Cognitive Computation (2021) 13:807–820
https://doi.org/10.1007/s12559-019-09656-w

Image Captioning with Memorized Knowledge

Hui Chen[1] · Guiguang Ding[1] · Zijia Lin[2] · Yuchen Guo[1] · Caifeng Shan[3] · Jungong Han[4]

Received: 12 November 2018 / Accepted: 29 May 2019 / Published online: 10 June 2019

Abstract

Image captioning, which aims to automatically generate text description of given images, has received much attention from researchers. Most existing approaches adopt a recurrent neural network (RNN) as a decoder to generate captions conditioned on the input image information. However, traditional RNNs deal with the sequence in a recurrent way, squeezing the information of all previous words into hidden cells and updating the context information by fusing the hidden states with the current word information. This may miss the rich knowledge too far in the past. In this paper, we propose a memory-enhanced captioning model for image captioning. We firstly introduce an external memory to store the past knowledge, i.e., all the information of generated words. When predicting the next word, the decoder can retrieve knowledge information about the past by means of a selective reading mechanism. Furthermore, to better explore the knowledge stored in the memory, we introduce several variants that consider different types of past knowledge. To verify the effectiveness of the proposed model, we conduct extensive experiments and comparisons on the well-known image captioning dataset MS COCO. Compared with the state-of-the-art captioning models, the proposed memory-enhanced captioning model shows a significant improvement in terms of the performance (improving 3.5% in terms of CIDEr). The proposed memory-enhanced captioning model, as demonstrated in the experiments, is more effective and superior to the state-of-the-art methods.

Keywords Image captioning · Attention · Memory · Encoder-decoder

Introduction

Vision and language are two advanced abilities of human beings, which are related to the cognition model in the human brain. However, it is hard to mimic such abilities by machines, i.e., enabling machines to "see" the scene and "describe" it with the natural language. Benefiting from the development of cognition science, more advanced models have been created, inspired by the human cognition system, e.g., convolutional neural network (CNN) [36]. And many cognition-related tasks have been better solved and obtained significant breakthroughs, e.g., image classification [63], machine translation [44], and image captioning [11].

To enable machines to understand the scene shown in the image and describe it in human languages, many efforts have been put into the image captioning field from both academia and industry. Image captioning is a challenging task for both understanding the visual contents and describing them in natural languages. In spite of the difficulties, it has a promising value in a wide range of applications, such as video tracking [32–35], childhood education [52], cross-view retrieval [13, 41], visual impairment rehabilitation [14], and sentiment analysis [38].

There are many pioneering works attempting to tackle this challenge [7, 54]. The majority of the previous works adopt the encoder-decoder framework, which has been proved to be effective in dealing with sequence generation tasks. Benefiting from the advancement in image classification [63], a CNN [28, 36] is usually used as an encoder to extract visual features for a given image, and then a recurrent neural network (RNN), especially long-short-term memory (LSTM), is used to generate sentences conditioned on these visual features. Vinyals et al. [54]

This work was supported by the National Key R&D Program of China (Nos. 2018YFC0806900) and the National Natural Science Foundation of China (Nos. 61571269).

✉ Guiguang Ding
dinggg@tsinghua.edu.cn

Hui Chen
jichenhui2012@gmail.com

[1] School of Software, Tsinghua University, Beijing, China

[2] Microsoft Research, Beijing, China

[3] Philips Research, Eindhoven, Netherlands

[4] WMG Data Science, University of Warwick, Coventry, UK

extracted a static vector to represent the image information and then injected it into the decoder. They trained the whole system in an end-to-end way and obtained a state-of-the-art performance. And to mimic the visual system of humans, visual attention mechanism [1, 6, 25, 45, 59] was introduced into this field. Instead of squeezing all the visual content into a vector, a feature map, which consists of a series of visual features, was extracted by CNNs. Then, an attention module was adopted to adaptively attend to different salient visual features under the supervision of the current context information.

Although impressive captioning performance has been achieved, the quality of captions generated by existing captioning models still remains unsatisfying. Generally, to generate a word, existing captioning models simply store the past information of all the generated words with the hidden states, which are two vectors in LSTM. Then, the generation process is performed in a recurrent way. However, due to the difficulties of the RNNs in capturing the knowledge in the long run, the previous knowledge will vanish gradually along the generation process. Therefore, the hidden state may not be adequate to represent past knowledge, resulting in a loss of the past knowledge, especially the knowledge of words far from the current word. Although LSTMs [22] are more powerful in memorizing the past knowledge than RNNs, they still struggle to remember words too far in the past [56]. In fact, the past information of words that are not close to the current word can be important for the word prediction and can boost the caption generation, because there may exist strong semantic dependencies between previous words and the current one. For example, in the sentence "a vase filled with red and green flowers", "vase" is more related to "flowers" than "red" or "green," although it is far from "flowers."

In this paper, we propose a memory-enhanced captioning model, which attempts to enhance the capability of memorizing the past knowledge for the captioning model via the memory mechanism. We name the proposed captioning model as memory with selective reading mechanism (MemSRM). Specifically, we incorporate an external memory into the encoder-decoder framework equipped with a visual attention module. The usage of the memory can help preserve the past knowledge of all previously generated words, and can selectively provide rich knowledge about the past generated words, especially those words not close to the current word. And in the proposed selective reading mechanism, when the decoder is about to make a prediction, the memory can perform a reading operation to extract a related feature representing the past knowledge by means of an attention module. With the aid of such an attention module, the memory is able to selectively retrieve the most relevant information to the input query information and thus can provide rich knowledge about the past. Such knowledge-related information will be aggregated with the current word information via a gated fusion unit further. And finally, the result of aggregation will be leveraged to predict the subsequent word.

Besides, to further explore the impact of memory, we enumerate different information of past knowledge, including visual knowledge, semantic knowledge, and a fusion of both kinds of knowledge. Similar to previous works [45, 59], the attention module equipped in the decoder can associate different words with different saliency visual features. Such attention features can be regarded as the visual knowledge corresponding the generated words, and thus can be stored in the memory and provide rich past visual information for the decoder. The semantic knowledge is related to the representation of the generated words directly. This kind of knowledge contains the semantic dependencies among words in a sentence, and can be used to provide rich knowledge about the semantic dependencies among previously generated words. And the fusion knowledge combines visual knowledge and semantic knowledge, further improving the knowledge capacity.

To verify the effectiveness of the proposed approach, we conduct extensive experiments and analyses on a well-known image captioning benchmark dataset, i.e., MS COCO. We also make comparisons to the state-of-the-art approaches. The experimental results well demonstrate that the proposed approach is more effective and superior to the state-of-the-art approaches.

Overall, the main contributions of our work are three-fold.

- We propose a memory-enhanced captioning model, named MemSRM, attempting to strengthen the captioning model in memorizing the past knowledge, especially about generated words far from the current word. By incorporating a selective reading mechanism (SRM), the introduced memory can provide the decoder with informative knowledge about the previously generated words, and thus boost the captioning performance.
- We further explore different kinds of past knowledge to be stored in the memory, including visual knowledge semantic knowledge, and the fusion of both. The different kinds of knowledge can provide rich information of past knowledge, and are proved to be effective to improve the performance in our experiments.
- We validate the effectiveness of the proposed memory-enhanced captioning model on the MS COCO image captioning dataset by conducting extensive experiments and analyses. Comparison experiments demonstrate that the proposed approach is more effective and superior to the state-of-the-art approaches.

The preliminary conference version of our work was presented in [3]. Compared with the conference paper, we enhance our work from the following three aspects. First, we provide a more comprehensive review of related work. Second, we extend the ways of the interaction between the memory and the decoder and adopt a gated fusion unit to compose the context information for predicting the next word using the knowledge information provided by the memory and the information about the current word. Third, we conduct more comparative experiments and enrich the analysis and discussion of results.

The rest of this paper is structured as follows. The "Related Work" section provides a comprehensive review of related works on image captioning. The "Encoder-Decoder Framework for Image Captioning" section gives an overview of the basic captioning model, i.e., the encoder-decoder framework with an attention mechanism. Details about the proposed memory-enhanced captioning model, including writing/reading operations and three different kinds of knowledge stored in the memory, are described in the "Image Captioning with Memory" section. Experimental results and analyses are presented in the "Experiments" section, followed by conclusions and potential future work in the "Conclusion" section.

Related Work

Many methods have been proposed to improve the performance of image captioning models. Generally, the existing image captioning algorithms can be divided into three categories, i.e., template-based approaches, transfer-based approaches, and neural network-based approaches.

Template-Based Approaches This category of approaches generally use templates or design a language model which fills in slots of a template. Farhadi et al. [17] modeled the co-occurrence relations among words to elaborately design the templates. And in [29], a conditional random field was utilized to capture the dependencies among different templates. More complicated models have also been applied in the image captioning task and shown to be effective in generating relatively flexible sentences. Mitchell et al. [48] exploited syntactic trees by leveraging syntactically informed word co-occurrence statistics. Elliott et al. [15] introduced a kind of visual dependency representation to capture the relationships between objects in an image. Despite being simple and intuitive, template-based methods are heavily hand-designed and not expressive enough to generate meaningful sentences.

Transfer-Based Approaches This category of approaches adopts cross-modal retrieval technique based on an assumption that a similar image could be described with similar or even identical captions. Generally, descriptions of a retrieved image are regarded as the reference caption directly for a given to-be-captioned query image. Gong et al. [19] and Micah et al. [23] aligned the visual information and the semantic information in a joint latent space, and then selected a description from the database which is close to the query image as the reference. Kuznetsova et al. [30, 31] proposed to retrieve related images as neighboring images, and then extracted segments from their captions and arranged a description as the final caption. Devlin et al. [10] simply searched the top-k most similar sentence and selected the best sentences by calculating the consensus score [9] of the corresponding captions.

Neural Network-Based Approaches Inspired by recent advances in machine translation [8, 44], most recent works focus on the neural network-based approaches. This category of approaches generally adopts the encoder-decoder [7, 54] framework, where a recurrent neural network (RNN) is employed to generate a caption based on the information of an image. The attention mechanism is often equipped in the state-ot-the-art captioning performance due to its capability of attending to different salient aspects of information related to input query information. Therefore, we here focus on the related works about captioning models with the attention mechanism.

- *Visual attention.* The visual attention model makes the image feature adaptive to the sentence context at hand [6]. Xu et al. [59] firstly introduced the visual attention into image captioning. Chen et al. [6] proposed spatial and channel-wise attention to attend to both salient region features and salient channels of features. Lu et al. [45] introduced a visual sentinel allowing attending to regions adaptively. Anderson et al. [1] combined both bottom-up attention and top-down attention to generate more informative image features, resulting in a better generation.
- *Semantic attention.* You et al. [62] firstly proposed to selectively attend to semantic concepts instead of visual regions. Jia et al. [24] used the global semantic correlation between images and captions to guide the decoder to generate better captions. Chen et al. [4] extracted attribute-level features for images and incorporated them into the decoder with the attention.

Recently, reinforcement learning has been introduced to improve performance and shown a promising success. Ranzato et al. [50] proposed Mixed Incremental Cross-Entropy Reinforce (MIXER) to directly optimize the evaluation metrics used at inference time. Self-critical sequence training (SCST) [51] rewarded the whole sentence and directly maximized the expected reward during training. Chen et al. [5] used the temporal-difference learning method to model the temporal information between consecutive actions. Liu et al. [42] used a monto rollout strategy to estimate the expected reward for each word in a sentence and directed the model to generate sentences with high rewards.

Our work is also related to [16, 26, 55]. Fakoor et al. [16] adopted the memory network to solve the locality problem for video captioning, where the memory network aims to acquire the global information of the sequential frames in videos. Kaiser et al. [26] employed the memory to capture the information of rare words in the training dataset. Wang et al. [55] proposed a memory-enhanced decoder for neural machine translation with an external memory, which shared the same idea as ours. Different from [55], we propose to utilize the memory to remember the past knowledge for image captioning task, which includes visual knowledge, apart from the semantic knowledge.

Encoder-Decoder Framework for Image Captioning

Most image captioning models follow the encoder-decoder framework, which generally consists of two components: a visual encoder and a sentence decoder. Given an RGB image I, an encoder firstly encodes the image into the visual representation, i.e., the feature map. Then the sentence decoder will generate a sentence word by word on the basis of the visual representation. In this section, we will introduce these important components to give an overview of the basic captioning model.

Visual Encoder The informativeness of the visual representations plays an important role in image captioning. Benefiting from the significant advances in image classification and object detection fields, convolutional neural networks (CNNs) are widely used in image captioning as the visual encoder. Generally, such CNNs are pre-trained on larger-scale labeled datasets, such as ImageNet, and can extract compact and representative features for images. Specifically, given a RGB image I, an encoder CNN will encode it into a feature map, $V = [v_1, v_2, ..., v_n]$, $v_i \in \mathbb{R}^D$.

Sentence Decoder The goal of the sentence decoder is to predict a series of consecutive words and compose them into a sentence $Y = \{y_0, y_1, ..., y_T\}$. It is crucial for the decoder to be aware of the transition between two successive words and the context information from previously generated words. Recurrent neural networks (RNNs) have been proved to be effective in modeling the context in a sentence, and thus usually be adopted as the decoder. However, the issue of exploding or vanishing gradient impedes the performance of generic RNNs. In image captioning, a variant of RNNs, i.e., long-short-term memory (LSTM), is usually employed to substitute the generic RNN due to its strengths in capturing long-term dependencies among words in a sentence. Specifically, at each time step t, given the current information x_t, the previous hidden state h_{t-1}, and context vector c_{t-1}, LSTM updates its parameters as follows:

$$\begin{aligned}
i_t &= \sigma(W_{ix}x_t + W_{ih}h_{t-1} + b_i)\\
f_t &= \sigma(W_{fx}x_t + W_{fh}h_{t-1} + b_f)\\
o_t &= \sigma(W_{ox}x_t + W_{oh}h_{t-1} + b_o)\\
c_t &= i_t \odot \phi(W_{zx}x_t + W_{zh}h_{t-1} + b_c) + f_t \odot c_{t-1}\\
h_t &= o_t \odot tanh(c_t)\\
q_t &= W_{qh}h_t
\end{aligned} \tag{1}$$

where i_t, f_t, and o_t denote the input gate, the forget gate, and the output gate, respectively. And $\odot$ means the element-wise multiplication. For the ease of the explanation, we use $h_t = \text{LSTM}(x_t, h_{t-1})$ to indicate the above process. In Fig. 1, we show a LSTM cell on the right.

Top-Down Model In this paper, we adopt the top-down [1] as our basic captioning model. Top-down adopts the attention mechanism to adaptively attend to different parts during decoding. As shown in Fig. 1, the word y_t will be

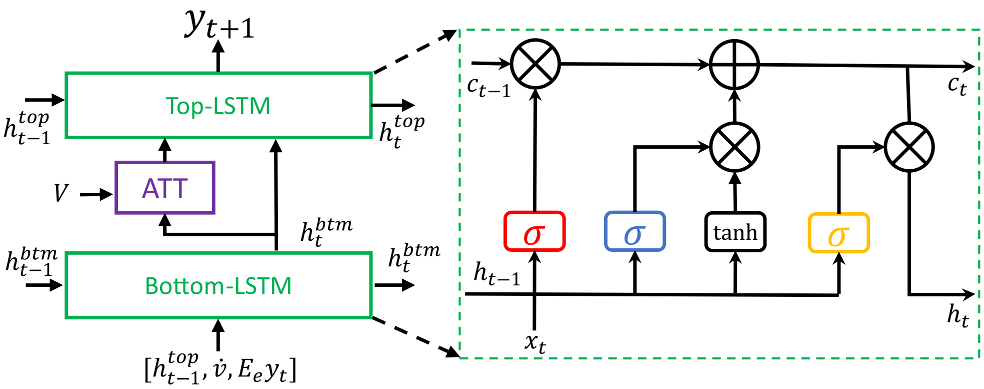


Fig. 1 Left: the top-down architecture. Right: a LSTM cell. In the green box, the input gate, the forget gate, and the output gate are colored by blue, red, and orange, respectively. Best viewed in color

fed into the bottom LSTM with the previous hidden states, i.e., h_{t-1}^{top} and h_{t-1}^{btm}, and the global image feature, i.e., $\dot{v}$. The hidden state of the bottom LSTM, i.e., h_t^{btm}, is updated as follows:

$$h_t^{btm} = \text{LSTM}([h_{t-1}^{top}, \dot{v}, E_e y_t], h_{t-1}^{btm}) \tag{2}$$

where E_e is an embedding dictionary to be learned, and y_t represents a one-hot vector, where only the position corresponding to the current word has the value, i.e., 1, while others positions are all zeros.

The bottom LSTM can be regarded as a transformation function for the embedding representations of words. Then, the attention module performs a soft attention function over the image's feature map V with the current word's representation h_t^{btm}. We first define a soft function denoted by $att()$ which takes a matrix $V = [v_1, v_2, ..., v_n]$, $v_i \in \mathbb{R}^D$ and a query vector h as inputs, and outputs a context vector c:

$$\begin{aligned} &c = \text{att}(V, h) = \sum_{i=1}^{k} \alpha_i v_i \\ &s.t. \quad \alpha = \text{softmax}(W_a \tanh(W_{av} V + (W_{ah} h)\mathbf{1}^T)) \end{aligned} \tag{3}$$

where W_a, W_{av}, and W_{ah} are parameters to be learned.

Then, the visual feature $\bar{v}_t$ produced by the attention module at the time step t can be written as follows:

$$\bar{v}_t = \text{att}(V, h_t^{btm}) \tag{4}$$

where V is the feature map of the image.

Finally, the top LSTM functions as a predictor which will generate a distribution over the whole vocabulary:

$$\begin{aligned} &h_t^{top} = \text{LSTM}([\bar{v}_t, h_t^{btm}], h_{t-1}^{top}) \\ &p(y_t | y_{0:t-1}, I) = \text{softmax}(W_p h_t^{top} + b_p) \end{aligned} \tag{5}$$

where W_p and b_p are parameters to be learned.

Image Captioning with Memory

In this section, we describe the proposed memory-enhanced captioning model. Specifically, in order to preserve the past knowledge, we use an external memory to store the information of all words generated before. And when decoding, an informative knowledge about the generated words will be provided by the memory to improve the prediction of the next word. An overview of the proposed captioning model enhanced by a memory is illustrated in Fig. 2. In the following sections, we first introduce the memory in detail, including how to access the memory (described in the "Memory Access" section) and how to construct the memory (described in the "Memory Construction" section). Then, a gated fusion unit is proposed in the "Gated Fusion Unit" section to incorporate the information of the past knowledge with the current word information provided by the decoder. Finally, the training and optimizing approaches are described in the "Training and Optimization" section.

Memory Access

The past knowledge can provide rich context information for the prediction. However, the current captioning model may miss such rich knowledge, especially knowledge about words far from the current word. We propose to enhance the capacity of preserving the knowledge for the captioning model by using an external memory. Specifically, we adopt an external memory, denoted by M, to store the knowledge about the image and the words that the model has learned. Such memory is described as an array of objects denoted by m_t. And when interacting with other components in the captioning model, i.e., the encoder, the decoder, and the attention module, the memory performs two access mechanisms depending on the direction of information flows. One is the writing operation, which updates the

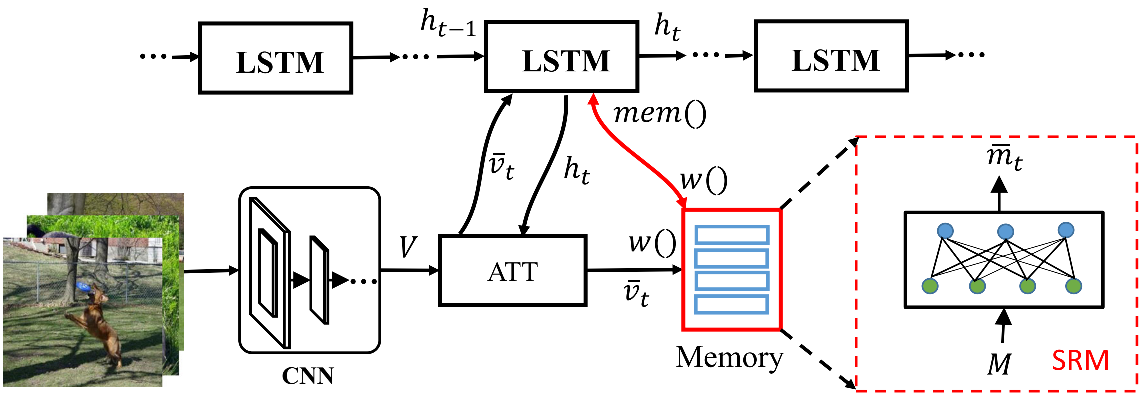


Fig. 2 The framework of the proposed memory-enhanced captioning model. $mem()$ is the reading operation, which adopts a selective reading mechanism (SRM), and $w()$ is the writing operation

content of the memory to maintain the information of the knowledge. The other is the reading operation, which reads the most informative feature related to the query input from the memory.

Writing The writing function brings in the knowledge which is desired to be memorized in the generation process. In the initial state, the memory M is empty. Along with the decoding process, it is filled with the past knowledge by extending its size. Specifically, we treat one position of memory as an empty slot, and directly insert the knowledge information into the slot one by one along the process:

$$m_t = f_t \tag{6}$$

where f_t is the knowledge at time step t, and m_t is the slot in the memory indexed by i.

We conduct different strategies to construct the knowledge f_t, which is described in the "Memory Construction" section in detail. Note that the decoder will read from the memory at every time step, which means that the size of memory may cause a burden on the efficiency of the decoding and training. However, in our experiment, we do not restrict the capacity of the memory considering that sentences in captioning dataset, e.g., MS COCO, are tailored with a relatively decent length.

Reading The reading operation, namely mem() for simplicity, selects the related knowledge for the prediction at the current time step. Since not all knowledge is helpful for the prediction, we need to filter the irrelevant information and return the most related information to the current information of knowledge at time step t, i.e., f_t. Here, we propose a selective reading mechanism (SRM) to achieve that goal.

It is intuitive that the relationship between two words depends on the similarity between their corresponding features in the latent space. The closer a word is to the other word, the more likely a strong dependency exists between them. Therefore, we propose the selective reading mechanism to attend to the most related knowledge adaptively when querying the memory. Specifically, by leveraging the attention mechanism described in Eq. 3, we first produce a relationship distribution of the past knowledge conditioned on the current query information by a softmax.

$$\beta = \mathrm{softmax}(W_b \tanh(W_{bm} M + (W_{bf} f_t)\mathbf{1}^T)) \tag{7}$$

where $M = \{m_0, m_1, ..., m_k\}$ and W_b, W_{bf} and W_{bh} are parameters to be learned.

Then, we obtain the knowledge feature $\bar{m}_t$ via weighting and summing all knowledge in the memory:

$$\bar{m} = \mathrm{mem}(M, f_t) = \sum_{i=1}^{k} \beta_i m_i \tag{8}$$

where m_i is the ith element in the memory. The selective reading mechanism can choose the most related past knowledge by assigning a larger weight to the corresponding element, which can mine relationships among words in the same sentence according to their relative similarities in the embedding space.

The previous work [16] used a recurrent layer to model the reading mechanism. However, as described in the

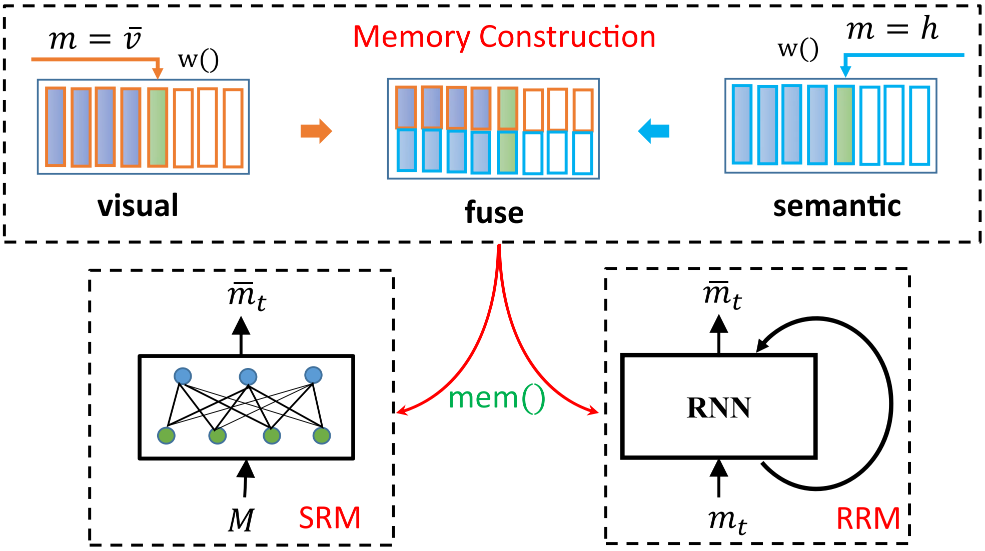


Fig. 3 Top: Three kinds of knowledge stored in the memory. Bottom: the proposed selective reading mechanism (SRM) and the compared recurrent reading mechanism (RRM)

"Introduction" section, the recurrent layer can lead to a loss of previous knowledge. On the contrary, the proposed selective reading mechanism can attend to all the previous knowledge selectively without regard to the distance. Therefore, to see the superiority of the proposed SRM, we regard the recurrent modeling strategy as one of the baselines. And for the ease of the explanation, we name the recurrent modeling strategy as recurrent reading mechanism (RRM). Practically, we maintain a LSTM module and feed the past knowledge into it step by step:

$$\begin{aligned} h_{t_m}^{mem} &= \mathrm{LSTM}(m_{t_m}, h_{t_m-1}^{mem}) \\ \bar{m} &= \mathrm{mem}(M, f_t) = \mathrm{LSTM}(f_t, h_{t-1}^{mem}) \end{aligned} \tag{9}$$

where $h_{t_m-1}^{mem}$, $h_{t_m}^{mem}$, and h_{t-1}^{mem} are the hidden state vectors of the LSTM module. t is the index of the time step of the decoder. Figure 3 shows the difference between the proposed SRM and the compared RRM.

Memory Construction

Since the memory aims to provide informative knowledge about the past generated words for the current decoder, it is important to allocate appropriate knowledge information in the memory. In this section, we will discuss how to construct the proposed memory by means of different kinds of past knowledge, which are the visual knowledge, the semantic knowledge, and the fusion knowledge, as illustrated in Fig. 3.

- **Visual knowledge**

  In image captioning models, the decoder usually adopts the visual attention model to attend to the visual information, i.e., the given image, at every time step, so as to generate captions that are consistent with the image's content. This visual attention mechanism can enable the decoder to adaptively perceive different visual aspects of an image related to the next word. Generally, the attention mechanisms model the salience of regions in a image with a posterior probability, i.e., $p(V|h)$ where V is the input region features of the image, and h is the hidden state of the decoder which can be also regarded to be related to the current word. Then, a weighted summing function is applied to aggregate the input region features, i.e., $\bar{v}_t = \sum_{i=0}^{n} p(v_i|h) * v_i, v_i \in V$. We can consider the attention mechanism as a fuzzy function which highlights the region we attend to and makes others blurry. And thus, the feature vector, $\bar{v}_t$, produced by the attention model at different time steps t can describe different visual aspects of the image implicitly, e.g., objects, attributes, and other semantic features. Such information can be considered as the visual knowledge that has learned by the decoder. Therefore, we regard the attention features, $\bar{v}_t$, as the visual knowledge and we store them in the memory. The writing function and reading function can be denoted as follows:

  $$\begin{aligned} \bar{m}_t^{vis} &= \mathrm{mem}(M^{vis}, h_t^{btm}) \\ m_t^{vis} &= \bar{v}_t \end{aligned} \tag{10}$$

  where $M^{vis} = \{m_0^{vis}, m_1^{vis}, ..., m_{t-1}^{vis}\}$ and h_t^{btm} is the hidden state of the bottom LSTM in the top-down captioning model.

- **Semantic knowledge**

  The context in semantic dependencies among words in a sentence plays an important role in language generation. Modeling such semantic context among words has brought great performance improvement for image captioning task, benefited from the usage of recurrent neural networks, especially LSTMs. However, due to the difficulties of the RNN in capturing the long term dependencies, it will forget the semantic information too far in the past gradually. Even though LSTMs, benefited from the gating mechanism, are more powerful in memorizing the past information than vanilla RNNs, they still have trouble remembering words too far in the past [56]. Note that the semantic information at different time steps cannot well represent words in a distance, but they can still adequately capture the semantic dependencies near the current word. Therefore, we store such local semantic information in an external memory as the semantic knowledge that the decoder has learned. In this way, we can keep all the semantic dependencies information that has learned before and conveniently access them along the generation process, so that the decoder can well perceive the past semantic knowledge and improve its generation. The writing function and reading function can be denoted as follows:

  $$\begin{aligned} \bar{m}_t^{sem} &= \mathrm{mem}(M^{sem}, h_t^{btm}) \\ m_t^{sem} &= h_t^{btm} \end{aligned} \tag{11}$$

  where $M^{sem} = \{m_0^{sem}, m_1^{sem}, ..., m_{t-1}^{sem}\}$.

- **Fusion knowledge**

  We also adopt a fusion strategy to combine both knowledges described above. Specifically, when updating the memory, we firstly aggregate the visual knowledge and the semantic knowledge with a non-linear function, and the result is written into the memory directly:

  $$\begin{aligned} u_t &= \sigma(W_u[\bar{v}_t, h_t^{btm}] + b_u) \\ m_t^{fuse} &= u_t \end{aligned} \tag{12}$$

where $[.,.]$ means concatenation. W_u and b_u are learnable parameters. σ is a non-linear function, which is a $relu$ function with dropout here.

The reading operation can be denoted as follows:

$$\bar{m}_t^{fuse} = \text{mem}(M^{fuse}, h_t^{btm}) \tag{13}$$

where $M^{fuse} = \{m_0^{fuse}, m_1^{fuse}, ..., m_{t-1}^{fuse}\}$

Gated Fusion Unit

After obtaining the knowledge feature $\bar{m}_t$, i.e., the output of the memory, we further adopt a gated fusion unit to aggregate the knowledge feature with the current feature for prediction. Specifically, firstly, we concatenate the knowledge feature, $\bar{m}_t$, and the current feature, h_t^{top}, followed by a non-linear transformation layer:

$$o_t = \phi(W_c[\bar{m}_t : h_t^{top}] + b_c) \tag{14}$$

where [:] means the concatenating function and $\bar{m}_t$ is the knowledge representation, which can be either $\bar{m}_t^{vis}$, $\bar{m}_t^{sem}$, or their combination $\bar{m}_t^{fuse}$ introduced above. ϕ is a non-linear function, which is a $tanh$ function here. W_c and b_c are learnable parameters. Then, we use a gate to decide how much the prediction depends on the current feature, h_t^{top}, and the fused knowledge feature, o_t, respectively:

$$\begin{aligned} g_t &= \varphi(W_g[\bar{m}_t : h_t^{top}] + b_g) \\ z_t &= g_t \odot o_t + (1 - g_t) \odot h_t^{top} \end{aligned} \tag{15}$$

where φ is a non-linear function, which is a $sigmoid$ function here. $\odot$ means the element-wise product. W_g and b_g are learnable parameters. g_t is a gating feature with the same dimension as h_t^{top}. Finally, z_t is used to generate a distribution over the whole vocabulary:

$$p(y_t | y_{0:t-1}, I) = \text{softmax}(W_z z_t + b_z) \tag{16}$$

where W_z and b_z are parameters to be learned.

Training and Optimization

The captioning model is encouraged to directly maximize the cross-entropy objective during training:

$$\theta^* = \arg\min_{\theta} \sum_{t=1}^{T} \log p(y_t | I, \theta, y_0, y_1, ..., y_{t-1}) \tag{17}$$

where T is the maximum length of sentences and θ denotes the parameters of the proposed network.

As revealed in [50], the cross-entropy objective requires the ground-truth sentence to be input into the captioning model during training. However, in the inference procedure, the sentence is generated by the model itself, dependent on the information of previously generated words, which causes an inconsistency between the training procedure and the inference procedure. To address this issue, reinforcement learning based objective is introduced into image captioning, which directly optimizes the non-differentiate NLP metrics and back-propagates the gradient through REINFORCE algorithms. Similar to the previous work [1], we also use the reinforcement learning strategy to optimize the proposed memory-enhanced captioning model during training. Specifically, after being pre-trained with the cross-entropy loss, we minimize the negative expected reward:

$$L_r(\theta) = -E_{Y \backsim \pi_\theta} r(Y) \tag{18}$$

where π_θ denotes the distribution of sentences and $Y = \{y_0, y_1, y_2, ..., y_T\}$ is a sentence sampled according to the distribution π_θ via Monte-Carlo sampling method. r is a metric evaluation (we use CIDEr as the optimized metric here).

To reduce the gradient variance and stabilize the training process, following [51], we introduce a baseline term and the gradient of the reinforcement learning objective can be approximated as:

$$\nabla_\theta L_r(\theta) = -(r(Y) - r(Y'))\nabla_\theta \log \pi_\theta(Y) \tag{19}$$

where Y' is a sentence generated by the greedy decoding algorithm used in the inference time.

The reinforcement learning objective can capacitate the captioning model automatically to explore the sentence distribution space. While training, the model tends to increase the probability of sentences with high scores and reduce the likelihood of those with lower scores, which can greatly improve the quality of the generated sentences.

Experiments

Dataset

To verify the effectiveness of the proposed memory-enhanced captioning model (MemSRM), we conduct a series of experiments on a popular image captioning dataset, named MS COCO, which is the largest dataset for image captioning. It consists of 82,783 training images and 40,504 validation images. And each image has at least 5 ground-truth sentences. For offline evaluation, we follow previous works [45, 59], and split the 123,287 images into three parts, i.e., 5000 for validation, 5000 for test, and the remaining for training. It also provides 40,775 images as the test set for online evaluation. We will report the results of both

Table 1 Performance on MS COCO. All models are trained with ResNet image feature

Model	Knowledge	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
Top-down [1]	no	76.6	–	–	34.0	26.5	54.9	111.1
Top-down (ours)	no	80.2	62.8	47.8	35.6	27.0	56.6	113.6
MemRRM	vis	80.1	63.0	48.1	35.9	27.5	57.1	117.3
	sem	80.2	63.3	48.2	36.3	27.5	56.9	117.2
	fuse	80.0	63.0	48.1	36.0	27.3	57.1	117.7
MemSRM	vis	80.6	63.6	48.6	36.5	27.7	57.3	118.8
	sem	80.3	63.1	48.1	35.8	27.7	57.1	117.8
	fuse	80.1	63.4	48.2	36	27.5	57.1	118.6

online and offline evaluation, and make comparisons with state-of-the-art approaches under these two settings, too.

Implementation Details

Following previous works [46], we filter out words that occur less than 5 times and trim captions to a maximum of 16 words. Eventually, we set up a vocabulary of 9487 words.

Image Features We use two kinds of image features.

- ResNet features: We use the ResNet-101 [21] pre-trained on ImageNet as the encoder CNN. We adopt the feature map of the final convolutional layer as the visual features. And we apply spatially average pooling so that the feature map has a fixed size of 14 × 14 × 2048.
- Bottom-up features [1]: We use a Faster R-CNN pre-trained on Visual Genome. To extract features, we firstly detect all objects in the image, then extract top 36 features in each image with highest probabilities, which ends up with a 36-by-2048 feature map for each image.

Training Details For all models, the hidden state size of the decoder LSTM is 1300. The embedding dimension of each word is fixed as 1000. We set the embedding dimension of the image feature to 1000 using a linear layer. For a fair comparison, when pre-trained under cross-entropy loss, all models are trained in an end-to-end way using the ADAM optimizer with a learning rate of 5×10^{-4} and a learning rate decay factor of 0.8. Batch size is set to 100. We pre-trained the captioning model using the cross-entropy objective function for 40 epochs and choose the best model on the validation set as the initial model for reinforcement learning. When trained using reinforcement learning objective function, the ADAM optimizer is adopted with an initial learning rate of 1×10^{-4} and a learning rate decay factor of 0.8. The reinforcement learning process is running for up to 60 epochs.

Test Strategy For test, we use the word sampled from the prediction of the model at the last time step. We apply beam search strategy to generate captions with higher probabilities. By default, we set the beam size as 3, which is commonly used in the previous works [45].

Evaluation Metrics To compared with other methods, we use the same evaluation metrics, including BLEU (B@1,B@2,B@3,B@4) [49], METEOR (MT) [2], ROUGE-L (RG) [39], and CIDEr (CD) [53]. Meanwhile, we use the MS COCO caption evaluation tool[1] to compute these metrics.

Quantitative Analysis

Evaluations of the Memory Tables 1 and 2 show the performance comparisons among variants of the proposed memory-enhanced captioning model and the baseline models, using ResNet image features and bottom-up image feature, respectively. Note that we re-implement our baseline model, i.e., top-down [1], and our re-implementation can achieve higher performance than that reported in [1]. Therefore, here, we make comparisons with both our re-implementation, i.e., top-down (ours) and top-down [1]. To verify the superior of the proposed selective reading mechanism, we report the performance of the memory-enhanced model with the recurrent reading mechanism, which is indicated by MemRRM in Tables 1 and 2.

As illustrated in Tables 1 and 2, for both kinds of image features, our proposed MemSRM can achieve superior performance than both baseline models.

Specifically, in Table 1, compared with top-down (ours), the proposed MemSRM, which adopts the selective reading mechanism to read from the memory, can obtain a maximum improvement of 0.9% and 5.2% in BLEU-4 and

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

Table 2 Performance on MS COCO. All models are trained with bottom-up image feature

Model	Knowledge	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
Top-down [1]	no	79.8	–	–	36.3	27.7	56.9	120.1
Top-down (ours)	no	81.2	64.3	49.3	37.0	27.9	58.0	121.4
MemRRM	vis	81.4	64.7	49.9	37.7	28.4	58.2	122.8
	sem	81.5	65.0	50.1	37.9	28.4	58.2	123.2
	fuse	81.4	64.8	49.8	37.5	28.4	58.3	122.8
MemSRM	vis	81.6	65.1	50.1	38.1	28.3	58.3	122.8
	sem	81.2	64.6	49.8	37.6	28.4	58.3	123.5
	fuse	81.4	64.8	49.9	37.7	28.5	58.4	123.4

CIDEr. And in Table 2, the maximum improvement is 1.1% and 2.1%, in terms of BLEU-4 and CIDEr, respectively. Better performance can be attributed to the introduced memory, which can enhance the ability to memorize past knowledge for the captioning model. Besides, if the memory adopts the recurrent modeling strategy as [16], i.e., MemRM, it also can achieve a performance improvement compared to top-down (ours), which further demonstrate the importance of the past knowledge for the current prediction. Besides, compared with MemRRM, for both the ResNet image feature and bottom-up image feature, MemSRM can obtain better performance than MemRRM

Table 3 Performance comparison of the proposed method with the state-of-the-art approaches on MS COCO

Models	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
Google NIC [54]	66.6	45.1	30.4	20.3	–	–	–
Deep VS [27]	62.5	45.0	32.1	23.0	19.5	–	66.0
R-LSTM [7]	76.1	59.6	45.0	33.7	25.7	55.0	102.9
R-LSTM [11]	76.5	60.3	45.8	34.4	26.4	55.7	106.4
m-RNN [47]	67.0	49.0	35.0	25.0	–	–	–
ERD [60]	–	–	–	29.8	24.0	–	89.5
ATT [62]	70.9	53.7	40.2	30.4	24.3	–	–
Hard-Attention [59]	71.8	50.4	35.7	25.0	23.4	–	–
Soft attention [59]	70.7	49.2	34.4	24.3	23.9	–	–
SCA-CNN [6]	71.9	54.8	41.1	31.1	25.0	53.1	95.2
Adaptive-Attention [45]	74.2	58.0	43.9	33.2	26.6	–	108.5
SCN-LSTM [18]	72.8	56.6	43.3	33.0	25.7	–	101.2
MSM [61]	73.0	56.5	42.9	32.5	25.1	53.8	98.6
RA+SF [25]	69.7	51.9	38.1	28.2	23.5	50.9	83.8
VS-LSTM [37]	78.9	63.4	48.1	36.3	27.3	–	120.8
Show and observe [4]	74.3	57.9	44.3	33.8	–	54.9	104.4
StackCap [20]	78.4	62.5	47.9	36.1	27.4	56.9	120.4
SCST [51]	–	–	–	35.4	27.1	56.6	117.5
SR-PL [43]	80.1	63.1	48.0	35.8	27.4	57.0	117.1
MIXER [50]	–	–	–	30.9	24.9	53.8	101.9
Top-down [1]	79.8	–	–	36.3	27.7	56.9	120.1
memory-att [3]	75.7	59.5	45.7	35.0	-	55.7	109.2
Ours	81.6	65.1	50.1	38.1	28.3	58.3	122.8
Ours[a]	81.9	65.5	50.7	38.4	28.7	58.7	125.5

[a] The results of ensemble models

Table 4 Evaluation performance of the proposed method on the online MS COCO testing server

	BLEU-1		BLEU-2		BLEU-3		BLEU-4		METEOR		ROUGE-L		CIDEr	
	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
Google NIC[†] [54]	71.3	89.5	54.2	80.2	40.7	69.4	30.9	58.7	25.4	34.6	53.0	68.2	94.3	94.6
m-RNN [47]	71.6	89.0	54.5	79.8	40.4	68.7	29.9	57.5	24.2	32.5	52.1	66.6	91.7	93.5
ERD [60]	72.0	90.0	55.0	81.2	41.4	70.5	31.3	59.7	25.6	34.7	53.3	68.6	96.5	96.9
MSM[†] [61]	73.9	91.9	57.5	84.2	43.6	74.0	33.0	63.2	25.6	35.0	54.2	70.0	98.4	100.3
R-LSTM [7]	75.1	91.3	58.3	83.3	43.6	72.7	32.3	61.6	25.1	33.6	54.1	68.8	96.9	98.8
ATT-FCN[†] [62]	73.1	90.0	56.5	81.5	42.4	70.9	31.6	59.9	25.0	33.5	53.5	68.2	94.3	95.8
Hard-Attention [62]	70.5	88.1	52.8	77.9	38.3	65.8	27.7	53.7	24.1	32.2	51.6	65.4	86.5	89.3
Adaptive-Attention[†] [45]	74.6	91.8	58.2	84.2	44.3	74.0	33.5	63.3	26.4	35.9	55.0	70.6	103.7	105.1
SCA-CNN [6]	71.2	89.4	54.2	80.2	40.4	69.1	30.2	57.9	24.4	33.1	52.4	67.4	91.2	92.1
Top-down [1]	80.2	95.2	64.1	88.8	49.1	79.4	36.9	68.5	27.6	36.7	57.1	72.4	117.9	120.5
memory-att [3]	75.5	92.7	59.2	85.2	45.5	75.4	34.8	64.8	27.2	36.7	55.8	71.4	106.9	106.7
Ours	80.8	95.3	64.3	89.0	49.5	79.8	37.5	69.7	28.0	36.9	57.9	73.0	118.8	121.5
Ours[a]	80.9	95.6	64.7	89.4	49.9	80.3	37.6	70.1	28.3	37.3	58.2	73.3	121.9	124.3

[a] The results of ensemble models

as a whole, which can reveal that the proposed selective reading mechanism is superior to the recurrent reading mechanism.

Considering three kinds of knowledge to be stored in the memory, we can see that three kinds of knowledge can bring performance improvement over the basic model. And as a whole, the visual knowledge can boost the captioning model more significantly than semantic knowledge and fusion knowledge when both kinds of image features.

Compared with the State-of-the-art Method We further compare our models with several state-of-the-art methods, as shown in Table 3. We divide the state-of-the-art approaches into four categories. The first category of approaches simply inject the visual feature, i.e., a static vector, into the decoder. This category includes Google NIC [54], Deep VS [27], R-LSTM [7, 11], m-RNN [47], and ERD [60], as listed in the second block of Table 3. The second category of approaches include ATT [62], Soft/Hard-Attention [59], SCA-CNN [6], and Adaptive-Attention [45], which leverages the visual attention mechanism to enable the decoder to capture different aspects of visual information. The third category involves SCN-LSTM [18], MSM [61], RA+SF [25], and Show and Observe [4], which boost the captioning model with semantic-level information, such as attributes. The fifth block in Table 3 shows the fourth category of approaches which enhance the captioning model with reinforcement learning, including VS-LSTM [37], StackCap [20], SCST [51], SR-PL [43], and MIXER [50]. We also list the performance of top-down reported in [1] and memory-att of our previous conference version [3]. We also compare our method with the state-of-the-arts on the server of MS COCO.[2] The comparison is shown in Table 4.

From Tables 3 and 4, we can see that our single model can obtain better results than all baseline models, and our ensemble model can improve the performance further. The results can well demonstrate the effectiveness and the superiority of the proposed method, compared with the state-of-the-art approaches.

Qualitative Analysis

Some examples of the generated captions are shown in Fig. 4. For image 10, the top-down model correctly recognizes the cat, but fails to recognize its position, i.e., “refrigerator” instead of “kitchen,” while the proposed memory-captioning models can describe the scene successfully. As for images 1, 2, 3, 8, and 9, the top-down captioning model struggles to describe the objects relations, such as “standing” for “man” and “surfboard” in image 1, “in front of” for “building” and “clock tower” in image 2, and “standing on top of” for “cat” and “glass bottle” in image 8, while the proposed models can generate accurate captions. Besides, the proposed model can also be kind of superior to the basic top-down model in terms of novelty and the precision of descriptions, e.g., “overlooking” the ocean in image 3,

[2] https://competitions.codalab.org/competitions/3221#results

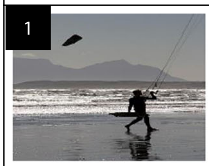
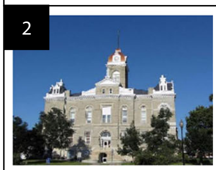
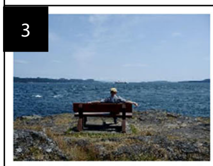
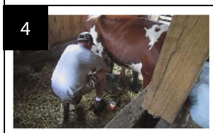
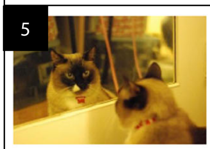
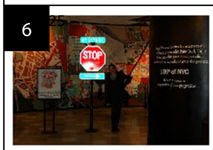
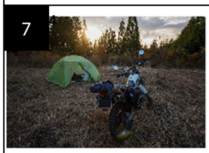
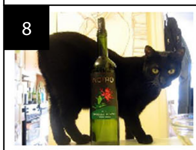
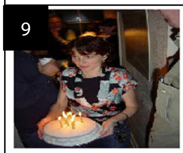
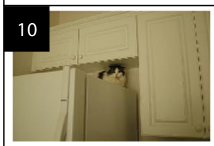
Image	Generated caption	Ground truth
1 	**topdown**: a man is standing on a surfboard in the ocean **MemRRM**: a man standing on the beach flying a kite **MemSRM**: caption: a man is flying a kite on the beach	The kite surfer is walking on the beach holding the kite. A man flying a kite over a sandy beach. A man flying a kite while walking on the beach. A man carrying something and flying a kite with his other hand. A man holding kite down by the ocean.
2 	**topdown**: a building with a clock tower in front of it **MemRRM**: a building with a clock tower on the top of **MemSRM**: a large building with a clock tower on the top'}	A concrete building with towers, a steep in the middle and a clock underneath. A large gray building with a clock tower surrounded by some trees. a big tower that is surrounded by trees A large gray building with a clock tower. A stone building that has a clock on the top.
3 	**topdown**: a couple of people sitting on a bench in the water **MemRRM**: a man sitting on a bench overlooking the ocean **MemSRM**: a man sitting on a bench next to the ocean	A man sitting on a bench right on a bay of water. A person sitting down a bench in front of the ocean. Man sitting on bench on rocky shore watching ship in distance. Man sitting on a bench overlooking the ocean. A man sitting on top of a bench near the ocean.
4 	**topdown**: a man is standing next to a cow **MemRRM**: a man is milking a cow in a barn **MemSRM**: a man is milking a cow in a fence	A man milking a cow during the day. A man milking a brown and white cow in barn. The guy with the white shirt and baseball cap is milking the cow. A man on a stool milking a cow. a man sitting on a stool milking a cow
5 	**topdown**: a cat sitting in front of a mirror **MemRRM**: a cat is looking at its reflection in a mirror **MemSRM**: a cat is looking at its reflection in a mirror	A cat looking at his reflection in the mirror. A cat that is looking in a mirror. A cat looking at itself in a mirror. A cat looking at itself adoringly in a mirror. A cat stares at itself in a mirror.
6 	**topdown**: a stop sign in front of a building **MemRRM**: a woman standing in front of a stop sign **MemSRM**: a woman standing in front of a stop sign	A woman is standing beside a stop sign in a museum. A woman standing next to a red stop sign with street signs. A woman is standing next to a stop sign A museum exhibit featuring a large stop sign. A woman leans on a lit up stop sign at an art exhibit.
7 	**topdown**: a motorcycle parked in the grass of a field **MemRRM**: a motorcycle parked in the grass with a tent **MemSRM**: a motorcycle parked in the grass next to a tent	A dirt bike parked near a tent in the woods. a green tent and a parked motorcycle and some trees Motorcycle parked in front of a tent with the sun going down behind them. a motorcycle parked next to a green tent in a field A parked motorcycle next to a green tent.
8 	**topdown**: a black cat standing on top of a glass bottle **MemRRM**: a black cat standing next to a bottle of wine **MemSRM**: a black cat standing on a table next to a bottle of wine	A black cat rubbing up against a bottle of wine. A black cat and a bottle of wine. A cat is walking past a bottle on a counter close up of a black cat neat a bottle of wine A black cat walks gingerly around an empty wine bottle.
9 	**topdown**: a group of people standing around a birthday cake **MemRRM**: a woman holding a birthday cake with candles on it **MemSRM**: a woman holding a birthday cake with candles on it	A woman holding a birthday cake with lit candles. a person holding a cake with lit candles A woman in a floral blouse carrying a cake. A woman is walking with a birthday cake at a party with people. A woman carrying a birthday cake with several lit candles on it.
10 	**topdown**: a cat sitting on the door of a kitchen **MemRRM**: a black and white cat sitting on top of a refrigerator **MemSRM**: a black and white cat sitting on top of a refrigerator	A cat in a kitchen on top of a refrigerator. A cat sitting on the top of a refrigerator hiding. A cat is standing on top of a fridge. A cat tucked between the top of a refrigerator and some cabinets. A cat that is sitting on top of a fridge.

Fig. 4 Examples of the proposed memory-enhanced models

"milking a cow in a barn" in image 4, and "looking at its reflection" in image 5.

Conclusion

Conventional captioning models usually adopt a recurrent neural network (RNN) to capture the long-term dependencies among words and generate the caption. However, it has been shown that RNNs have difficulties in remembering the knowledge too far in the past, even with some advanced techniques, e.g., the gating mechanism. In this paper, we propose a memory-enhanced captioning model for image captioning, named MemSRM, which can enhance the decoder's capability of memorizing knowledge by introducing an extern memory. Specifically, different kinds of knowledge, i.e., the visual knowledge, the semantic knowledge, and the fusion of both, are stored in the external memory. And for decoding, with the memory reading mechanism, i.e., the selective reading mechanism (SRM), the external memory can provide a feature representation of the past knowledge to help the prediction of the next words. To verify the effectiveness of the proposed model, we conduct extensive experiments and comparisons on a well-known image captioning dataset, i.e., MS COCO. In comparison with other state-of-the-art captioning models, the proposed MemSRM shows a substantial advantage in terms of performance improvement, which can well demonstrate that the proposed captioning model is effective and superior. For future studies, we plan to explore the writing mechanisms and apply the proposed memory-enhanced approach in other tasks, for example, image classification [12, 40], visual retrieval [57, 58].

Compliance with Ethical Standards

Ethical Approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S, Zhang L. Bottom-up and top-down attention for image captioning and vqa. arXiv:1707.07998. 2017.
2. Banerjee S, Lavie A. Meteor: an automatic metric for mt evaluation with improved correlation with human judgments. In Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. 2005. vol. 29, p. 65–72.
3. Chen H, Ding G, Lin Z, Guo Y, Han J. Attend to knowledge: memory-enhanced attention network for image captioning. In: International Conference on Brain Inspired Cognitive Systems. Springer; 2018. p. 161–71.
4. Chen H, Ding G, Lin Z, Zhao S, Han J. Show, observe and tell: attribute-driven attention model for image captioning. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18. International Joint Conferences on Artificial Intelligence Organization; 2018. p. 606–12.
5. Chen H, Ding G, Zhao S, Han J. Temporal-difference learning with sampling baseline for image captioning. AAAI Conference on Artificial Intelligence. 2018.
6. Chen L, Zhang H, Xiao J, Nie L, Shao J, Chua TS. Sca-cnn: spatial and channel-wise attention in convolutional networks for image captioning CVPR. 2017.
7. Chen M, Ding G, Zhao S, Chen H, Liu Q, Han J. Reference based LSTM for image captioning AAAI. 2017.
8. Cho K, Van Merriënboer B, Gülçehre Ç, Bahdanau D, Bougares F, Schwenk H, Bengio Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In Conference on Empirical Methods on Natural Language processing. 2014. p. 1724–34. 2014.
9. Devlin J, Cheng H, Fang H, Gupta S, Deng L, He X, Zweig G, Mitchell M. Language models for image captioning: the quirks and what works. In Annual Meeting of the Association for Computational Linguistics. 2015. p. 100–5. 2015.
10. Devlin J, Gupta S, Girshick R, Mitchell M, Zitnick CL. Exploring nearest neighbor approaches for image captioning. arXiv:1505.04467. 2015.
11. Ding G, Chen M, Zhao S, Chen H, Han J, Liu Q. Neural image caption generation with weighted training and reference. Cognitive Computation. 2018. https://doi.org/10.1007/s12559-018-9581-x.
12. Ding G, Guo Y, Chen K, Chu C, Han J, Dai Q. Decode: deep confidence network for robust image classification. IEEE Transactions on Image Processing. 2019.
13. Ding G, Guo Y, Zhou J, Gao Y. Large-scale cross-modality search via collective matrix factorization hashing. TIP. 2016;25(11):5427–40.
14. Dodds A. Rehabilitating blind and visually impaired people: a psychological approach. Springer. 2013.
15. Elliott D, Keller F. Image description using visual dependency representations. In Conference on Empirical Methods on Natural Language Processing. 2013. p. 1292–302.
16. Fakoor R, Mohamed Ar, Mitchell M, Kang SB, Kohli P. Memory-augmented attention modelling for videos. arXiv:1611.02261. 2016.
17. Farhadi A, Hejrati M, Sadeghi MA, Young P, Rashtchian C, Hockenmaier J, Forsyth D. Every picture tells a story: generating sentences from images. In European Conference on Computer Vision. 2010. p. 15–29.
18. Gan Z, Gan C, He X, Pu Y, Tran K, Gao J, Carin L, Deng L. Semantic compositional networks for visual captioning. In CVPR. 2017.
19. Gong Y, Wang L, Hodosh M, Hockenmaier J, Lazebnik S. Improving image-sentence embeddings using large weakly annotated photo collections. In European Conference on Computer Vision. 2014. p. 529–45.
20. Gu J, Cai J, Wang G, Chen T. Stack-captioning: coarse-to-fine learning for image captioning. In AAAI. 2018.
21. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016;00:770–778.
22. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735–1780.
23. Hodosh M, Young P, Hockenmaier J. Framing image description as a ranking task: data, models and evaluation metrics. J Artif Intell Res. 2013;47:853–99.
24. Jia X, Gavves E, Fernando B, Tuytelaars T. Guiding the long-short term memory model for image caption generation. In IEEE

International Conference on Computer Vision. 2015. p. 2407–15. 2015.
25. Jin J, Fu K, Cui R, Sha F, Zhang C. Aligning where to see and what to tell: image caption with region-based attention and scene factorization. arXiv:1506.06272. 2015.
26. Kaiser L, Nachum O, Roy A, Bengio S. Learning to remember rare events CVPR. 2017.
27. Karpathy A, Li FF. Deep visual-semantic alignments for generating image descriptions. In IEEE Conference on Computer Vision and Pattern Recognition. 2015. p. 3128–37.
28. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems. 2012. p. 1097–105.
29. Kulkarni G, Premraj V, Dhar S, Li S, Choi Y, Berg A, Berg T. Baby talk: understanding and generating simple image descriptions. In IEEE Conference on Computer Vision and Pattern Recognition. 2011. p. 1601–8.
30. Kuznetsova P, Ordonez V, Berg A, Berg T, Choi Y. Collective generation of natural image descriptions. In Annual Meeting of the Association for Computational Linguistics. 2012. p. 359–68.
31. Kuznetsova P, Ordonez V, Berg T, Choi Y. Treetalk: composition and compression of trees for image descriptions. Trans Assoc Comput Ling. 2014;2(10):351–62.
32. Lan X, Ma A, Yuen PC, Chellappa R. Joint sparse representation and robust feature-level fusion for multi-cue visual tracking. IEEE Trans Image Process. 2015;24(12):5826.
33. Lan X, Ye M, Shao R, Zhong B, Yuen PC, Zhou H. Learning modality-consistency feature templates: a robust rgb-infrared tracking system. IEEE Trans Ind Electron. 2019:1–1. https://doi.org/10.1109/TIE.2019.2898618.
34. Lan X, Ye M, Zhang S, Zhou H, Yuen PC. Modality-correlation-aware sparse representation for RGB-infrared object tracking. Pattern Recogn Lett. 2018. https://doi.org/10.1016/j.patrec.2018.10.002.
35. Lan X, Zhang S, Yuen PC, Chellappa R. Learning common and feature-specific patterns: a novel multiple-sparse-representation-based tracker. IEEE Trans Image Process. 2018;27(4):2022–37.
36. Li J, Zhang Z, He H. Hierarchical convolutional neural networks for EEG-based emotion recognition. Cogn Comput. 2018;10(2):368–80.
37. Li N, Chen Z. Image captioning with visual-semantic LSTM. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18. International Joint Conferences on Artificial Intelligence Organization; 2018. p. 793–799.
38. Li Y, Pan Q, Yang T, Wang S, Tang J, Cambria E. Learning word representations for sentiment analysis. Cogn Comput. 2017;843–851.
39. Lin CY, Hovy E. Automatic evaluation of summaries using n-gram co-occurrence statistics. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. Association for Computational Linguistics; 2003. p. 71–78.
40. Lin Z, Ding G, Han J, Shao L. End-to-end feature-aware label space encoding for multilabel classification with many classes. IEEE Trans Neural Netw Learn Syst. 2018;29(6):2472–87.
41. Lin Z, Ding G, Han J, Wang J. Cross-view retrieval via probability-based semantics-preserving hashing. IEEE Transactions on Cybernetics. 2016.
42. Liu S, Zhu Z, Ye N, Guadarrama S, Murphy K. Improved image captioning via policy gradient optimization of spider. In: Proceedings of the IEEE International Conference on Computer Vision. 2017. p. 873–81.
43. Liu X, Li H, Shao J, Chen D, Wang X. Show, tell and discriminate: image captioning by self-retrieval with partially labeled data. arXiv:1803.08314. 2018.
44. Liu Y, Vong C, Wong P. Extreme learning machine for huge hypotheses re-ranking in statistical machine translation. Cogn Comput. 2017;9(2):285–94.
45. Lu J, Xiong C, Parikh D, Socher R. Knowing when to look: adaptive attention via a visual sentinel for image captioning. 2017.
46. Luo R, Price B, Cohen S, Shakhnarovich G. Discriminability objective for training descriptive captions. arXiv:1803.04376. 2018.
47. Mao J, Xu W, Yang Y, Wang J, Yuille AL. Deep captioning with multimodal recurrent neural networks (m-RNN). In International Conference on Learning Representations. 2015.
48. Mitchell M, Han X, Dodge J, Mensch A, Goyal A, Berg A, Yamaguchi K, Berg T, Stratos K, Daumé HIII. Midge: generating image descriptions from computer vision detections. In Conference of the European Chapter of the Association for Computational Linguistics. 2012. p. 747–56.
49. Papineni K, Roukos S, Ward T, Zhu WJ. Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational linguistics. Association for Computational Linguistics; 2002. p. 311–8.
50. Ranzato M, Chopra S, Auli M, Zaremba W. Sequence level training with recurrent neural networks. arXiv:1511.06732. 2015.
51. Rennie SJ, Marcheret E, Mroueh Y, Ross J, Goel V. Self-critical sequence training for image captioning CVPR. 2016.
52. Roopnarine J, Johnson JE. Approaches to early childhood education. Merrill/Prentice Hall. 2013.
53. Vedantam R, Lawrence Zitnick C, Parikh D. Cider: consensus-based image description evaluation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015. p. 4566–75.
54. Vinyals O, Toshev A, Bengio S, Erhan D. Show and tell: a neural image caption generator. InCVPR. 2015 p. 3156–64.
55. Wang M, Lu Z, Li H, Liu Q. Memory-enhanced decoder for neural machine translation. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 2016. p. 278–86.
56. Weston J, Chopra S, Bordes A. Memory networks. arXiv:1410.3916. 2014.
57. Wu G, Han J, Guo Y, Liu L, Ding G, Ni Q, Shao L. Unsupervised deep video hashing via balanced code for large-scale video retrieval. IEEE Trans Image Process. 2019;28(4):1993–2007.
58. Wu G, Han J, Lin Z, Ding G, Zhang B, Ni Q. Joint image-text hashing for fast large-scale cross-media retrieval using self-supervised deep learning. IEEE Transactions on Industrial Electronics. 2018.
59. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, Zemel R, Bengio Y. Show, attend and tell: neural image caption generation with visual attention. In ICML. 2015. p. 2048–57.
60. Yang Z, Yuan Y, Wu Y, Salakhutdinov R, Cohen WW. Encode, review, and decode: reviewer module for caption generation NIPS. 2016.
61. Yao T, Pan Y, Li Y, Qiu Z, Mei T. Boosting image captioning with attributes. arXiv:1611.01646. 2016.
62. You Q, Jin H, Wang Z, Fang C, Luo J. Image captioning with semantic attention. In IEEE Conference on Computer Vision and Pattern Recognition. 2016. p. 4651–59. 2016.
63. Zhong G, Yan S, Huang K, Cai Y, Dong J. Reducing and stretching deep convolutional activation features for accurate image classification. Cogn Comput. 2018;10(1):179–86.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.