Generating Radiology Reports via Memory-driven Transformer

Zhihong Chen♠♡,   Yan Song♠♡†,   Tsung-Hui Chang♠♡,   Xiang Wan♡

♠The Chinese University of Hong Kong (Shenzhen)
♡Shenzhen Research Institute of Big Data
♠zhihongchen@link.cuhk.edu.cn
♠{songyan, changtsunghui}@cuhk.edu.cn
♡wanxiang@sribd.cn

Abstract

Medical imaging is frequently used in clinical practice and trials for diagnosis and treatment. Writing imaging reports is time-consuming and can be error-prone for inexperienced radiologists. Therefore, automatically generating radiology reports is highly desired to lighten the workload of radiologists and accordingly promote clinical automation, which is an essential task to apply artificial intelligence to the medical domain. In this paper, we propose to generate radiology reports with memory-driven Transformer, where a relational memory is designed to record key information of the generation process and a memory-driven conditional layer normalization is applied to incorporating the memory into the decoder of Transformer. Experimental results on two prevailing radiology report datasets, IU X-Ray and MIMIC-CXR, show that our proposed approach outperforms previous models with respect to both language generation metrics and clinical evaluations. Particularly, this is the first work reporting the generation results on MIMIC-CXR to the best of our knowledge. Further analyses also demonstrate that our approach is able to generate long reports with necessary medical terms as well as meaningful image-text attention mappings.[1]

1 Introduction

Radiology report generation, which aims to automatically generate a free-text description for a clinical radiograph (e.g., chest X-ray), has emerged as a prominent attractive research direction in both artificial intelligence and clinical medicine. It can greatly expedite the automation of workflows and improve the quality and standardization of health care. Recently, there are many methods proposed

---

†Corresponding author.
[1]Our code and the best performing models are released at https://github.com/zhjohnchan/R2Gen.
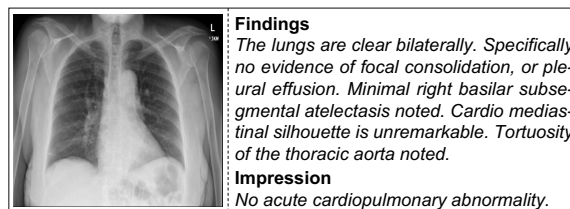

Figure 1: An example chest X-ray image and its report including findings and impression.

in this area (Jing et al., 2018; Li et al., 2018; Johnson et al., 2019; Liu et al., 2019; Jing et al., 2019).

Practically, a significant challenge of radiology report generation is that radiology reports are long narratives consisting of multiple sentences. As illustrated by Figure 1, a radiology report generally consists of a section of findings which describes medical observations, including both normal and abnormal features, as well as an impression or concluding remark summarizing the most prominent observations. Therefore, applying conventional image captioning approaches (Vinyals et al., 2015; Anderson et al., 2018) may be insufficient for radiology report generation, as such approaches are designed to briefly describe visual scenes with short sentences. The ability to provide accurate clinical descriptions for a radiograph is of the highest priority, which places a higher demand on the generation process. Nevertheless, despite the difficulties posed by these evident length and accuracy requirements, radiology reports do have their own distinctive characteristics. An important feature to note is their highly patternized nature, as illustrated by the sample report described above (Figure 1). On the basis of this patternization, many approaches have been proposed to address the challenges of radiology report generation. For example, Liu et al. (2019) found that a simple retrieval-based method could achieve a comparative performance for this task. Li et al. (2018) combined retrieval-based and generation-based methods with manually extracted

templates. Although promising results may be obtained by the retrieval-based approaches, they are still limited in the preparation of large databases, or the explicit construction of template lists to determine the patterns embedded in various reports.

In this paper, we propose to generate radiology reports via memory-driven Transformer. In detail, a relational memory (RM) is proposed to record the information from previous generation processes and a novel memory-driven conditional layer normalization (MCLN) is designed to incorporate the relational memory into Transformer (Vaswani et al., 2017). As a result, similar patterns in different medical reports can be implicitly modeled and memorized during the generation process, which thereby can facilitate the decoding of Transformer and is capable of generating long reports with informative content. Experimental results on two benchmark datasets confirm the validity and effectiveness of our approach, where Transformer with RM and MCLN achieves the state-of-the-art performance on all datasets. To summarize, the contributions of this paper are four-fold:

- We propose to generate radiology reports via a novel memory-driven Transformer model.
- We propose a relational memory to record the previous generation process and the MCLN to incorporate relational memory into layers in the decoder of Transformer.
- Extensive experiments are performed and the results show that our proposed models outperform the baselines and existing models.
- We conduct analyses to investigate the effect of our model with respect to different memory sizes and show that our model is able to generate long reports with necessary medical terms and meaningful image-text attention mappings.

2 The Proposed Method

Generating radiology reports is essentially an image-to-text generation task, for which there exist several solutions (Vinyals et al., 2015; Xu et al., 2015; Anderson et al., 2018; Cornia et al., 2019).

We follow the standard sequence-to-sequence paradigm for this task. In doing so, we treat the input from a radiology image as the source sequence $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_S\}, \mathbf{x}_s \in \mathbb{R}^d$, where $\mathbf{x}_s$ are patch features extracted from visual extractors and d the size of the feature vector. The corresponding report is the target sequence $Y = \{y_1, y_2, ..., y_T\}, y_t \in \mathbb{V}$, where y_t are the generated

tokens, T the length of generated tokens and $\mathbb{V}$ the vocabulary of all possible tokens. An overview of our proposed model is shown in Figure 2, where the details are illustrated in following subsections.

2.1 The Model Structure

Our model can be partitioned into three major components, i.e., the visual extractor, the encoder and the decoder, where the proposed memory and the integration of the memory into Transformer are mainly performed in the decoder. The overall description of the three components and the training objective of the task is detailed below.

Visual Extractor Given a radiology image Img, its visual features $\mathbf{X}$ are extracted by pre-trained convolutional neural networks (CNN), e.g., VGG (Simonyan and Zisserman, 2015) or ResNet (He et al., 2016), and the encoded results are used as the source sequence for all subsequent modules. The process is formulated as:

$$\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_S\} = f_v(Img) \qquad (1)$$

where $f_v(\cdot)$ represents the visual extractor.

Encoder In our model, we use the standard encoder from Transformer, where the outputs are the hidden states $\mathbf{h}_i$ encoded from the input features $\mathbf{x}_i$ extracted from the visual extractor:

$$\{\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_S\} = f_e(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_S) \qquad (2)$$

where $f_e(\cdot)$ refers to the encoder.

Decoder The backbone decoder in our model is the one from Transformer, where we introduce an extra memory module to it by improving the original layer normalization with MCLN for each decoding layer as shown in Figure 2. Therefore the decoding process can be formalized as

$$y_t = f_d(\mathbf{h}_1, ..., \mathbf{h}_S, \text{MCLN}(\text{RM}(y_1, ..., y_{t-1}))) \qquad (3)$$

where $f_d(\cdot)$ refers to the decoder and the details of the memory (RM) and MCLN are presented in following subsections.

Objective Given the aforementioned structure, the entire generation process can be formalized as a recursive application of the chain rule

$$p(Y|Img) = \prod_{t=1}^{T} p(y_t|y_1, ..., y_{t-1}, Img) \qquad (4)$$

where $Y = \{y_1, y_2, ..., y_T\}$ is the target text sequence. The model is then trained to maximize
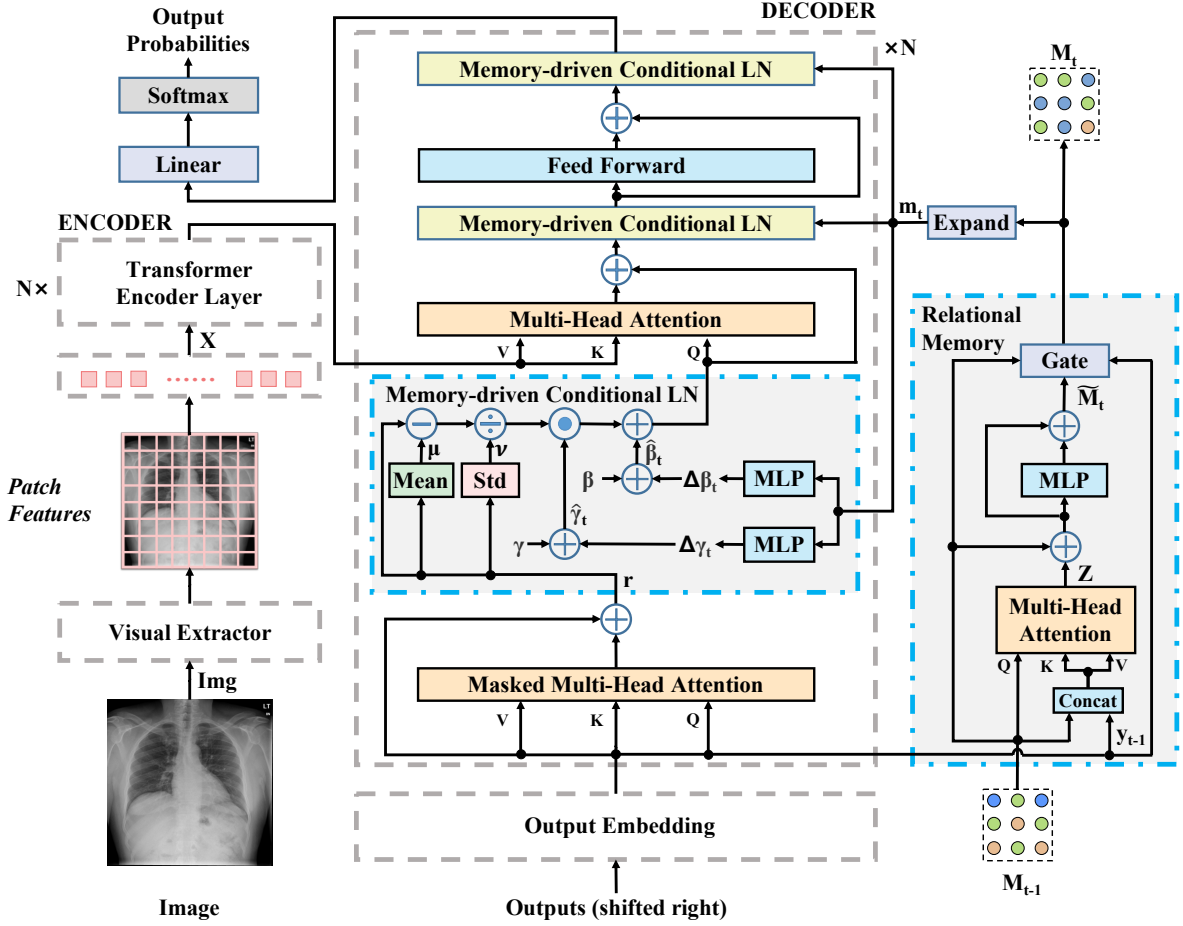
Figure 2: The overall architecture of our proposed model, where the visual extractor, encoder and decoder are shown in gray dash boxes and the details of the visual extractor and encoder are omitted. The relational memory and memory conditional layer-normalization are illustrated in grey solid boxes with blue dash lines.

$P(Y|Img)$ through the negative conditional log-likelihood of Y given the Img:

$$\theta^* = \arg\max_{\theta} \sum_{t=1}^{T} \log p(y_t|y_1, ..., y_{t-1}, Img; \theta)$$

(5)

where θ is the parameters of the model.

2.2 Relational Memory

For any relevant Img, they may share similar patterns in their reports and they can be used as good references for each other to help the generation process. As shown in Figure 1, patterns such as "*The lungs are clear bilaterally*" and "*no evidence of focal consolidation, or pleural effusion*" always appear in the reports of similar images and are shown simultaneously. To exploit such characteristics, we propose to use an extra component, i.e., relational memory, to enhance Transformer to learn from the patterns and facilitate computing the interactions among patterns and the generation process.

In doing so, the relational memory uses a matrix

to transfer its states over generation steps, where the states record important pattern information with each row (namely, memory slot) representing some pattern information.[2] During the generation, the matrix is updated step-by-step with incorporating the output from previous steps. Then, at time step t, the matrix from the previous step, M_{t-1}, is functionalized as the query and its concatenations with the previous output serve as the key and value to feed the multi-head attention module. Given H heads used in Transformer, there are H sets of queries, keys and values via three linear transformations, respectively. For each head, we obtain the query, key and value in the relational memory through $\mathbf{Q} = \mathbf{M}_{t-1} \cdot \mathbf{W_q}$, $\mathbf{K} = [\mathbf{M}_{t-1}; \mathbf{y}_{t-1}] \cdot \mathbf{W_k}$ and $\mathbf{V} = [\mathbf{M}_{t-1}; \mathbf{y}_{t-1}] \cdot \mathbf{W_v}$, respectively, where $\mathbf{y}_{t-1}$ is the embedding of the last output (at step $t-1$); $[\mathbf{M_{t-1}}; \mathbf{y}_{t-1}]$ is the row-wise concatenation of $\mathbf{M}_{t-1}$ and $\mathbf{y}_{t-1}$. $\mathbf{W_q}$, $\mathbf{W_k}$ and $\mathbf{W_v}$ are

---

[2]Note that the rows (memory slots) and patterns do not follow one-to-one mapping, where the entire matrix serves as a whole unit to deliver the pattern information.
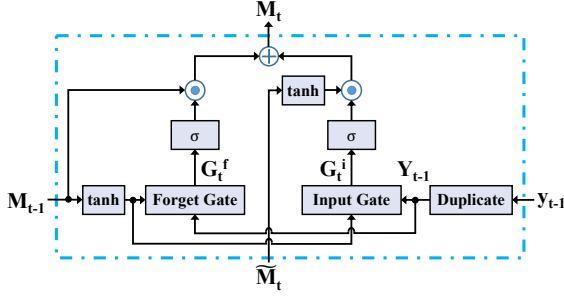
Figure 3: The illustration of the gate mechanism.

Dataset	IU X-Ray			MIMIC-CXR		
	Train	Val	Test	Train	Val	Test
Image #	5,226	748	1,496	368,960	2,991	5,159
Report #	2,770	395	790	222,758	1,808	3,269
Patient #	2,770	395	790	64,586	500	293
Avg. Len.	37.56	36.78	33.62	53.00	53.05	66.40

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 average word-based length (Avg. Len.) of reports.

the trainable weights of linear transformation of the query, key and value, respectively. Multi-head attention is used to model $\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$ so as to depict relations of different patterns. As a result,

$$\mathbf{Z} = \text{softmax}(\mathbf{Q}\mathbf{K}^\top / \sqrt{d_k}) \cdot \mathbf{V} \qquad (6)$$

where d_k is the dimension of $\mathbf{K}$, and $\mathbf{Z}$ the output of the multi-head attention module. Consider that the relational memory is performed in a recurrent manner along with the decoding process, it potentially suffers from gradient vanishing and exploding. We therefore introduce residual connections and a gate mechanism. The former is formulated as

$$\tilde{\mathbf{M}}_t = f_{mlp}(\mathbf{Z} + \mathbf{M}_{t-1}) + \mathbf{Z} + \mathbf{M}_{t-1} \qquad (7)$$

where $f_{mlp}(\cdot)$ refers to the multi-layer perceptron (MLP). The detailed structure of the gate mechanism in the relational memory is shown in Figure 3, where the forget and input gates are applied to balance the inputs from $\mathbf{M}_{t-1}$ and $\mathbf{y}_{t-1}$, respectively. To ensure that $\mathbf{y}_{t-1}$ can be used for computation with $\mathbf{M}_{t-1}$, it is extended to a matrix $\mathbf{Y}_{t-1}$ by duplicating it to multiple rows. Therefore, the forget and input gate are formalized as

$$\mathbf{G}_t^f = \mathbf{Y}_{t-1}\mathbf{W}^f + \tanh(\mathbf{M}_{t-1}) \cdot \mathbf{U}^f \qquad (8)$$

$$\mathbf{G}_t^i = \mathbf{Y}_{t-1}\mathbf{W}^i + \tanh(\mathbf{M}_{t-1}) \cdot \mathbf{U}^i \qquad (9)$$

where $\mathbf{W}^f$ and $\mathbf{W}^i$ are trainable weights for $\mathbf{Y}_{t-1}$ in each gate; similarly, $\mathbf{U}^f$ and $\mathbf{U}^i$ are the trainable weights for $\mathbf{M}_{t-1}$ in each gate. The final output of the gate mechanism is formalized as

$$\mathbf{M}_t = \sigma(\mathbf{G}_t^f) \odot \mathbf{M}_{t-1} + \sigma(\mathbf{G}_t^i) \odot \tanh(\tilde{\mathbf{M}}_t) \qquad (10)$$

where $\odot$ refers to the Hadamard product and σ the sigmoid function and $\mathbf{M}_t$ is the output of the entire relational memory module at step t.

2.3 Memory-driven Conditional Layer Normalization

Although memory shows its effectiveness in many NLP tasks (Sukhbaatar et al., 2015; Lample et al.,

2019), it is by default applied to encoding with rather isolated designs. However, given that text generation is a dynamic process and largely affected by the output at each decoding step, memory is expected to be closely integrated to the decoder.

Therefore, we propose a novel MCLN and use it to incorporate the relational memory to enhance the decoding of Transformer. Recall that in the conventional Transformer, to improve generalization, γ and β are two crucial parameters for scaling and shifting the learned representations,[3] respectively. Thus we propose to incorporate the relational memory via MCLN by feeding its output $\mathbf{M}_t$ to γ and β. Consequently, this design takes the benefit from the memory while preventing it from influencing too many parameters of Transformer so that some core information for generation is not affected.

As shown in Figure 2, in each Transformer decoding layer, we use three MCLNs, where the output of the first MCLN is functionalized as the query to be fed into the following multi-head attention module together with the hidden states from the encoder as the key and value. To feed each MCLN, at step t, the output of the relational memory $\mathbf{M}_t$ is expanded into a vector $\mathbf{m}_t$ by simply concatenating all rows from $\mathbf{M}_t$. Then, an MLP is used to predict a change $\Delta\gamma_t$ on γ_t from $\mathbf{m}_t$, and update it via

$$\Delta\gamma_t = f_{mlp}(\mathbf{m}_t) \qquad (11)$$

$$\hat{\gamma}_t = \gamma + \Delta\gamma_t \qquad (12)$$

Similarly, $\Delta\beta_t$ and $\hat{\beta}_t$ are performed by

$$\Delta\beta_t = f_{mlp}(\mathbf{m}_t) \qquad (13)$$

$$\hat{\beta}_t = \beta + \Delta\beta_t \qquad (14)$$

Afterwards, the predicted $\hat{\beta}_t$ and $\hat{\gamma}_t$ are applied to the mean and variance results of the multi-head

---

[3]In detail, γ is used to amplify the values in the learned representation and β provides a bias adjustment to them.

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	-	-	-	-
	+RM	0.444	0.283	0.196	0.141	0.179	0.364	8.9%	-	-	-
	+RM+MCLN	**0.470**	**0.304**	**0.219**	**0.165**	**0.187**	**0.371**	**17.6%**	-	-	-
MIMIC-CXR	BASE	0.314	0.192	0.127	0.090	0.125	0.265	-	0.331	0.224	0.228
	+RM	0.330	0.200	0.133	0.095	0.128	0.265	3.7%	0.325	0.243	0.249
	+RM+MCLN	**0.353**	**0.218**	**0.145**	**0.103**	**0.142**	**0.277**	**12.1%**	**0.333**	**0.273**	**0.276**

Table 2: The performance of all baselines and our full model on the test sets of IU X-RAY and MIMIC-CXR datasets with respect to NLG and CE metrics. BL-n denotes BLEU score using up to n-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 performance of all models is averaged from five runs.

self-attention from the previous generated outputs:

$$f_{mcln}(\mathbf{r}) = \hat{\gamma}_t \odot \frac{\mathbf{r} - \mu}{\upsilon} + \hat{\beta}_t \qquad (15)$$

where $\mathbf{r}$ refers to the output from the previous module; μ and υ are the mean and standard deviation of $\mathbf{r}$, respectively. The result $f_{mcln}(\mathbf{r})$ from MCLN is then fed to the next module (for the 1st and 2nd MCLN) or used as the final output for generation (for the 3rd MCLN).

3 Experiment Settings

3.1 Datasets

We conduct our experiments on two datasets, which are described as follows:

- **IU X-RAY** (Demner-Fushman et al., 2016)[4]: a public radiography dataset collected by Indiana University with 7,470 chest X-ray images and 3,955 reports.
- **MIMIC-CXR** (Johnson et al., 2019)[5]: the largest radiology dataset to date that consists of 473,057 chest X-ray images and 206,563 reports from 63,478 patients.

For both datasets, we follow Li et al. (2018) to exclude the samples without reports. Then we apply their conventional splits. Specifically, IU X-RAY is partitioned into train/validation/test set by 7:1:2 of the entire dataset, and MIMIC-CXR's official split is adopted. The statistics of the datasets are shown in Table 1, with the numbers of images, reports, patients and the average length of reports.

3.2 Baseline and Evaluation Metrics

To compare with our proposed model, the following ones are used as the main baselines:

- **BASE**: this is the vanilla Transformer, with three layers, 8 heads and 512 hidden units without other extensions and modifications.
- **BASE+RM**: this is a simple alternative of our proposed model where the relational memory is directly concatenated to the output of the Transformer ahead of the softmax at each time step. This baseline aims to demonstrate the effect of using memory as an extra component instead of integration within the Transformer.

In addition, we also compare our model with those in previous studies, including conventional image captioning models, e.g., **ST** (Vinyals et al., 2015), **ATT2IN** (Rennie et al., 2017), **ADAATT** (Lu et al., 2017), **TOPDOWN** (Anderson et al., 2018), and the ones proposed for the medical domain, e.g., **COATT** (Jing et al., 2018), **HRGR** (Li et al., 2018) and **CMAS-RL** (Jing et al., 2019).

The performance of the aforementioned models is evaluated by conventional natural language generation (NLG) metrics and clinical efficacy (CE) metrics[6]. The NLG metrics[7] include BLEU (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2011) and ROUGE-L (Lin, 2004). For clinical efficacy metrics, we use the CheXpert (Irvin et al., 2019)[8] to label the generated reports and compare the results with ground truths in 14 different categories related to thoracic diseases and support devices. Precision, recall and F1 are used to evaluate model performance for these metrics.

DATA	MODEL	NLG METRICS						CE METRICS		
		BL-1	BL-2	BL-3	BL-4	MTR	RG-L	P	R	F1
IU X-RAY	ST♮	0.216	0.124	0.087	0.066	-	0.306	-	-	-
	ATT2IN♮	0.224	0.129	0.089	0.068	-	0.308	-	-	-
	ADAATT♮	0.220	0.127	0.089	0.068	-	0.308	-	-	-
	COATT♮	0.455	0.288	0.205	0.154	-	0.369	-	-	-
	HRGR♮	0.438	0.298	0.208	0.151	-	0.322	-	-	-
	CMAS-RL♮	0.464	0.301	0.210	0.154	-	0.362	-	-	-
	OURS	**0.470**	**0.304**	**0.219**	**0.165**	0.187	**0.371**	-	-	-
MIMIC-CXR	ST♯	0.299	0.184	0.121	0.084	0.124	0.263	0.249	0.203	0.204
	ATT2IN♯	0.325	0.203	0.136	0.096	0.134	0.276	0.322	0.239	0.249
	ADAATT♯	0.299	0.185	0.124	0.088	0.118	0.266	0.268	0.186	0.181
	TOPDOWN♯	0.317	0.195	0.130	0.092	0.128	0.267	0.320	0.231	0.238
	OURS	**0.353**	**0.218**	**0.145**	**0.103**	**0.142**	**0.277**	**0.333**	**0.273**	**0.276**

Table 3: Comparisons of our full model with previous studies on the test sets of IU X-RAY and MIMIC-CXR with respect to NLG and CE metrics. ♮ refers to that the result is directed cited from the original paper and ♯ represents our replicated results by their codes.

3.3 Implementation Details

We adopt the ResNet101 (He et al., 2016) pre-trained on Imagenet (Deng et al., 2009) as the visual extractor to extract patch features with the dimension of each feature set to 2,048. Note that for IU X-RAY, we use two images of a patient as input to ensure consistency with the experiment settings of previous work. The Transformer in our proposed model and all baselines are randomly initialized. For relational memory, its dimension and the number of heads in multi-head attention are set to 512 and 8, respectively, and the number of memory slots is set to 3 by default. For MCLN, we use two MLPs to obtain $\Delta\gamma$ and $\Delta\beta$ where they do not share parameters. The model is trained under cross entropy loss with ADAM optimizer (Kingma and Ba, 2015). We set the learning rate to 5e-5 and 1e-4 for the visual extractor and other parameters, respectively. We decay such rate by a factor of 0.8 per epoch for each dataset and set the beam size to 3 to balance the generation effectiveness and efficiency. Note that the aforementioned hyper-parameters are obtained by evaluating the models on the validation sets of the two datasets.

4 Results and Analyses

4.1 Effect of Relational Memory

To illustrate the effectiveness of our proposed method, we experiment with the aforementioned baselines on the two benchmark datasets. The results are reported in Table 2, with BASE+RM+MCLN representing our full model (same below).

There are several observations. First, on NLG metrics, both BASE+RM and BASE+RM+MCLN outperform the vanilla Transformer (BASE) on both datasets, which confirms the validity of incorporating memory into the decoding process in Transformer because that highly-patternized text in radiology reports are reasonably modeled to some extent. Second, our full model achieves the best performance over all baselines on different metrics, and it particularly outperforms BASE+RM with significant improvement, which clearly indicates the usefulness of MCLN in incorporating memory rather than other ways of integration. Third, on NLG metrics, when comparing between the datasets, the performance gains from two memory-driven models (i.e., BASE+RM and BASE+RM+MCLN) over BASE on IU X-RAY are larger than that of MIMIC-CXR. The reason behind might be that the IU X-RAY is relatively small and patterns among different reports in this dataset are more consistent so that our model helps more with the proposed memory. Fourthly, on the CE metrics on MIMIC-CXR, our full model shows the same trend as that for NLG metrics, where it outperforms all its baselines in terms of precision, recall and F1. This observation is important because higher NLG scores do not always result in higher clinical scores (e.g., the precision of BASE+RM on CE is lower than that of BASE), so

$|\mathcal{S}|$	PARA.	BL-1	BL-2	MTR	RG-L
1	76.6M	0.350	0.217	0.141	0.278
2	81.4M	0.355	0.215	0.141	0.278
3	86.1M	**0.360**	**0.223**	**0.144**	0.279
4	90.8M	0.354	0.217	0.142	**0.280**

Table 4: NLG scores of our full model on the MIMIC-CXR test set when different memory slots are used. PARA. denotes the number of parameters.

that the performance from CE further confirms the effectiveness of our method, whereas compared to BASE+RM, MCLN is able to leverage memory in a rather fine-grained way and thus better produce reasonable descriptions for clinical abnormalities.

4.2 Comparison with Previous Studies

We compare our full model (denoted as OURS) with existing models on the same datasets, with all results reported in Table 3 on both NLG and CE metrics. There are several observations drawn from different aspects. First, Transformer confirms its superiority to sequence-to-sequence structures in this task, which is illustrated by the comparison between our models (all baselines and our full model) and ST. Our full model also outperforms conventional image captioning models, e.g., ATT2IN, ADAATT and TOPDOWN, which are designed to generate a short piece of text for an image. This observation confirms that designing a specific model for long report generation is necessary for this task. Second, memory shows its effectiveness in this task when compared with those complicated models, e.g., HRGR uses manually extracted templates. Particularly, although on the two datasets, reinforcement learning (CMAS-RL) is proved to be the best solution with a careful design of adaptive rewards, our model achieves the same goal with a simpler method. Third, It is noticed that there are studies, e.g., HRGR, requires to utilize extra information for this task and our full model outperforms them without such requirements. This observation indicates that an appropriate end-to-end design (such as RM and MCLN) of using memory in Transformer can alleviate the need for extra resources to enhance this task.

4.3 Analysis

We analyze several aspects of our model regarding its hyper-parameters and generation results.
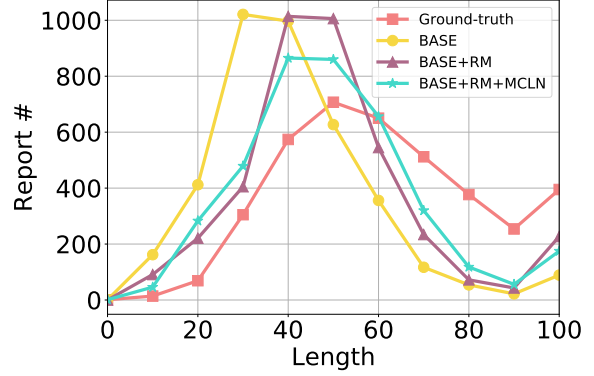

Figure 4: The length distributions of the generated reports on the MIMIC-CXR test set from BASE, BASE+RM and BASE+RM+MCLN, as well as the ground-truth.

Memory Size To show the impacts of the memory size, we train RM with different numbers of memory slots, i.e., $|\mathcal{S}| \in \{1, 2, 3, 4\}$ and the results on MIMIC-CXR are shown in Table 4. In general, since memory size controls how much information is preserved in the past generation steps, it is confirmed in the observation that enlarging memory size by the number of slots results in better overall performance, with $|\mathcal{S}| = 3$ achieving the best results. Still, we notice that the overall performance drops when $|\mathcal{S}| = 4$, which indicates that too large memory may introduce redundant and invalid information so as to negatively affect the generation process. Although enlarging memory size results in increasing parameter numbers, it is demonstrated that there are not too many parameters (comparing to the total number of parameters) introduced whenever adding one slot in the memory. This observation suggests that the proposed model is effective and efficient in learning with memory for the radiology report generation task.

Report Length In addition to NLG and CE metrics, another important criterion to evaluate generation models is the length of generated reports comparing to the ground-truth. In doing so, we categorize all reports generated on the MIMIC-CXR test set into 10 groups (within [0, 100] with interval of 10) according to their round-down lengths and draw curves for their numbers in each category for BASE, BASE+RM and BASE+RM+MCLN, as well as the ground-truth. The results are presented in Figure 4. Overall, more reports generated from BASE+RM and BASE+RM+MCLN are longer than that from BASE and their length distributions are closer to the ground-truth reports, which thus leads to better evaluation results on NLG metrics. The reason behind might be that the memory provides
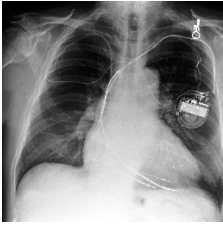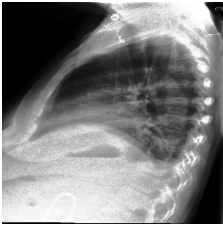
Figure 5: Illustrations of reports from ground-truth, BASE and BASE+RM+MCLN models for two X-ray chest images. To better distinguish the content in the reports, different colors highlight different medical terms.

more detailed information for the generation process so that the decoder tends to produce more diversified outputs than the original Transformer. Particularly, when comparing BASE+RM+MCLN and BASE+RM, the length distribution of the former generated reports is closer to the ground-truth, which can be explained by that, instead of applying memory to the final output, leveraging memory at each layer in Transformer is more helpful and thus controls the decoding process in a fine-grained way. The above observations show that both memory and the way of using it are two important factors to enhance radiology report generation.

Case Study To further investigate the effectiveness of our model, we perform qualitative analysis on some cases with their ground-truth and generated reports from different models. Figure 5 shows two examples of front and lateral chest X-ray images from MIMIC-CXR and such reports, where different colors on the texts indicate different medical terms. It is observed in these cases that BASE+RM+MCLN is able to generate descriptions aligned with that written by radiologists with similar content flow. For example, in both cases, patterns in the generated reports follow the structure that starting from reporting abnormal findings (e.g., "*cardiac silhouette*" and "*lung volumes*"), and then concluding with potential diseases (e.g., "*pleural effusion*" and "*atelectasis*"). In addition, for the necessary medical terms in the ground-truth reports, BASE+RM+MCLN covers almost all of them in its generated reports while vanilla Transformer did much worse, e.g., the key terms "*enlarged cardiac silhouette*", "*atelectasis*" and "*small pleural effusion*" in the two examples are not generated.

To further investigate different models qualitatively, we randomly select a chest X-ray on the MIMIC-CXR test set and visualize the image-text attention mappings from BASE and BASE+RM+MCLN. Figure 6 shows the intermediate image-text correspondences for several words from the multi-head attentions in the first layer of the decoders. It is observed that BASE+RM+MCLN is better at aligning the locations with the indicated disease or parts. This observation suggests that our model not only enhances the power of radiology report generation, but also improves the interaction between the images and the generated texts.

Error Analysis To analyze the errors from our model, especially in targeting the low CE scores, it is found that the class imbalance is severe on the datasets and affects the model training and inference, where majority voting is observed in the generation process. For example, on MIMIC-CXR, consolidation only accounts for 3.9% in the training set so that the trained model only recognizes that 2.9% results in this case compared with the ground truth 6.3%. Thus how to address the data bias problem is a possible future work to improve the accuracy of the generated radiology reports.

5 Related Work

The most popular related task to ours is image captioning (Vinyals et al., 2015; Xu et al., 2015; Anderson et al., 2018; Wang et al., 2019), which aims to describe images with sentences. Different from them, radiology report generation requires much longer generated outputs, and possesses other features such as patterns, so that this task has its own characteristics requiring particular solutions. For example, Jing et al. (2018) proposed a co-attention

Ground-truth: *There are no old films available for comparison. The heart is moderately enlarged. There is a right ij cordis with tip in the upper svc. There is mild pulmonary vascular re-distribution but no definite infiltrates or effusion.*

Original Image | *"cardiomegaly"* | *"pulmonary"* | *"pleural"* | *"right"*

BASE: *As compared to the previous radiograph there is no relevant change. Moderate **cardiomegaly** with mild fluid overload but no overt **pulmonary** edema. No **pleural** effusions. No pneumonia. Unchanged **right** internal jugular vein catheter.*

Original Image | *"right"* | *"heart"* | *"pleural"* | *"lungs"*

BASE+RM+MCLN: *A **right** internal jugular central venous catheter terminates in the mid svc. The **heart** is moderately enlarged. The mediastinal and hilar contours are within normal limits. There is no pneumothorax or large **pleural** effusion. The **lungs** appear clear.*
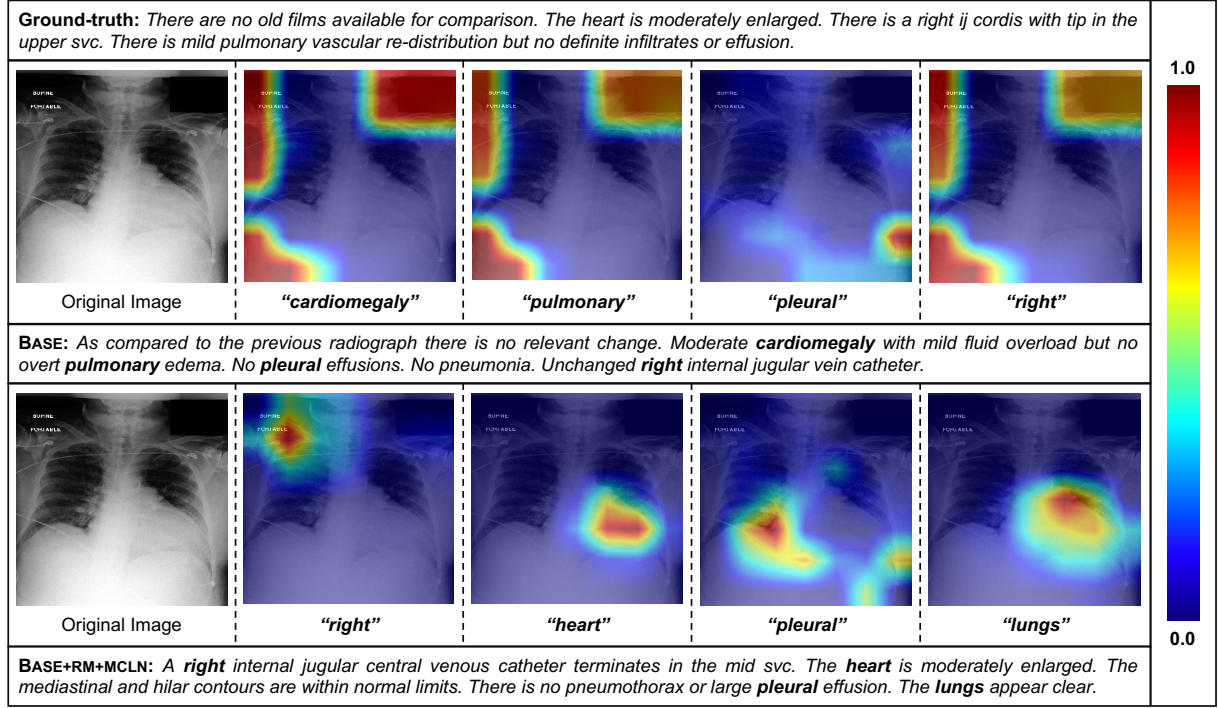
Figure 6: Visualizations of image-text attention mappings between a specific chest X-ray and generated reports from BASE and BASE+RM+MCLN, respectively. Colors from blue to red represent the weights from low to high.

mechanism and leveraged a hierarchical LSTM to generate reports. Li et al. (2018, 2019) proposed to use a manually extracted template database to help generation with bunches of special techniques to utilize templates. Liu et al. (2019) proposed an approach with reinforcement learning to maintain the clinical accuracy of generated reports. Compared to these studies, our model offers an alternative solution to this task with an effective and efficient enhancement of Transformer via memory.

Extra knowledge (e.g., pre-trained embeddings (Song et al., 2017; Song and Shi, 2018; Zhang et al., 2019) and pretrained models (Devlin et al., 2019; Diao et al., 2019)) can provide useful information and thus enhance model performance for many NLP tasks (Tian et al., 2020a,b,c). Specifically, memory and memory-augmented neural networks (Zeng et al., 2018; Santoro et al., 2018; Diao et al., 2020; Tian et al., 2020d) are another line of related research, which can be traced back to Weston et al. (2015), which proposed memory networks to leverage extra information for question answering; then Sukhbaatar et al. (2015) improved it with an end-to-end design to ensure the model being trained with less supervision. Particularly for Transformer, there are also memory-based methods proposed. For example, Lample et al. (2019) proposed to solve the under-fitting problem of Transformer by introducing a product-key layer that is

similar to a memory module. Banino et al. (2020) proposed MEMO, an adaptive memory to reason over long-distance texts. Compared to these studies, the approach proposed in this paper focuses on leveraging memory for decoding rather than encoding, and presents a relational memory to learn from previous generation processes as well as patterns for long text generation. To the best of our knowledge, this is the first study incorporating memory for decoding with Transformer and applied for a particular task, which may provide a reference for studies in the line of this research.

6 Conclusion

In this paper, we propose to generate radiology reports with memory-driven Transformer, where a relational memory is used to record the information from previous generation processes and a novel layer normalization mechanism is designed to incorporate the memory into Transformer. Experimental results on two benchmark datasets illustrate the effectiveness of the memory by either concatenating it with the output or integrating it with different layers of the decoder by MCLN, which obtains the state-of-the-art performance. Further analyses investigate how memory size affects model performance and show that our model is able to generate long reports with necessary medical terms and meaningful image-text attention mappings.

References

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.

Andrea Banino, Adrià Puigdomènech Badia, Raphael Köster, Martin J Chadwick, Vinicius Zambaldi, Demis Hassabis, Caswell Barry, Matthew Botvinick, Dharshan Kumaran, and Charles Blundell. 2020. MEMO: A Deep Network for Flexible Combination of Episodic Memories. *arXiv preprint arXiv:2001.10913*.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2019. M^2: Meshed-Memory Transformer for Image Captioning. *arXiv preprint arXiv:1912.08226*.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Shizhe Diao, Jiaxin Bai, Yan Song, Tong Zhang, and Yonggang Wang. 2019. ZEN: Pre-training Chinese Text Encoder Enhanced by N-gram Representations. *arXiv preprint arXiv:1911.00720*.

Shizhe Diao, Yan Song, and Tong Zhang. 2020. Keyphrase Generation with Cross-Document Attention. *arXiv preprint arXiv:2004.09800*.

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.

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.

Guillaume Lample, Alexandre Sablayrolles, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2019. Large Memory Layers with Product Keys. In *Advances in Neural Information Processing Systems*, pages 8546–8557.

Christy Y Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. 2019. Knowledge-Driven Encode, Retrieve, Paraphrase for Medical Image Report Generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6666–6673.

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.

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.

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.

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

Adam Santoro, Ryan Faulkner, David Raposo, Jack Rae, Mike Chrzanowski, Theophane Weber, Daan Wierstra, Oriol Vinyals, Razvan Pascanu, and Timothy Lillicrap. 2018. Relational recurrent neural networks. In *Advances in neural information processing systems*, pages 7299–7310.

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

Yan Song, Chia-Jung Lee, and Fei Xia. 2017. Learning Word Representations with Regularization from Prior Knowledge. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 143–152.

Yan Song and Shuming Shi. 2018. Complementary Learning of Word Embeddings. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 4368–4374.

Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. 2015. End-To-End Memory Networks. In *Advances in neural information processing systems*, pages 2440–2448.

Yuanhe Tian, Yan Song, Xiang Ao, Fei Xia, Xiaojun Quan, Tong Zhang, and Yonggang Wang. 2020a. Joint Chinese Word Segmentation and Part-of-speech Tagging via Two-way Attentions of Auto-analyzed Knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8286–8296.

Yuanhe Tian, Yan Song, and Fei Xia. 2020b. Supertagging Combinatory Categorial Grammar with Attentive Graph Convolutional Networks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Yuanhe Tian, Yan Song, Fei Xia, and Tong Zhang. 2020c. Improving Constituency Parsing with Span Attention. In *Findings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Yuanhe Tian, Yan Song, Fei Xia, Tong Zhang, and Yonggang Wang. 2020d. Improving Chinese Word Segmentation with Wordhood Memory Networks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8274–8285.

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.

Jason Weston, Sumit Chopra, and Antoine Bordes. 2015. Memory Networks. *CoRR*, abs/1410.3916.

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.

Jichuan Zeng, Jing Li, Yan Song, Cuiyun Gao, Michael R Lyu, and Irwin King. 2018. Topic Memory Networks for Short Text Classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3120–3131.

Hongming Zhang, Jiaxin Bai, Yan Song, Kun Xu, Changlong Yu, Yangqiu Song, Wilfred Ng, and Dong Yu. 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)*, pages 5250–5259.