Reinforced Cross-modal Alignment for Radiology Report Generation

Han Qin[♠], **Yan Song**[♠†]
[♠]The Chinese University of Hong Kong (Shenzhen)
[♠]hanqin@link.cuhk.edu.cn [♠]songyan@cuhk.edu.cn

Abstract

Medical images are widely used in clinical decision-making, where writing radiology reports is a potential application that can be enhanced by automatic solutions to alleviate physicians' workload. In general, radiology report generation is an image-text task, where cross-modal mappings between images and texts play an important role in generating high-quality reports. Although previous studies attempt to facilitate the alignment via the co-attention mechanism under supervised settings, they suffer from lacking valid and accurate correspondences due to no annotation of such alignment. In this paper, we propose an approach with reinforcement learning (RL) over a cross-modal memory (CMM) to better align visual and textual features for radiology report generation. In detail, a shared memory is used to record the mappings between visual and textual information, and the proposed reinforced algorithm is performed to learn the signal from the reports to guide the cross-modal alignment even though such reports are not directly related to how images and texts are mapped. Experimental results on two English radiology report datasets, i.e., IU X-Ray and MIMIC-CXR, show the effectiveness of our approach, where the state-of-the-art results are achieved. We further conduct human evaluation and case study which confirm the validity of the reinforced algorithm in our approach.[1]

1 Introduction

Radiology report generation aims to automatically generate a free-text description from a specific clinical radiograph (e.g., chest X-ray), which can significantly alleviate the burden of radiologists and thus improve the quality and standardization of health care. With the advantages of its applications, radiology report generation has become an interesting research topic attracted in both artificial intelligence and clinical medicine. Recently, to generate more accurate reports, approaches based on deep learning techniques are adapted to this task and have achieved great success (Jing et al., 2018; Li et al., 2018; Liu et al., 2019).

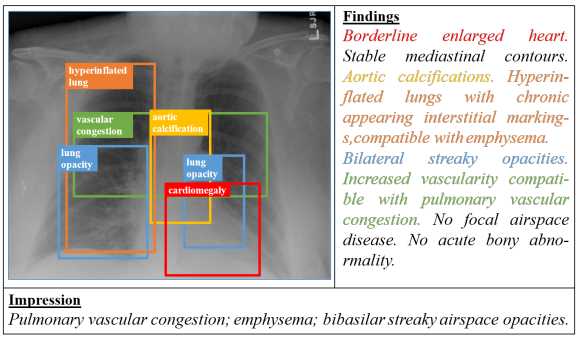

Figure 1: A chest X-ray image with its report, where aligned visual and textual features are linked by colors.

To effectively perform radiology report generation, most existing studies adopted conventional encoder-decoder architectures with convolutional neural networks (CNN) as the encoder and recurrent neural networks (e.g., LSTM (Hochreiter and Schmidhuber, 1997), GRU (Cho et al., 2014)) or non-recurrent neural networks (e.g., Transformer (Vaswani et al., 2017)) as the decoder. Considering that there is alignment between radiographs and their corresponding doctor-written reports (such as the mappings demonstrated in Figure 1 where visual and textual features representing the same content are highlighted in the same color), the ability of a model to learn such alignment is the key to achieve outstanding performance. To model the alignment information, Jing et al. (2018) proposed a co-attention mechanism to explicitly learn the linking between visual features in the radiographs and the semantic information in the corresponding doctor-written text reports, where the model is trained to generate text sequences via maximum likelihood estimation (MLE). However, one chal-

[†]Corresponding author.
[1]Our code for this paper is released at https://github.com/cuhksz-nlp/R2GenRL.

Findings of the Association for Computational Linguistics: ACL 2022, pages 448 - 458
May 22-27, 2022

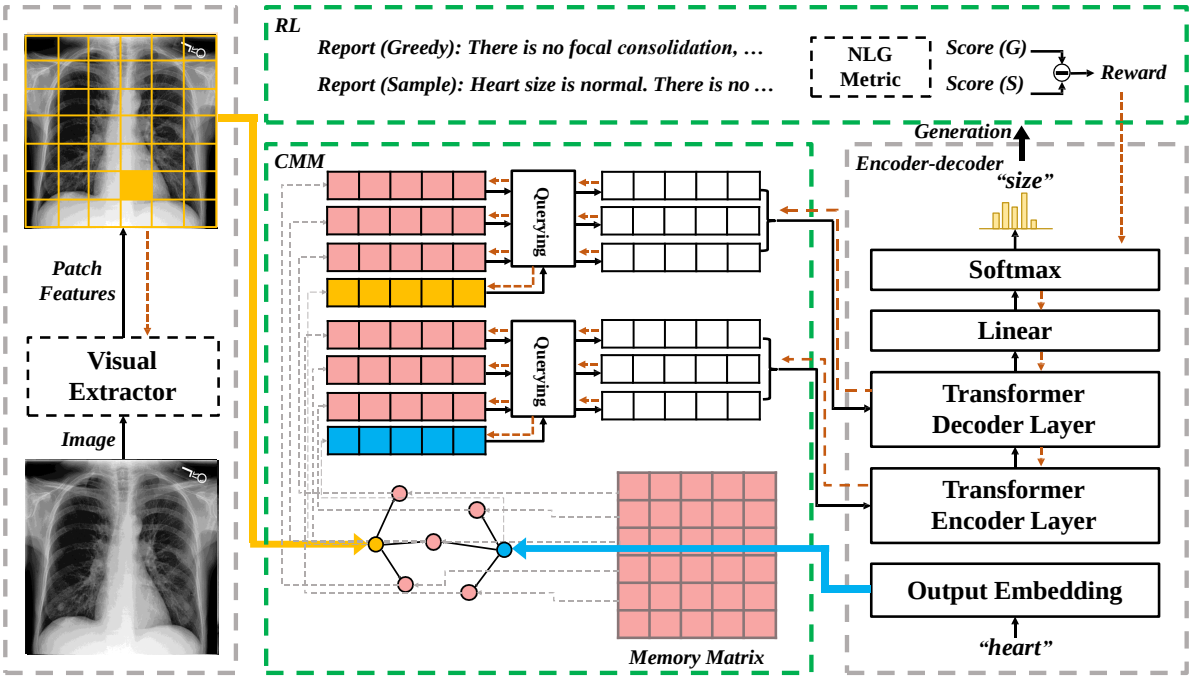

Figure 2: The overall architecture of our proposed approach, where the visual extractor, encoder, and decoder are shown in gray dash boxes with the details omitted. The CMM and reward computation process of RL are illustrated in green dash boxes, and the orange dash arrows indicate back-propagation of gradients from training the model. The orange, blue, and red nodes in CMM denote the vector representations of visual features, textual features, and memories, respectively.

lenge to learn the alignment is that there is no annotated alignment for such research to perform supervised learning to accurately map cross-modal information, so that normal learning procedure may not fit this scenario. To address this challenge, reinforcement learning (RL) is a potential solution, because it is able to guide the learning process of the cross-modal alignment with appropriate supervision from carefully designed rewards. Although there are studies following this paradigm by using RL to perform report generation, they focused on other aspects of this task rather than facilitating the mappings for cross-modal information. For example, Li et al. (2018) designed sentence- and word-level rewards to guide the model to choose to either retrieve a template sentence or generate a new sentence, and Jing et al. (2019) ultilized multi-agent RL to capture the imbalanced distribution between abnormality and normality. Therefore, RL on cross-modal alignment is expected to be studied and has the potential for further improvements.

In this paper, we propose to enhance radiology report generation via reinforced cross-modal alignment to alleviate the requirement of annotated supervision while facilitate the interactions across modalities (i.e., images and texts). In detail, our approach is based on Chen et al. (2021b), where a cross-modal memory (CMM) module is used to stores the cross-modal information that bridges the visual and textual features. Based on CMM, the proposed RL algorithm is applied to leverage the signals from natural language generation (NLG) metrics, i.e., BLEU (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2011) and ROUGE (Lin, 2004), to guide the cross-modal mappings so as to better matching features from images and texts as well as have a direct target of learning outcome for report generation. Experimental results confirm the validity of our approach, which outperforms strong baselines and achieves the state-of-the-art performance on two widely used benchmark datasets, i.e., IU X-Ray (Demner-Fushman et al., 2016) and MIMIC-CXR (Johnson et al., 2019). Moreover, we perform human evaluation and case study to further illustrate the validity of the proposed RL in our approach.

2 The Proposed Approach

Following previous studies (Jing et al., 2018; Li et al., 2018; Liu et al., 2019; Chen et al., 2021b), we treat radiology report generation as a sequence-to-sequence task, where the source sequence are patch features $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_s, ..., \mathbf{x}_S\}$ from an input image, where $\mathbf{x}_s \in \mathbb{R}^d$ is extracted by

visual extractors, and the target sequence $\mathbf{Y} = \{y_1, y_2, ..., y_t, ..., y_T\}$ is the corresponding report to the image, with $y_t \in \mathbb{V}$ being generated tokens, T the length of the report and $\mathbb{V}$ the vocabulary of all possible tokens. On the top of the general sequence-to-sequence paradigm, we add CMM which allows the proposed RL to take the signal from $\mathbf{Y}$ and use it to guide cross-modal mappings for $\mathbf{X}$ and $\mathbf{Y}$. An overview of our proposed approach is presented in Figure 2, where the details for different parts are illustrated as follows.

2.1 The Overall Generation Pipeline

In general, our model is composed of three major components, i.e., the visual extractor, the CMM, and the encoder-decoder part, where the CMM is dynamically integrated into the encoding and decoding process. The high-level descriptions of the three components are explained below.

Visual Extractor The visual features $\mathbf{X}$ of a radiology image $\mathbf{I}$ are extracted by pre-trained convolutional neural networks (CNN), such as VGG (Simonyan and Zisserman, 2015) or ResNet (He et al., 2016). Normally, an image is decomposed into regions with equal size (i.e., patches), and the extracted features from patches are then expanded into a sequence by simply concatenating all features from each row in a row-by-row manner, where the process is formulated by

$$\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_s, ..., \mathbf{x}_S\} = f_v(\mathbf{I}) \quad (1)$$

with $f_v(\cdot)$ representing the visual extractor. Then the result is used as the source sequence for all subsequent modules.

Cross-modal Memory Memories are widely used to model the associations between different types of features through the mapping of keys and values and its effectiveness in doing so is demonstrated in many previous natural language processing studies (Miller et al., 2016; Xu et al., 2019; Tian et al., 2020; Cornia et al., 2020; Nie et al., 2020; Wu et al., 2021; Chen et al., 2021a). Therefore, we use CMM, which is based on Chen et al. (2021b), to record potentially shared information of visual and textual features in the memory so that the entire learning process is able to explicitly map corresponding parts in images and texts within a unified representation space. Formally, given a source sequence $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_s, ..., \mathbf{x}_S\}$ from an image, we feed it to the CMM module to obtain the memory correspondences $\{\mathbf{r}_{\mathbf{x}_1}, \mathbf{r}_{\mathbf{x}_2}, ..., \mathbf{r}_{\mathbf{x}_s}, ..., \mathbf{r}_{\mathbf{x}_S}\}$ for the visual features. Similarly, the given the generated text sequence $\{y_1, y_2, ..., y_{t-1}\}$ for $\mathbf{I}$ with embedding $\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{t-1}\}$ is also fed to CMM to form memory correspondences for textual features $\{\mathbf{r}_{\mathbf{y}_1}, \mathbf{r}_{\mathbf{y}_2}, ..., \mathbf{r}_{\mathbf{y}_{t-1}}\}$.

Encoder-Decoder The encoder-decoder in our model is built upon standard Transformer. In detail, the encoding process is formulated as

$$\{\mathbf{z}_1, \mathbf{z}_2 ..., \mathbf{z}_S\} = f_e(\mathbf{r}_{\mathbf{x}_1}, \mathbf{r}_{\mathbf{x}_2}, ..., \mathbf{r}_{\mathbf{x}_S}) \quad (2)$$

where $f_e(\cdot)$ is the encoder. With the encoded results, the decoding process is formulated by

$$y_t = f_d(\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_S, \mathbf{r}_{\mathbf{y}_1}, \mathbf{r}_{\mathbf{y}_2}, ..., \mathbf{r}_{\mathbf{y}_{t-1}}) \quad (3)$$

where $f_d(\cdot)$ is the decoder and y_t the generated token at the current time step.

2.2 Cross-modal Memory

In our approach, CMM serves as an intermediate medium to connect the visual and textual features and thus allows the model to automatically learn the cross-modal mappings without relying on gold annotated alignments. Specifically, CMM contains a memory matrix[2] $\mathbf{M} = [\mathbf{m}_1, ..., \mathbf{m}_i, ..., \mathbf{m}_\mathcal{N}]$ that consists of $\mathcal{N}$ memory vectors ($\mathbf{m}_i$ is the i-th memory vector) to align the visual and textual features. It applies multi-thread[3] alignments to the visual features (i.e., $\mathbf{x_s}$), the textual features (i.e., $\mathbf{y}_t$), and the memory vectors (i.e., $\mathbf{m_i}$), where the alignments in all threads follow the same procedure.

In detail, in each thread, it firstly maps $\mathbf{x_s}$, $\mathbf{y}_t$, and $\mathbf{m_i}$ to the alignment space $\mathbf{x}'_s$, $\mathbf{y}'_t$, $\mathbf{k}_i$ through three trainable matrices (i.e., $\mathbf{W}_x$, $\mathbf{W}_y$, and $\mathbf{W}_k$), respectively, which is formally represented by

$$\mathbf{x}'_s = \mathbf{x}_s\mathbf{W}_x, \ \mathbf{y}'_t = \mathbf{y}_t\mathbf{W}_y, \ \mathbf{k}_i = \mathbf{m}_i\mathbf{W}_k \quad (4)$$

Next, for each visual feature (i.e., $\mathbf{x}'_s$), CMM computes the distances (denoted by $d_{s,i}$) between $\mathbf{x}'_s$ all memory vectors (i.e., $\mathbf{k}_i$) in the alignment space by $d_{s,i} = \mathbf{x}'_s \cdot \mathbf{k}_i$ and extracts the closest $\mathcal{K}$ memory vectors (i.e., keys in the memories) which are denoted as $[\mathbf{k}_{s,1}, \cdots, \mathbf{k}_{s,j}, \cdots, \mathbf{k}_{s,\mathcal{K}}]$. Then, CMM finds the corresponding memory vectors $\mathbf{m}_{s,j}$ of the keys $\mathbf{k}_{s,j}$ in the memory matrix and uses a trainable matrix $\mathbf{W}_v$ to map $\mathbf{m}_{s,j}$ to its corresponding value vectors ($\mathbf{v}_{s,j}$) through $\mathbf{v}_{s,j} = \mathbf{m}_{s,j} \cdot \mathbf{W}_v$

[2]One way to obtain the memory matrix $\mathbf{M}$ is to randomly initialize it and then update it during the training process.

[3]Thread number can be arbitrarily set in experiments.

Afterwards, we compute the weighted sum of the value vectors and obtain the output $\mathbf{r}_{\mathbf{x}_s}$ for $\mathbf{x}_s$ by

$$\mathbf{r}_{\mathbf{x}_s} = \Sigma_{j=1}^{\mathcal{K}} w_{s,j} \mathbf{v}_{s,j} \tag{5}$$

where the weight $w_{s,j}$ is computed through

$$w_{s,j} = \frac{\exp(\mathbf{x}'_s \cdot \mathbf{k}_{s,j})}{\Sigma_{j=1}^{\mathcal{K}} \exp(\mathbf{x}'_s \cdot \mathbf{k}_{s,j})} \tag{6}$$

The same procedure is applied to the textual features and obtain the output $\mathbf{r}_{\mathbf{y}_t}$ for $\mathbf{y}_t$. Finally, CMM concatenates the output $\mathbf{r}_{\mathbf{x}_s}$ and $\mathbf{r}_{\mathbf{y}_t}$ from all threads and feeds them to the encoder-decoder structure in our model.

2.3 Reinforced Cross-modal Alignment

Although CMM provides a "soft" mechanism to facilitate the linking between visual and textual features, there is still no annotated alignment to guide an accurate learning process, which is a common problem exists in previous work (Jing et al., 2018). To address this problem, we propose to use RL to provide appropriate supervision from NLG evaluation metrics to search for better mappings between features from different modalities. In doing so, we treat the genration model as the **agent** that interacts with an external **environment** (visual and textual features). Therefore, all parameters of our approach, θ, define a **policy** p_θ that results in an **action** (i.e., the prediction of the next word). Upon generating the end-of-sequence (EOS) token, the agent uses a **reward** r based on evaluation metrics, e.g., BLEU, METEOR and ROUGE, etc., where the reward r_t for the action at step t is the improvement on the evaluation metric by generating the the next word y_t, which is formally expressed by

$$r_t = r(\mathbf{Y}_t) - r(\mathbf{Y}_{t-1}) \tag{7}$$

where $\mathbf{Y}_t = \{y_1, y_2, ..., y_t\}$ and $\mathbf{Y}_{t-1} = \{y_1, y_2, ..., y_{t-1}\}$. Therefore, the entire reward R of generating $\mathbf{Y} = \{y_1, y_2, ..., y_t, ..., y_T\}$ is the sum of r_t:

$$R = \sum_{t=1}^{T} r(\mathbf{Y}_t) - r(\mathbf{Y}_{t-1}) = r(\mathbf{Y}) \tag{8}$$

Then the model is trained to maximize the expected reward $\mathbb{E}_{\mathbf{Y}\sim p_\theta}[r(\mathbf{Y})]$ from the generated report $\mathbf{Y}$ via a sampling strategy (e.g., sampling by probabilities). Based on $\mathbb{E}_{\mathbf{Y}\sim p_\theta}[r(\mathbf{Y})]$, the loss of our entire approach is defined as

$$L(\theta) = -\mathbb{E}_{\mathbf{Y}\sim p_\theta}[r(\mathbf{Y})] \tag{9}$$

with the gradient of $L(\theta)$ for θ computed using the REINFORCE algorithm (Williams, 1992) via

$$\nabla_\theta L(\theta) = -\mathbb{E}_{\mathbf{Y}\sim p_\theta}[r(\mathbf{Y}) \nabla_\theta \log p_\theta(\mathbf{Y})] \tag{10}$$

Then, we approximate the expectation (i.e., the expected gradient) through a single Monte-Carlo sample $\mathbf{Y}$ from p_θ:

$$\nabla_\theta L(\theta) \approx -r(\mathbf{Y}) \nabla_\theta \log p_\theta(\mathbf{Y}) \tag{11}$$

However, the gradient estimated from the above process is of high variance. To maintain the stability of the RL, we follow Rennie et al. (2017a) to reduce such variance by introducing a reference reward b.[4] Therefore, Eq. (10) is formalized as

$$\nabla_\theta L(\theta) = -\mathbb{E}_{\mathbf{Y}\sim p_\theta}[(r(\mathbf{Y}) - b) \nabla_\theta \log p_\theta(\mathbf{Y})] \tag{12}$$

with the expected gradient approximated by

$$\nabla_\theta L(\theta) \approx -(r(\mathbf{Y}) - b) \nabla_\theta \log p_\theta(\mathbf{Y}) \tag{13}$$

Note that, in our approach, b is obtained by computing the NLG metric (e.g., BLEU-4) of the generated report using greedy sampling during inferencing at the training stage. As a result, any actions (i.e., result in some generated $\mathbf{Y}$) that returns higher $r(\mathbf{Y})$ than b drives the following learning process to take as more such actions as possible.

3 Experiment Settings

3.1 Datasets

In our experiments, we use two conventional benchmark datasets, i.e., IU X-RAY (Demner-Fushman et al., 2016)[5] from Indiana University and MIMIC-CXR (Johnson et al., 2019)[6] from the Beth Israel Deaconess Medical Center. The IU X-RAY is a

[4] b is normally a constant (i.e., a reference reward value) obtained from higher rewards by sampling all possible actions. Note that the introduction of b does not change the expected gradient (Eq. (10)), proved by

$$\begin{aligned}\mathbb{E}_{\mathbf{Y}\sim p_\theta}[b \nabla_\theta \log p_\theta(\mathbf{Y})] &= b \sum_{\mathbf{Y}} \nabla_\theta p_\theta(\mathbf{Y}) \\ &= b \nabla_\theta \sum_{\mathbf{Y}} p_\theta(\mathbf{Y}) \\ &= b \nabla_\theta 1 \\ &= 0\end{aligned}$$

[5] https://openi.nlm.nih.gov/

[6] https://physionet.org/content/mimic-cxr/2.0.0/

Dataset	IU X-Ray			MIMIC-CXR		
	Train	Val	Test	Train	Val	Test
Image #	5.2K	0.7K	1.5K	369.0K	3.0K	5.2K
Report #	2.8K	0.4K	0.8K	222.8K	1.8K	3.3K
Patient #	2.8K	0.4K	0.8K	64.6K	0.5K	0.3K
Avg. Len.	37.6	36.8	33.6	53.0	53.1	66.4

Table 1: The statistics of the two benchmark datasets w.r.t. their training, validation and test sets, including the numbers of images, reports and patients, and the averaged word-based length (Avg. Len.) of reports.

Hyper-Parameter	Value
Batch Size	**8**, 10, 16, 32
LR (visual extractor)	1e-5, 3e-5, **5e-5**, 1e-4
LR (encoder-decoder)	5e-5, **1e-4**, 3e-4, 5e-4

Table 2: The hyper-parameters tested in tuning our models, where LR (visual extractor) and LR (encoder-decoder) represent the learning rates for the visual extractor and the encoder-decoder. The bold values illustrate the best hyper-parameter configuration for both IU X-Ray and MIMIC-CXR.

relatively small dataset with 7,470 chest X-ray images and 3,955 corresponding reports; the MIMIC-CXR is the largest public radiography dataset with 473,057 chest X-ray images and 206,563 reports.

Following the experiment settings from previous studies (Li et al., 2018; Jing et al., 2019; Chen et al., 2020), we exclude the samples without reports for both datasets. For IU X-Ray, we use the same split (i.e., 70%/10%/20% for train/validation/test set) as that in Li et al. (2018) and for MIMIC-CXR we adopt its official split. Table 1 show the statistics of all datasets in terms of the numbers of images, reports, patients and the average length of reports with respect to train/validation/test sets.

3.2 Baseline and Evaluation Metrics

To examine our proposed model, we use three baselines for comparison in our experiments. The first, namely **Base**, is the backbone encoder-decoder used in our full model, i.e., a three-layer Transformer model with 8 heads and 512 hidden units without other extensions. The second, namely **Base+RL** is the Transformer model with the same architecture of **Base**, where reinforcement learning is applied to training the model.[7] The third, namely **Base+CMM**, is the Transformer model with the same backbone architecture of **Base** and CMM, without RL.

For evaluation, we follow Chen et al. (2020) to evaluate the above models by two types of metrics, namely, conventional natural language generation (NLG) metrics and clinical efficacy (CE) metrics[8]. The NLG metrics[9] include BLEU (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2011) and ROUGE-L (Lin, 2004). For CE metrics, the CheXpert (Irvin et al., 2019)[10] is applied to label the generated reports and compare the results with ground truths in 14 different categories related to thoracic diseases and support devices. We use precision, recall, and F1 scores to evaluate model performance for CE metrics.

3.3 Implementation Details

To ensure consistency with previous studies (Li et al., 2018; Chen et al., 2020), we use two images for each patient as the input for report generation on IU X-Ray and one image for MIMIC-CXR. For visual extractor, we adopt the ResNet101 (He et al., 2016) pretrained on ImageNet (Deng et al., 2009) to extract patch features with 512 dimensions for each feature. For the encoder-decoder backbone, considering the quality of text representation significantly determines the model performance (Radford et al., 2018; Song and Shi, 2018; Lewis et al., 2020; Song et al., 2021), we use Transformer (Vaswani et al., 2017), which has demonstrated its superior in modeling text in many natural language processing tasks, as the encoder-decoder and randomly initialize its parameters. For the memory matrix in CMM, its dimension and the number of memory vectors $\mathcal{N}$ are set to 512 and 2048, respectively, with random initialization. In addition, the thread number and the $\mathcal{K}$ in CMM are set to 8 and 32, respectively. We train our model using MLE for 30 epochs to regularize the action space before the RL is applied. Afterwards, we start RL using the Adam optimizer (Kingma and Ba, 2015). Table 2 reports the hyper-parameters tested in tuning our models for the two datasets. For each dataset, we try all combinations of the hyper-parameters and use the one achieving the highest BLEU-4 on the validation sets of IU X-Ray and MIMIC-CXR. For example, the best performing learning rates of

[7] This baseline verifies the effectiveness of reinforcement learning on the same structure of **Base** without CMM.

[8] Note that CE metrics only apply to MIMIC-CXR because the labeling schema of CheXpert is designed for MIMIC-CXR, which is different from that of IU X-Ray.

[9] https://github.com/tylin/coco-caption

[10] https://github.com/MIT-LCP/mimic-cxr/tree/master/txt/chexpert

Data	Model	NLG Metrics							CE Metrics		
		BL-1	BL-2	BL-3	BL-4	MTR	RG-L	Avg. Δ	P	R	F1
IU X-Ray	Base	0.396	0.254	0.179	0.135	0.164	0.342	-	-	-	-
	+RL	0.446	0.290	0.212	0.167	0.194	0.356	15.2%	-	-	-
	+CMM	0.474	0.309	0.224	0.173	0.195	0.376	20.6%	-	-	-
	+CMM+RL	**0.494**	**0.321**	**0.235**	**0.181**	**0.201**	**0.384**	**25.2%**	-	-	-
MIMIC-CXR	Base	0.314	0.192	0.127	0.090	0.125	0.265	-	0.331	0.224	0.228
	+RL	0.357	0.219	0.146	0.104	0.139	0.274	12.1%	0.325	0.267	0.271
	+CMM	0.365	0.222	0.147	0.104	0.142	0.272	13.2%	0.329	0.285	0.280
	+CMM+RL	**0.381**	**0.232**	**0.155**	**0.109**	**0.151**	**0.287**	**19.1%**	**0.342**	**0.294**	**0.292**

Table 3: NLG and CE evaluations of different models on the test sets of IU X-Ray and MIMIC-CXR datasets. BL-n denotes BLEU score using up to 4-grams; MTR and RG-L denote METEOR and ROUGE-L, respectively. The average improvement over all NLG metrics compared to Base is also presented in the "Avg. Δ" column.

the visual extractor and other parameters are set to 5×10^{-5} and 1×10^{-4}, respectively, and we decay them by 0.8 per epoch for all datasets.

4 Results and Analysis

4.1 Overall Results

The experimental results of different models on the two benchmark datasets are reported in Table 3 where Base, Base+CMM, and Base+RL represent the aforementioned baselines and Base+CMM+RL represents our full model.

There are several observations. First, all models with CMM consistently outperform Base and Base+RL on both datasets with respect to all NLG metrics, which confirms the advantage of incorporating cross-modal memory into Transformer-based models. Second, the comparison between models with and without RL (i.e., Base vs. Base+RL and Base+CMM vs. Base+CMM+RL) on different metrics confirms the effectiveness of using RL to train such generation model, where models with RL outperforms the ones without RL on all evaluation metrics. This observation indicates that RL has its superiority to map essential features from images and texts with distant (even irrelevant) signals (i.e., NLG metrics) so as to produce better radiology reports. Third, in particular, our full model Base+CMM+RL outperforms all other models by a large margin on both datasets with respect to all metrics, although the other baselines have already achieved outstanding performance, which indicates the effectiveness of the design of reinforced cross-modal alignment. This observation further confirms that, under the RL setting, CMM is able to better search for the cross-modal alignment in the memory representation space without explicit supervision and provides a more accurate feature correspondence in generating high-quality reports.

4.2 Comparison with Previous Studies

We further compare our full model (i.e. Base+CMM+RL) with existing studies on the same datasets, and report the results (in terms of NLG and CE metrics) in Table 4. It is observed that our approach outperforms all previous studies. Particularly, compared with previous studies that also use RL (e.g., HRGR and CMAS-RL), our approach focuses on using RL to leverage the signals from NLG metrics so as to update the whole model, whereas their approaches focus on using RL to improve the decision-making of sentence template utilization and abnormality detection. In addition, compared with Chen et al. (2021b) that uses memory-based approach to align cross-modal information, our approach is able to outperform their approach with the help of the proposed RL mechanism, which demonstrates the effectiveness of our approach to further enhance the cross-modal modeling. Further more, the overall comparison indicates that it is of great potential in exploiting informative patterns among images and their texts for report generation without requiring any external resources, while previous studies (e.g., CoAtt and HRGR) rely on extra information (e.g., private datasets for visual extractor pretraining) for this task.

4.3 Human Evaluation

We employ human evaluation to further evaluate the effect of different modules (i.e., CMM and RL) in our proposed model. In detail, we randomly select 100 chest X-ray images and their ground truth reports from the test set of MIMIC-CXR, as

Data	Model	NLG Metrics						CE Metrics		
		BL-1	BL-2	BL-3	BL-4	MTR	RG-L	P	R	F1
IU X-Ray	ST‡ (Vinyals et al., 2015)	0.216	0.124	0.087	0.066	-	0.306	-	-	-
	Att2in‡ (Rennie et al., 2017b)	0.224	0.129	0.089	0.068	-	0.308	-	-	-
	AdaAtt‡ (Lu et al., 2017)	0.220	0.127	0.089	0.068	-	0.308	-	-	-
	CoAtt‡ (Jing et al., 2018)	0.455	0.288	0.205	0.154	-	0.369	-	-	-
	Hrgr‡ (Li et al., 2018)	0.438	0.298	0.208	0.151	-	0.322	-	-	-
	Cmas-RL‡ (Jing et al., 2019)	0.464	0.301	0.210	0.154	-	0.362	-	-	-
	R2Gen‡ (Chen et al., 2020)	0.470	0.304	0.219	0.165	-	0.371	-	-	-
	CA‡ (Liu et al., 2021c)	0.492	0.314	0.222	0.169	0.193	0.381	-	-	-
	CMCL‡ (Liu et al., 2021a)	0.473	0.305	0.217	0.162	0.186	0.378	-	-	-
	PPKED‡ (Liu et al., 2021b)	0.483	0.315	0.224	0.168	-	0.376	-	-	-
	R2GenCMN‡ (Chen et al., 2021b)	0.475	0.309	0.222	0.170	0.191	0.375	-	-	-
	Ours (CMM+RL)	**0.494**	**0.321**	**0.235**	**0.181**	**0.201**	**0.384**	-	-	-
MIMIC -CXR	ST◇ (Vinyals et al., 2015)	0.299	0.184	0.121	0.084	0.124	0.263	0.249	0.203	0.204
	Att2in◇ (Rennie et al., 2017b)	0.325	0.203	0.136	0.096	0.134	0.276	0.322	0.239	0.249
	AdaAtt◇ (Lu et al., 2017)	0.299	0.185	0.124	0.088	0.118	0.266	0.268	0.186	0.181
	Topdown◇ (Anderson et al., 2018)	0.317	0.195	0.130	0.092	0.128	0.267	0.320	0.231	0.238
	R2Gen‡ (Chen et al., 2020)	0.353	0.218	0.145	0.103	0.142	0.270	0.333	0.273	0.276
	CA‡ (Liu et al., 2021c)	0.350	0.219	0.152	**0.109**	**0.151**	0.283	-	-	-
	CMCL‡ (Liu et al., 2021a)	0.344	0.217	0.140	0.097	0.133	0.281	-	-	-
	PPKED‡ (Liu et al., 2021b)	0.360	0.224	0.149	0.106	0.149	0.284	-	-	-
	R2GenCMN‡ (Chen et al., 2021b)	0.353	0.218	0.148	0.106	0.142	0.278	0.334	0.275	0.278
	Ours (CMM+RL)	**0.381**	**0.232**	**0.155**	**0.109**	**0.151**	**0.287**	**0.342**	**0.294**	**0.292**

Table 4: Comparisons of our proposed models (i.e., CMM+RL) with previous studies on the test sets of IU X-Ray and MIMIC-CXR with respect to NLG and CE metrics. Herein, ‡ marks the results that are directed cited from their paper and ◇ represents the results of our runs with their released codes.

Model	Cor.	Flu.	Cov.	Avg.
Base	5.0	13.0	8.0	8.7
+RL	9.0	23.0	15.0	15.7
+CMM	21.0	30.0	20.0	23.7
+CMM+RL	**65.0**	**34.0**	**57.0**	**52.0**

Table 5: The results of human evaluation for different models. Cor., Flu. and Cov. are abbreviations of correctness, language fluency, and content coverage, respectively, with Avg. denoting the average of them.

well as the reports generated from the baselines and our model. Five human experts who are familiar with radiology are asked to choose the best reports among the generated and the ground truth reports. Following Li et al. (2018), the assessment criterion used in our experiments include correctness, language fluency, and content coverage. The results are reported in Table 5. Overall, Base+CMM+RL outperforms all baselines with a more satisfying result from humans in terms of all the criterion. In particular, Base+CMM+RL significantly outperforms other baselines on correctness and coverage, which further confirms that the reinforced cross-modal alignment helps our approach generate more accurate and comprehensive reports.

4.4 Case Study

To further investigate the effect of our model, we perform a case study on the generated reports from different models (i.e., Base, Base+RL, Base+CMM, and Base+CMM+RL) with an example input chest X-ray image chosen from the test set of MIMIC-CXR. Figure 3 shows the example image with its ground-truth report, and the generation outputs from different models. For each model, we also demonstrate the mappings between regions of the image and words/phrases in the generated text, where the intensity[11] of the mappings is illustrated on the images with different colors.

The observations are drawn from two different aspects. First, Base+CMM and Base+CMM+RL is able to generate descriptions aligned with the ground-truth, which confirms the effectiveness of CMM. For example, for the medical findings in the ground-truth reports (i.e., "*low lung volumes*", "*heart size is mildly enlarged*", "*atelectasis*", and "*vascular congestion*"), Base+CMM and Base+CMM+RL covers many key words in the findings (e.g., "*low lung volumes*", "*heart size*", "*atelectasis*" and "*vascular congestion*") in the gen-

[11] The intensity is measured by the attention scores extracted by the first layer of the decoder.

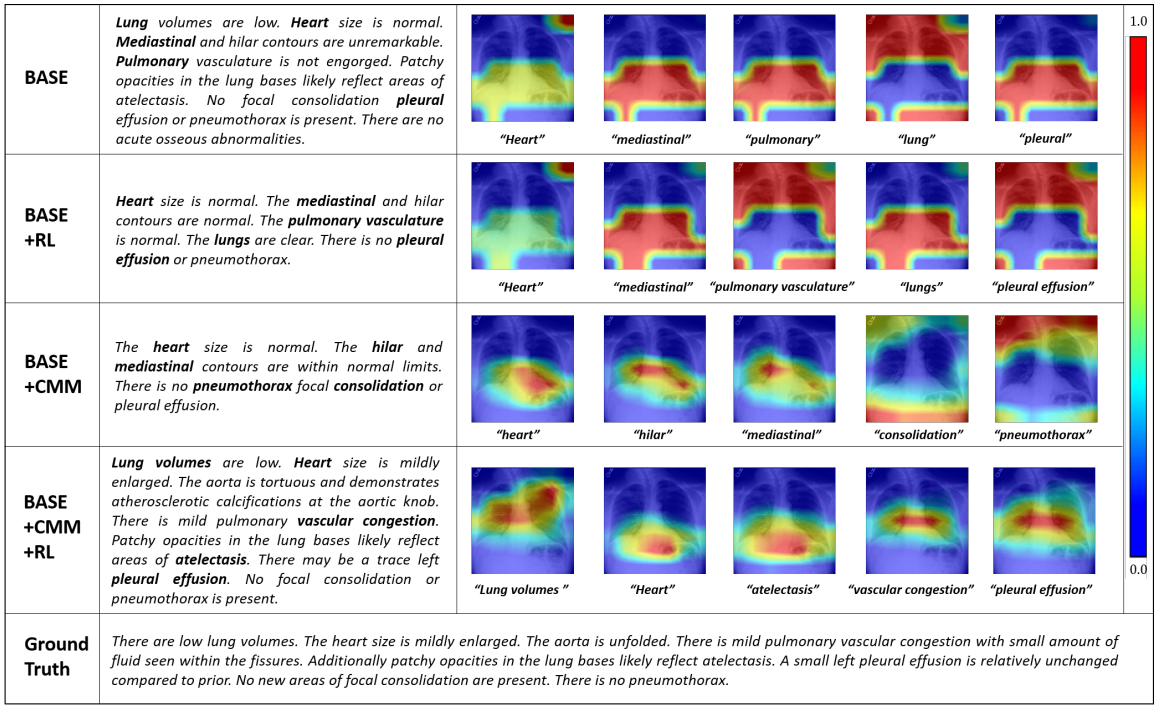

Figure 3: Visualizations of image-text mappings between particular regions (indicated by colored weights) of a chest X-ray image and words/phrases from its reports generated by BASE, BASE+RL, BASE+CMM, , and BASE+CMM+RL, respectively. The color spectrum indicates the value of weight in the range of [0, 1].

erated reports whereas BASE can cover few of them. Second, the validity of RL in aligning the cross-modal features can be observed from the fact that BASE+CMM+RL is able to generate relatively more accurate reports (e.g. "*Heart size is mildly enlarged*") than BASE+CMM (e.g. "*Heart size is normal*") because the former obtains better visual-textual mappings. For example, the abnormality (i.e., "*pleural effusion*") presented in chest X-ray image is covered by the generated report from BASE+CMM+RL and the corresponding region on the image is precisely associated with to the texts.

5 Related Work

The task that is the most relevant to ours is image captioning, which aims to generation text captions that describe the content of the given images (Vinyals et al., 2015; Xu et al., 2015; Anderson et al., 2018; Wang et al., 2019). Being one of its applications and extensions to the medical domain, radiology report generation aims to depicting radiology images with professional texts (Liu et al., 2019; Huang et al., 2019; Miura et al., 2020; Zhang et al., 2020; Alfarghaly et al., 2021; Nooralahzadeh et al., 2021; Najdenkoska et al., 2021; Wang et al., 2021). In general, existing approaches for radiology report generation were mainly designed and proposed to better align images and texts or to exploit highly-patternized features of texts. For example, Jing et al. (2018) proposed a co-attention mechanism to simultaneously explore visual and semantic information with a multi-task learning framework; Li et al. (2018) introduced a template database to incorporate patternized information; Chen et al. (2020) improved generation process by applying a memory-driven Transformer to model patternized information; Chen et al. (2021b) proposes memory-based module to model the cross-modal information. In addition, there are studies that use RL to perform report generation (e.g., Jing et al. (2019) utilized multi-agent RL to capture the imbalanced distribution between abnormality and normality), but their focus is not to utilize RL for text and image alignment. Compared to previous studies, our model offers an effective alternative for radiology reports generation, where a soft intermediate layer with RL is provided to facilitate the mappings between visual and textual features, which allows one to produce more accurate descriptions for radiology images.

6 Conclusion

In this paper, we propose an RL approach based on CMM to better align visual and textual features for radiology report generation. In detail, a shared memory is used to store the cross-modal informa-

tion and RL is applied to leverage the signals from NLG metrics to guide cross-modal mappings so as to better link features from images and texts. The experimental results on two benchmark datasets (i.e., IU X-Ray and MIMIC-XCR) demonstrate the effectiveness of our model, which achieves the state-of-the-art performance on both datasets. Human evaluation and the case study further confirm that our approach is able to generate high-quality reports with meaningful image-text alignment.

Acknowledgements

This work is supported by Shenzhen Science and Technology Program under the project "Fundamental Algorithms of Natural Language Understanding for Chinese Medical Text Processing" (JCYJ20210324130208022) and the Natural Science Foundation of Guangdong Province, China, under the project "Deep Learning based Chinese Combination Category Grammar Parsing and its Application in Relation Extraction". It is also supported by Shenzhen Institute of Artificial Intelligence and Robotics for Society under the project "Automatic Knowledge Enhanced Natural Language Understanding and Its Applications" (AC01202101001).

References

Omar Alfarghaly, Rana Khaled, Abeer Elkorany, Maha Helal, and Aly Fahmy. 2021. Automated radiology report generation using conditioned transformers. *Informatics in Medicine Unlocked*, 24:100557.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 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*, pages 6077–6086.

Guimin Chen, Yuanhe Tian, Yan Song, and Xiang Wan. 2021a. Relation Extraction with Type-aware Map Memories of Word Dependencies. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*.

Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. 2021b. 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)*, pages 5904–5914, Online.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020. Generating Radiology Reports via Memory-driven Transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1439–1449.

Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-memory Transformer for Image Captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10578–10587.

Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. In *Proceedings of the sixth workshop on statistical machine translation*, pages 85–91.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Xin Huang, Fengqi Yan, Wei Xu, and Maozhen Li. 2019. Multi-attention and Incorporating Background Information Model for Chest X-ray Image Report Generation. *IEEE Access*, 7:154808–154817.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. 2019. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597.

Baoyu Jing, Zeya Wang, and Eric Xing. 2019. Show, Describe and Conclude: On Exploiting the Structure Information of Chest X-ray Reports. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6570–6580.

Baoyu Jing, Pengtao Xie, and Eric Xing. 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)*, pages 2577–2586.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. MIMIC-CXR: A large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. *CoRR*, abs/1412.6980.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online.

Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. 2018. Hybrid Retrieval-Generation Reinforced Agent for Medical Image Report Generation. In *Advances in neural information processing systems*, pages 1530–1540.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81.

Fenglin Liu, Shen Ge, and Xian Wu. 2021a. Competence-based multimodal curriculum learning for medical report generation. In *ACL*, volume 1, page 3.

Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. 2021b. Exploring and Distilling Posterior and Prior Knowledge for Radiology Report Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13753–13762.

Fenglin Liu, Changchang Yin, Xian Wu, Shen Ge, Ping Zhang, and Xu Sun. 2021c. Contrastive attention for automatic chest x-ray report generation. *arXiv preprint arXiv:2106.06965*.

Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. 2019. Clinically Accurate Chest X-Ray Report Generation. In *Machine Learning for Healthcare Conference*, pages 249–269.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. 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*, pages 375–383.

Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. 2016. Key-Value Memory Networks for Directly Reading Documents. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1400–1409.

Yasuhide Miura, Yuhao Zhang, Emily Bao Tsai, Curtis P Langlotz, and Dan Jurafsky. 2020. Improving factual completeness and consistency of image-to-text radiology report generation. *arXiv preprint arXiv:2010.10042*.

Ivona Najdenkoska, Xiantong Zhen, Marcel Worring, and Ling Shao. 2021. Variational topic inference for chest x-ray report generation. *arXiv preprint arXiv:2107.07314*.

Yuyang Nie, Yuanhe Tian, Yan Song, Xiang Ao, and Xiang Wan. 2020. Improving Named Entity Recognition with Attentive Ensemble of Syntactic Information. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4231–4245.

Farhad Nooralahzadeh, Nicolas Perez Gonzalez, Thomas Frauenfelder, Koji Fujimoto, and Michael Krauthammer. 2021. Progressive Transformer-Based Generation of Radiology Reports. *arXiv preprint arXiv:2102.09777*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-training.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017a. Self-critical Sequence Training for Image Captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7008–7024.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017b. Self-critical Sequence Training for Image Captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7008–7024.

Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR*, abs/1409.1556.

Yan Song and Shuming Shi. 2018. Complementary Learning of Word Embeddings. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4368–4374.

Yan Song, Tong Zhang, Yonggang Wang, and Kai-Fu Lee. 2021. ZEN 2.0: Continue Training and Adaption for N-gram Enhanced Text Encoders. *arXiv preprint arXiv:2105.01279*.

Yuanhe Tian, Wang Shen, Yan Song, Fei Xia, Min He, and Kenli Li. 2020. Improving Biomedical Named Entity Recognition with Syntactic Information. *BMC Bioinformatics*, 21:1471–2105.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and Tell: A Neural Image Caption Generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Weixuan Wang, Zhihong Chen, and Haifeng Hu. 2019. Hierarchical Attention Network for Image Captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8957–8964.

Yixin Wang, Zihao Lin, Jiang Tian, Yang Zhang, Jianping Fan, Zhiqiang He, et al. 2021. Confidence-guided radiology report generation. *arXiv preprint arXiv:2106.10887*.

Ronald J Williams. 1992. Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning. *Machine learning*, 8(3-4):229–256.

Jie Wu, Ian Harris, and Hongzhi Zhao. 2021. Spoken Language Understanding for Task-oriented Dialogue Systems with Augmented Memory Networks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 797–806, Online.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *International conference on machine learning*, pages 2048–2057.

Kun Xu, Yuxuan Lai, Yansong Feng, and Zhiguo Wang. 2019. Enhancing Key-Value Memory Neural Networks for Knowledge Based Question Answering. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2937–2947, Minneapolis, Minnesota.

Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan Yuille, and Daguang Xu. 2020. When radiology report generation meets knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12910–12917.