MIMIC-CXR and IU X-Ray, respectively. For MIMIC-CXR, the maximum length of the report is set as 100 and for IU X-Ray it is 60. We employ the pre-trained

[2]Includes four special tokens [*CLS*], [*PAD*], [*SEP*] and [*UNK*].

uncased BERT-base model (Devlin et al. 2019) as the BERT encoders, which has 12 encoder layers, and the hidden dimension is 768. The dimension of memory vectors is the same as the hidden dimension. For MIMIC-CXR, number of vectors in the memory matrix η is set to 100, *HeadsNum* is set to 4, and for IU X-Ray they are 64 and 2, respectively. During the inference stage, we adopt a beam search strategy with a beam size of 5 for sampling reports.

We train our model under cross entropy loss, the learning rates of the visual extractor and other parameters are set to $5e-5$ and $1e-4$, respectively. The AdamW (Loshchilov and Hutter 2019) optimizer is adopted with a weight decay of 0.01. All parameters are randomly initialized except the pre-training parameters, and all random seeds are fixed. All experiments are run on the Nvidia Geforce 3090 GPUs. For MIMIC-CXR, the batch size is set as 48, the average training time is 1.38 hours per epoch, with a total of 30 epochs trained. The average inference time is 0.19 seconds per radiograph. And for IU X-Ray, the batch size is set as 24, the training time is 2.1 hours for 30 epochs and the average inference time is 0.14 seconds per radiograph.

Results and Analyses

In this section, we compared the performance of some representative models with our model on MIMIC-CXR and IU X-Ray. The comparison results are shown in Table 1, the model testing indicators included in it are all from their published literature. Models in the first group are mainly generation-based, including R2Gen (Chen et al. 2020), CMN (Chen et al. 2021), PPKED (Liu et al. 2021a), Ali-Trans (You et al. 2021) and M2KT (Yang et al. 2023). Models in the second group are retrieval-based, including HRGR (Li et al. 2018), KERP (Li et al. 2019) and KGAE (Liu et al. 2021b).

Overall, on MIMIC-CXR, our model achieves the best results on BLEU-1, BLEU-2, BLEU-3 and the second best results on BLEU-4. And for IU X-Ray, our model achieves the best results on BLEU-2, Meteor and the second best results on BLEU-1, BLEU-3 and ROUGE-L. In detail, our model outperforms the generation-based models in almost all metrics on both datasets, which shows that our alignment method is effective. Additionally, our model achieves better performance on MIMIC-CXR than IU X-Ray. This is owing to that MIMIC-CXR is much lager than IU X-Ray and provides richer cross-modal information. But if the retrieval-based models are also taken into account, the limitation of our model on some metrics (e.g. BLEU-4) are revealed. This may come from the fact that our model are not so good at generating long phrases as retrieval-based models.

For the Clinical Efficacy metrics, as shown in Table 2, our model achieves the best results on Precision, Recall and F1 scores, which demonstrates that reports generated by our model are not only coherent but also accurate in clinical diagnosis.

Effect of Memory Alignment Module

We explore the effect of our Memory Alignment (MA) module through ablation studys, and compare it with the Cross-Modal Memory on the same backbone. The result is shown in Table 3, all listed models are trained on MIMIC-CXR under the same settings.

Ablation Study. In this part, we tested three models on MIMIC-CXR for ablation study. Model (a) serves as the baseline, comprising a visual feature extractor, BERT encoders, and a generation head. Model (b) adds a MA based on model (a), which using only visual embeddings as input of MA. Model (d) is our full model. Compared to the baseline which generates reports according to the visual features only, (b) and (d) gain additional cross-modal knowledge from the memory alignment embeddings to assist in generation. As a result, they both outperform the baseline, with an average improvement on all NLG metrics by 7.91% and 10.29%. Model (d) performs better than (b) after adding the positional embedding to the input, with an average improvement of 2.17% across all metrics, indicating better alignment performance. This is because the positional embedding contains some cross-modal knowledge learnt by the generator, which may be able to add some "common sense" to the alignment. We believe that this can further improve the robustness of the model as well as the alignment.

Comparison with the Cross-Modal Memory. In this part, we compare our model with the Cross-Modal Memory (CMM) from three aspects: parameter quantity, alignment costs and NLG metrics. Model (c) is the combination of our baseline and the CMM, whose structure and hyperparameters are consistent with those described in (Chen et al. 2021).

In terms of parameter quantity, the memory matrix in CMM consists of 2,048 memory vectors, while ours consists of only 100 memory vectors, accounting for about 4.9% of the former. Furthermore, there is a huge alignment cost gap between the two methods, since our method needs to be computed only once during the generation process, while the memory matrix of CMM needs to be queried every time for a newly generated word. Under the same conditions, the average inference time of Model (a), Model (c) and Model (d) is 0.19s, 0.48s and 0.19s per radiograph, respectively. Our model has almost no impact on the inference time, while the Model (c) takes about 2.5 times longer to do it. At the same time, our model outperform the Model (c) with an average improvement on all NLG metrics by 6.8%. In summary, on the same backbone, our method achieves better results with fewer parameters and alignment costs. This is because in our approach, the memory matrix only needs to store the cross-modal knowledge and some additional knowledge from the generator can be utilized.

Impact of the Memory Matrix Size

We trained our model in different memory vector numbers η from 16 to 256 on the MIMIC-CXR to study its impact on the results. As shown in Table 4, when η is relatively small (e.g. 16 and 32), the model performs poorly, means that the alignment module even has some negative effects. This may due to the fact that the alignment module can only give some similar outputs due to insufficient parameters, which in turn affects the utilization of visual features. On the other hand, when η is too large (e.g. 256), the results also decreased due to overfitting or other reasons. The model achieves the best

Models	Years	MIMIC-CXR						IU X-Ray					
		B@1	B@2	B@3	B@4	M	R	B@1	B@2	B@3	B@4	M	R
R2Gen	2020	0.353	0.218	0.145	0.103	0.128	0.267	0.470	0.304	0.219	0.165	0.187	0.371
CMN	2021	0.353	0.218	0.148	0.108	0.142	0.277	0.475	0.309	0.222	0.170	0.191	0.375
PPKED	2021	0.360	0.224	0.149	0.106	0.149	0.284	0.483	0.315	0.224	0.168	-	0.376
Ali-Trans	2021	0.378	0.235	0.156	0.112	**0.158**	0.283	0.484	0.313	0.225	0.173	0.204	0.379
M2KT	2023	0.386	0.237	0.157	0.111	-	0.274	0.497	0.319	0.230	0.174	-	0.399
HRGR	2018	-	-	-	-	-	-	0.438	0.298	0.208	0.151	-	0.322
KERP	2019	-	-	-	-	-	-	0.482	0.325	0.226	0.162	-	**0.339**
KGAE	2021	0.369	0.231	0.156	**0.118**	0.153	**0.295**	**0.512**	0.327	**0.240**	**0.179**	0.195	0.383
Ours		**0.396**	**0.244**	**0.162**	0.115	0.151	0.274	0.501	**0.328**	0.230	0.170	**0.213**	0.386

Table 1: Result on the test set of MIMIC-CXR and IU X-Ray. B@n for BLEU-n, R for ROUGE-L, and M for METEOR. Models in the first group are mainly generation-based, and in the second group are retrieval-based. Bolded numbers are the best results, and underlined ones are the second best.

Models	P	R	F1
TOPDOWN (Anderson et al. 2018)	0.322	0.239	0.249
R2Gen (Chen et al. 2020)	0.333	0.273	0.276
CMN (Chen et al. 2021)	0.334	0.275	0.278
RL-CMN (Qin and Song 2022)	0.342	0.294	0.292
KGAE (Liu et al. 2021b)	0.389	0.362	0.355
Ours	**0.411**	**0.398**	**0.389**

Table 2: Results of the CE metrics on the MIMIC-CXR.P for Precision, R for Recall

Model	MA		CMM	MIMIC-CXR			
	v	p		B@1	B@4	M	R
(a)				0.357	0.101	0.136	0.270
(b)	✓			0.389	0.110	0.149	**0.275**
(c)			✓	0.369	0.106	0.141	0.267
(d)	✓	✓		**0.396**	**0.115**	**0.151**	0.274

Table 3: Result of the ablation studys and comparison experiment on MIMIC-CXR. v and p refer to visual embeddings and positional embeddings respectively, and the CMM refers to the Cross-Modal Memory.

performance when η is 100, and the performance around it is not much different. Therefore, in order to achieve good results, a suitable size of the memory matrix is required. However, this does not mean that the module is sensitive to its size, the performance can be significantly improved within a reasonable range of η.

Visual Analysis

We perform a case in Figure 4 to qualitative analysis the effect of the alignment between visual and textual features. The figure contains a radiology image with a ground-truth report taken from MIMIC-CXR, and two reports generated by the baseline and our model, respectively. For further analysis, we show some activation maps of the generated words (bold and italic), which can qualitatively demonstrate the alignment effect of the model. The darker the color in the activation maps, the more attention the model pays to this area.

η	BLEU-1	BLEU-4	M	R	Δ
16	0.346	0.092	0.125	0.253	-6.5%
32	0.355	0.098	0.135	0.264	-1.6%
64	0.370	0.102	0.139	0.264	1.2%
100	**0.396**	**0.115**	**0.151**	**0.274**	9.3%
128	0.391	0.111	0.148	0.272	7.2%
256	0.384	0.108	0.143	0.269	4.8%

Table 4: Model with different memory vector numbers η. The Δ in the table refers to the average implementation on the baseline.

Note that our model has different attention regions for different words, shows that there exist some corresponding relationships between visual and textual features. For examlpe, the word "*cardiomediastinal*" corresponds well to the area of the heart. However, the baseline focuses almost on a fixed area, which means it is more likely to ignore the image and randomly generate high-frequency sentences. The report generated by the baseline model contains some relevant findings, but it also includes many irrelevant high-frequency expressions, such as "*As compared to the previous...*". Though high-frequency expressions may help in achieving high NLG metrics, they are not always in line with the actual situation. While the report generated by our model contains many expressions (underlined) that are the same or similar to the ground-truth, such as "*no focal consolidation pleural effusion or pneumothorax*". This Shows that the model can generate better and accurate reports with the aids of the alignment.

Conclusion

In this paper, we propose a Memory Alignment module, which can align the visual and textual features through a memory matrix, and assist the model to generate a coherent and informative report. Visual analysis demonstrates that our model can effectively align the information between im-

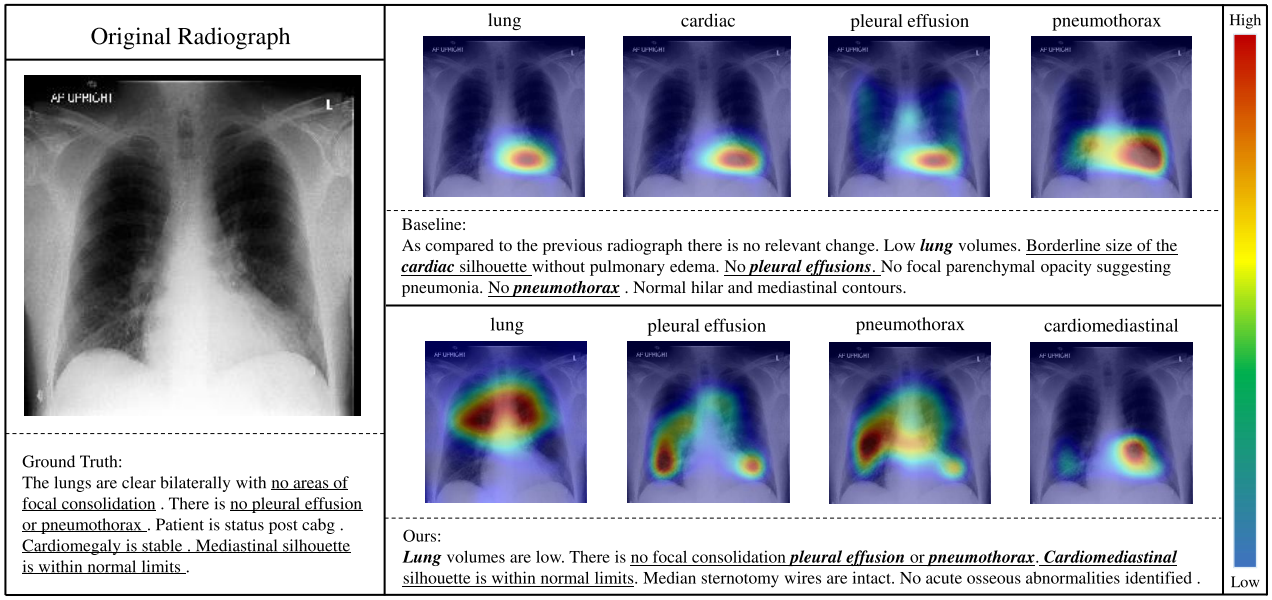

Figure 4: Visualization of the case, the left is the original image and report from MIMIC-CXR, and the right are the reports with partial activation maps generated by the baseline and our model respectively. The activation map corresponds to the word above it, the darker the color in the activation maps, the more attention the model pays to that area.

ages and text, and generate more accurate reports. The comparison experiments with other alignment approaches proves that our Memory Alignment module is simpler and more effective. In addition, our model demonstrated excellent performance on both public datasets, further validating the effectiveness of the proposed Memory Alignment module.

Acknowledgments

This work was supported by Beijing Natural Science Foundation(7222086).

References

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Aneja, J.; Deshpande, A.; and Schwing, A. G. 2018. Convolutional Image Captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chen, Q.; Deng, C.; and Wu, Q. 2022. Learning Distinct and Representative Modes for Image Captioning. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 9472–9485. Curran Associates, Inc.

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 (Volume 1: Long Papers)*, 5904–5914. Online: Association for Computational Linguistics.

Chen, Z.; Song, Y.; Chang, T.-H.; 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)*, 1439–1449. Online: Association for Computational Linguistics.

Demner-Fushman, D.; Kohli, M. D.; Rosenman, M. B.; Shooshan, S. E.; Rodriguez, L.; Antani, S.; Thoma, G. R.; and McDonald, C. J. 2015. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2): 304–310.

Denkowski, M.; and Lavie, A. 2011. Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. 85–91.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805.

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)*, 2261–2269.

Irvin, J.; Rajpurkar, P.; Ko, M.; Yu, Y.; Ciurea-Ilcus, S.; Chute, C.; Marklund, H.; Haghgoo, B.; Ball, R. L.; Shpanskaya, K. S.; Seekins, J.; Mong, D. A.; Halabi, S. S.; Sandberg, J. K.; Jones, R.; Larson, D. B.; Langlotz, C. P.; Patel, B. N.; Lungren, M. P.; and Ng, A. Y. 2019. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. *CoRR*, abs/1901.07031.

Jing, B.; Xie, P.; and Xing, E. 2018. On the Automatic Generation of Medical Imaging Reports. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2577–2586. Melbourne, Australia: Association for Computational Linguistics.

Johnson, A. E. W.; Pollard, T. J.; Berkowitz, S. J.; Greenbaum, N. R.; Lungren, M. P.; ying Deng, C.; Mark, R. G.; and Horng, S. 2019. MIMIC-CXR: A large publicly available database of labeled chest radiographs. *ArXiv*, abs/1901.07042.

Li, C. Y.; Liang, X.; Hu, Z.; and Xing, E. P. 2018. Hybrid Retrieval-Generation Reinforced Agent for Medical Image Report Generation. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, 1537–1547. Red Hook, NY, USA: Curran Associates Inc.

Li, C. Y.; Liang, X.; Hu, Z.; and Xing, E. P. 2019. Knowledge-Driven Encode, Retrieve, Paraphrase for Medical Image Report Generation. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press. ISBN 978-1-57735-809-1.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of summaries. 10.

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

Liu, F.; You, C.; Wu, X.; Ge, S.; wang, S.; and Sun, X. 2021b. Auto-Encoding Knowledge Graph for Unsupervised Medical Report Generation. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 16266–16279. Curran Associates, Inc.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. arXiv:1711.05101.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, 311–318. USA: Association for Computational Linguistics.

Qin, H.; and Song, Y. 2022. Reinforced Cross-modal Alignment for Radiology Report Generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, 448–458. Dublin, Ireland: Association for Computational Linguistics.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. C.; and Fei-Fei, L. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3): 211–252.

Tran, A.; Mathews, A.; and Xie, L. 2020. Transform and Tell: Entity-Aware News Image Captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

Xu, K.; Ba, J. L.; Kiros, R.; Cho, K.; Courville, A.; Salakhutdinov, R.; Zemel, R. S.; and Bengio, Y. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, 2048–2057. JMLR.org.

Xue, Y.; Xu, T.; Rodney Long, L.; Xue, Z.; Antani, S.; Thoma, G. R.; and Huang, X. 2018. Multimodal Recurrent Model with Attention for Automated Radiology Report Generation. In Frangi, A. F.; Schnabel, J. A.; Davatzikos, C.; Alberola-López, C.; and Fichtinger, G., eds., *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, 457–466. Cham: Springer International Publishing. ISBN 978-3-030-00928-1.

Yan, B.; and Pei, M. 2022. Clinical-BERT: Vision-Language Pre-training for Radiograph Diagnosis and Reports Generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(3): 2982–2990.

Yang, S.; Wu, X.; Ge, S.; Zheng, Z.; Zhou, S. K.; and Xiao, L. 2023. Radiology report generation with a learned knowledge base and multi-modal alignment. *Medical Image Analysis*, 86: 102798.

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 de Bruijne, M.; Cattin, P. C.; Cotin, S.; Padoy, N.; Speidel, S.; Zheng, Y.; and Essert, C., eds., *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, 72–82. Cham: Springer International Publishing. ISBN 978-3-030-87199-4.