Recurrent Relational Memory Network for Unsupervised Image Captioning

Dan Guo, Yang Wang*, Peipei Song* and Meng Wang

Key Laboratory of Knowledge Engineering with Big Data (HFUT), Ministry of Education
School of Computer Science and Information Engineering, Hefei University of Technology (HFUT)
{guodan, yangwang}@hfut.edu.cn, {beta.songpp, eric.mengwang}@gmail.com

Abstract

Unsupervised image captioning with no annotations is an emerging challenge in computer vision, where the existing arts usually adopt GAN (Generative Adversarial Networks) models. In this paper, we propose a novel memory-based network rather than GAN, named Recurrent Relational Memory Network ($R^2M$). Unlike complicated and sensitive adversarial learning that non-ideally performs for long sentence generation, $R^2M$ implements a concepts-to-sentence memory translator through two-stage memory mechanisms: fusion and recurrent memories, correlating the relational reasoning between common visual concepts and the generated words for long periods. $R^2M$ encodes visual context through unsupervised training on images, while enabling the memory to learn from irrelevant textual corpus via supervised fashion. Our solution enjoys less learnable parameters and higher computational efficiency than GAN-based methods, which heavily bear parameter sensitivity. We experimentally validate the superiority of $R^2M$ than state-of-the-arts on all benchmark datasets.

1 Introduction

Traditional image captioning [Yao et al., 2019; Huang et al., 2019] requires full supervision of image-caption pairs annotated by humans. However, such full supervision is ridiculously expensive to acquire in cross-modal datasets. Recently, substantial researches tend to flexible constrained caption tasks, such as unpaired captioning [Gu et al., 2019; Guo et al., 2019] and unsupervised captioning [Feng et al., 2019] with weak or no supervised cues. It is challenging to leverage the independent image set and sentence corpus to train a reliable image captioning model; worse still, image captions usually cover specified or insufficient topics, e.g., the well-known benchmark MSCOCO images [Lin et al., 2014] just cover 80 object categories, raising up the challenges to generate rich semantical and grammatical sentences.

There are merely two unsupervised methods test on the disjointed image and text corpus data. [Feng et al., 2019]

*Corresponding authors.

Figure 1: The basic idea of our solution; that is to learn common concepts co-occurred in both image set and text corpus. We propose a light memory network named $R^2M$ to memorize and translate concepts to a sentence. $R^2M$ first imitates humans to listen to salient words and make sentences, exhibiting supervised learning (SPL) on text corpus. Then it locates visual concepts in images and makes sentences, representing unsupervised learning (UPL). The proposed memory demonstrates effective semantic reasoning for sequential learning.

As illustrated in Fig.1, $R^2M$ explores the latent relevant
semantic learning with the memory network, so as to enjoy the flexible and augmented memory capacity for both vision and natural language processing tasks [Pei et al., 2019; Fan et al., 2019]. Our intuition is that memory is proficient at storing and retrieving relational contexts to correlate input information [Huang and Wang, 2019], while inhibits vanishing gradients [Santoro et al., 2018; Fu et al., 2019]. As illustrated in Fig. 2, \( R^2M \) consists of two-stage memories, fusion memory (FM) and recurrent memory (RM). The relational reasoning based on FM and RM in our captioning process, not only considers the dependencies between words and common concepts, but also distills the useful context into the memory, retaining for long periods. Technically, \( R^2M \) performs the recurrent relational reasoning through multi-head self-attention and a relational gate (detailed in Section 2.2).

So far, FM and RM memories resolve the relational reasoning for text generation. As inspired, as shown in Figs.1 and 2, we develop a joint exploitation of supervised learning (SPL) and unsupervised learning (UPL) on the disjoint datasets. In particular, the SPL strategy is to learn the memories and make sentences from several salient words that separately appeared in text corpus, while the UPL training is to fine-tune the memories and make much more appropriate sentences about the visual context in the image. The supervised training on text corpus incorporates inductive semantic bias into the language model training. Turning to visual concepts in images without supervision cues, we explore a semantic matching (hinge-based triplet ranking) loss \( L^M_{tr} \) and reconstruction loss. These two losses encourage the cross-modal similarity score of image, along with the generated sentence to be larger than that of the other sentences. For \( L^M_{tr} \), we distinguish the negatives \((I', C_1)\) and \((I, C')\) from a positive image pair \((I, C)\), which is discussed in Section 2.4.

The major contributions are summarized as follows:

- Orthogonal to GAN-based architectures for unsupervised image captioning, we propose a novel light Recurrent Relational Memory Network (\( R^2M \)), which merely utilizes the attention-based memory (detailed in Section 2.2) to perform the relational semantics reasoning and reconstruction.

- A joint exploitation of Supervised learning on text corpus and Unsupervised learning on images is proposed. We optimize the cross-modal semantic alignment and reconstruction via an unsupervised manner to achieve a novel concepts-to-sentence translation.

- The proposed \( R^2M \) achieves better performances than state-of-the-arts on all the current unsupervised datasets: MSCOCO paired Shutterstock captions, Flickr30k paired MSCOCO captions and MSCOCO paired GCC (Google’s Conceptual Captions).

2 Proposed Method

In this section, we formally discuss our proposed \( R^2M \). The overall architecture of \( R^2M \) is depicted in Fig.2, which consists of three modules: encoder, decoder and reconstructor.

2.1 \( R^2M \) Encoder

We first discuss the encoder. A visual dictionary \( \mathcal{D} \) is learned ahead by using Faster R-CNN [Huang et al., 2017] trained on a public dataset OpenImages-v4 [Krasin et al., 2017; Kuznetsova et al., 2018] to cover the majority of common visual concepts in daily conversations, which is used to filter out visual concepts \( V = \{v_i\}_{i=1}^k \) of image \( I \) or sentence \( S \). After that, visual concepts \( V \) are randomly and sequentially incorporated into LSTM with their word embeddings, leading to the encoded vector \( v = v^1 \) or \( v^2 \) from \( I \) or \( S \).

2.2 \( R^2M \) Decoder

Details of decoder are illustrated in Fig.3. The effect of \( R^2M \).Decoder is to generate grammatical and semantical sentences from a few discrete words, e.g., translating “man” and “motorcycle” to “a man riding on the back of a motorcycle”. The set of visual concepts has no available grammar
and syntax contexts. Based on that, we train the model to think, infer and talk about as human beings. To address this issue, we propose a memory-based decoder, which not only considers the correlation between visual concepts and current generated word, but also captures the temporal dependencies and distills the underlying memory information.

**Relation Learning I: Fusion Memory (FM)**

As shown in Fig.3, the fusion memory (FM) in the decoder phase is used to learn the relationship between visual concepts and generated words, while recurrent memory (RM) in both decoder and reconstructor recurrently updates the memory to deliver useful semantics. At time step \( t \), FM learns the implicit relationship between the encoded concept vector \( v \in \mathbb{R}^d \) and previous generated word \( w_{t-1} \in \mathbb{R}^d \). We adopt a row-wise concatenation to acquire a joint feature matrix \( x_t = \begin{bmatrix} v; w_{t-1} \end{bmatrix} \in \mathbb{R}^{2 \times d} \), upon which multi-head self-attention [Vaswani et al., 2017] is performed. The intuition is to explore the correlation between \( v \) and \( w_{t-1} \). We consider the influences: \( v \rightarrow v, v \rightarrow w_{t-1}, w_{t-1} \rightarrow v \), and \( w_{t-1} \rightarrow v \). They are performed by the dot-product of query and key transformers of \( x_t \) as follows:

\[
A_{v \rightarrow w_{t-1}} = \begin{bmatrix} v \rightarrow v; v \rightarrow w_{t-1}; w_{t-1} \rightarrow v; w_{t-1} \rightarrow v \end{bmatrix} \in \mathbb{R}^{2 \times 2},
\]

\[
= \text{softmax} \left( \begin{bmatrix} x_q U_q & x_k U_k \end{bmatrix}^T / \sqrt{\lambda_1} \right),
\]

where \( U_q, U_k \in \mathbb{R}^{d \times d} \) are parameters of linear transformations of \( x_t \) (query and key); \( \lambda_1 \) is a scaling factor to balance the fusion attention distribution.

The cross interaction between \( v \) and \( w_{t-1} \) is calculated based on both attended weights and values as follows:

\[
\bar{x}_t = A_{v \rightarrow w_{t-1}} \cdot \begin{bmatrix} v; w_{t-1} \end{bmatrix} \cdot U_v \in \mathbb{R}^{2 \times d},
\]

where \( U_v \in \mathbb{R}^{d \times d} \) is another learnable parameter of linear transformations of \( x_t \) (value).

To ensure diverse and comprehensive attention guidance, we fuse \( v \) and \( w_{t-1} \) from \( H \) perspectives. There are \( H \) heads of independent attention executions. The outputs are concatenated into a new matrix \( x' \) as follows:

\[
x' = \left[ \bar{x}_t^1 \, \bar{x}_t^2 \right]^{\text{H}}_{h=1} = \left[ \bar{x}_t^1, \cdots, \bar{x}_t^H \right] \in \mathbb{R}^{2 \times (H \cdot d_v)},
\]

where \( \llbracket \) denotes column-wise concatenation. Finally, we use a fully-connection (linear) layer to convert the matrix \( x' \) into a fusion-aware feature \( f_t \) below:

\[
f_t = FC(x') \in \mathbb{R}^d.
\]

**Relation Learning II: Recurrent Memory (RM)**

Observing \( f_t \) at the \( t \)-th time step, recurrent memory recurrently learns a decoded memory variable \( M_t^d \) as shown in Fig.4. To distill the information worthy to retain in memory, we apply a relational gate for the recurrent memory updating among sequential learning. First, the multi-head self-attention is recycled to model latent transformers of previous memory state \( M_{t-1}^d \) and fusion-aware feature \( f_t \), where \( M_t^d \) is initialized with zero-padding. Note that we merely focus on the memory variation itself. The query is related to \( M_{t-1}^d \), key and value refer to \( [M_{t-1}^d; f_t] \), implying that the joint effect of \( [M_{t-1}^d; f_t] \) is learned under the guidance of \( M_{t-1}^d \). In this part, the detailed dimensions of parameters are shown in Fig.4.

\[
A_{f_t \rightarrow M_{t-1}^d} = \begin{bmatrix} M_{t-1}^d \rightarrow M_{t-1}^d; f_t \rightarrow M_{t-1}^d \end{bmatrix} = \text{softmax} \left( \begin{bmatrix} M_{t-1}^d \cdot W_q^h \cdot [f_t]W_k^h \end{bmatrix} / \sqrt{\lambda_2} \right),
\]

\[
M_t^d = A_{f_t \rightarrow M_{t-1}^d} \cdot \begin{bmatrix} [M_{t-1}^d; f_t]W_k^h \end{bmatrix}^{\text{H}}_{h=1},
\]

where \( W_q^h, W_k^h \in \mathbb{R}^{d \times d_v} \) and \( W_v^h \in \mathbb{R}^{d \times d_v} \) are learnable parameters, and \( \lambda_2 \) is the scaling factor to balance the attention distribution in RM.

**Relational Gate** To model the temporal dependencies among the adjacent memories, we update the memory state in a gated recurrent manner. Specifically, we apply input gate \( g_i,t \) and forget gate \( g_f,t \) to balance the memory updating from the current memory gain \( M_t^d \), and original memory \( M_{t-1}^d \), respectively. Both \( g_i,t \) and \( g_f,t \) are affected by \( f_t \) and \( M_{t-1}^d \):

\[
\begin{cases}
\ g_i,t = \sigma(W_i \cdot f_t + U_i \cdot \text{tanh}(M_{t-1}^d) + b_i), \\
\ g_f,t = \sigma(W_f \cdot f_t + U_f \cdot \text{tanh}(M_{t-1}^d) + b_f), \\
\ M_t^d = g_i,t \odot \text{tanh}(M_{t-1}^d) + g_f,t \odot M_{t-1}^d,
\end{cases}
\]

where \( \odot \) and \( \sigma \) denote dot product and sigmoid functions. Based on the updated memory \( M_t^d \), RM outputs the word \( w_t \):

\[
w_t = \text{argmax} \{ \text{softmax}(W_d \cdot M_t^d) \},
\]

where \( W_d \) is a learnable matrix that maps \( M_t^d \) to a vector with the dictionary size.

**2.3 R²M. Reconstructor**

So far, the decoder yields a pipeline to translate discrete visual concepts into a formal sentence. Here, we attempt to
where function $M$ treated as the reconstructed vector of concepts, denoted as $|||$. We train the concepts-to-sentence decoder $\tilde{L}_S$ supervised on Text Corpus.

2.4 Training

Supervision Learning on Text Corpus

We train the log-likelihood of the generated sentences with original corpus sentences:

$$L_{XE} = -\sum_{t=1}^{len} \log p(w_t|w_{t-1}).$$ (10)

For $R^2 M.Reconstructor$, there is the reconstructed vector $\tilde{v}^S$ corresponding to sentence $S$. We align it in an unsupervised mode. The full objective on text corpus is:

$$L_S = L_{XE} + \beta L_{rec}^S$$ (11)

where $L_{rec}^S = ||v^S - \tilde{v}^S||_{L2}^2$, $\beta$ is a hyper-parameter, and $||.||_{L2}$ denotes the L2-norm loss.

Unsupervised Visual Alignment on Images

The remaining question is how to achieve a better generalization ability with no supervision cues for image captioning? To answer this question, we adopt a hinge-based triplet ranking loss $L_M^t$, which encourages the semantic relevance of $(I, C_I)$ to be much larger than other negative examples. We choose the hardest negatives $I'$ and $C'$ for a positive pair $(I, C_I)$, and perform $L_M^t$ as follows:

$$L_M^t = [m - S(I, C_I) + S(I', C_I) +]_{+} + [m - S(I, C_I) + S(I, C'_I)]_{+}$$

s.t., $I' = \arg\max_{I'} S(I', C_I), C' = \arg\max S(I, C'_I),$

where $[x]_{+} = \max(x, 0), S(\cdot)$ is the similarity function calculated by inner product, and $m$ serves as a margin parameter. $S(I, C_I) = S(\tilde{v}^I, \tilde{v}^I)$, where $\tilde{v}^I$ is the visual feature of image $I$ extracted by Inception-V4 [Szegedy et al., 2017] and $\tilde{v}^I$ is the reconstructed vector by the RM unit. For computational efficiency, we search the negatives $I'$ and $C'$ with each mini-batch instead of the entire training set.

Besides, the image reconstruction loss $L_{rec}^I$ is utilized to train the model. The full objective on images is:

$$L_I = L_{rec}^I + \gamma L_{rec}^I,$$ (13)

where $L_{rec}^I = ||v^I - \tilde{v}^I||_{L2}^2$ and $\gamma$ is a hyper-parameter.

3 Experiments

3.1 Dataset and Metrics

We test all the existing unsupervised image captioning datasets, including (1) MSCOCO images [Lin et al., 2014] paired with Shutterstock captions [Feng et al., 2019]; and (2) Flickr30k images [Young et al., 2014] paired with MSCOCO captions and (3) MSCOCO images paired with Google’s Conceptual Captions (GCC) [Sharma et al., 2018; Laina et al., 2019]. In the test splits of datasets, each image has five ground-truth captions.

3.2 Implementation Details

We split each image set and filter captions as Feng et al., 2019; Laina et al., 2019. The visual dictionary $D$ in Fig.2 is collected by a pre-trained Faster R-CNN [Huang et al., 2017] OpenImages-v4 [Krasin et al., 2017; Kuznetsova et al., 2018]. We merge the visual concepts in $D$ and words in training captions into a large vocabulary, to cover the majority of the to-be-generated words. The vocabulary sizes of the three datasets are 18,679/11,335/10,652, respectively, initialized with zero-padding, and $len$ is the length of the generated caption $C$. The last output of $R^2 M.Reconstructor$ is treated as the reconstructed vector of concepts, denoted as $\tilde{v}^I$ or $\tilde{v}^S$ corresponding to image $I$ or sentence $S$.

2.4 Training

Supervision Learning on Text Corpus

We train the concepts-to-sentence decoder $R^2 M.Decoder$ by maximizing log-likelihood of the generated sentences with original corpus sentences:

$$L_{XE} = -\sum_{t=1}^{len} \log p(w_t|w_{t-1}).$$ (10)

For $R^2 M.Reconstructor$, there is the reconstructed vector $\tilde{v}^S$ corresponding to sentence $S$. We align it in an unsupervised mode. The full objective on text corpus is:

$$L_S = L_{XE} + \beta L_{rec}^S$$ (11)

where $L_{rec}^S = ||v^S - \tilde{v}^S||_{L2}^2$, $\beta$ is a hyper-parameter, and $||.||_{L2}$ denotes the L2-norm loss.
Table 1: Performance comparison with the state-of-the-art methods. The best performance is marked with bold face.

Table 2: Ablation studies of $R^2M$ with different memory settings. The best performance is marked with bold face.

Table 2: Ablation studies of $R^2M$ with different losses. The best performance is marked with bold face.

3.3 Experimental Results and Analysis

Comparison with the State-of-the-arts

$R^2M$ exhibits large improvements across all the metrics. Both UC-GAN [Feng et al., 2019] and SME-GAN [Lama et al., 2019] rely on complicated GAN training strategies, whereas ours $R^2M$ is a memory solution. As shown in Table 1, $R^2M$ upgrades BLEU-4 (B-4) by 14.3%, 48.1% and 27.7% on three datasets, where BLEU-4 involves 4-gram phrases. It implies the stronger capacity of $R^2M$ to learn long-range dependencies than others. $R^2M$ also raises CIDEr/SPICE, from 28.6/9.1 to (29.0/9.1), 9.9/7.5 on MSCOCO & Shutterstock, respectively.

Ablation Study of $R^2M$

To verify each component in $R^2M$, we propose the ablation study. (1) Effect of FM. Compared to the entire $R^2M$, the performance of “D w/o FM” drops significantly, e.g., with 18.3%, 17.7% and 18.4% reduction of CIDEr (C) on three datasets in Table 2. FM effectively implements the implicit correlation between visual concept vector and word embedding. (2) Effect of memory in RM. For Table 2, either “D, R or D&R w/o Memory in RM” suffers from worse performance, e.g., on dataset MSCOCO→GCC, dropping the CIDEr from 29.3 to 27.1, 28.1 and 27.8. RM excels at storing and retrieving information across time than classical LSTM, to effectively handle sequential learning. (3) Effect of Loss. In each block diagram of Table 3, the first line records the result of model trained with only $L_{XE}$ on text corpus. Note that this baseline is competitive and outperforms the existing methods. The SPICE (S) is increased by around 11.1%, 9.3% and 25.7% compared on three datasets. By gradually incorporating $L^v_{rec} I^v_{rec}$ and $L^r_{rec}$ the model performs much better. The CIDEr gradually raises from 25.4 to 27.0, 28.9 and 29.0 on MSCOCO→Shutterstock. Especially after the assistance of semantic matching loss $L^s_M$, the CIDEr is significantly improved, nearly 7.0%, 6.6% and 3.2% on all datasets.

Qualitative Results

Visualization of Attention Weights in FM & RM. Fig.5 illustrates an example of memory learning in $R^2M$.Decoder, which is interpretable. FM displays the average weight of multi-head attention, while RM offers $H = 2$ heads attention. With the beginning token <$start$> at $t=0$ and the encoded concept vector $v$, FM pays more attention to the richer semantics $v$. And at time $t = 2$, FM focuses much more on the previous word $portraits_{t=1}$ as $portraits$ is the first generated concept and deserves more attention. Then, we discuss the interpretation of RM. Taking previous word $sunglasses_{t=9}$ as an example, it affects the generation of $sunglasses_{t=10}$ more influentially than the memory $M^d_t$. However, at $t = 11$, under the previous cue $sunglasses_{t=10}$, the model infers a relational conjunction $and_{t=11}$ by mainly recalling $M^d_{11}$. The same situation holds at the last time $umbrella_{t=14}$, there is no relevance cues to be found from $M^d_{14}$, the model decides to terminate the entire generation process.
Field offer another example in Fig.6 (c). Both the new words are not detected visual concepts. Following the textual cues learning from text corpus, $R^2 M$ acquires the knowledge to infer a new concept-independent concept $beach$; however, it is irrelevant. After unsupervised visual alignment learning, the caption finally outputs a new word $field$ instead of $beach$. $R^2 M$ is effective to infer promising descriptions about images without annotated captions.

Figure 5: Visualization results of attention weights in FM and RM units of $R^2 M$. $Decoder$. The weights reflect how much attention the model pays to each input variable.

Figure 6: Qualitative examples of $R^2 M$. Visual concepts are detected by Faster R-CNN. Words in \{red, brown, blue, green\} fonts mark the incorrectly recognized, irrelevant, and new detected concepts, and the adjective, respectively.

**Visualization of Generated Captions.** We detect visual concepts and their scores by Faster R-CNN. As shown in Fig.6 (a), $phone$ is an incorrectly detected object with a high score 0.75. While performing training on text corpus with $L_{*X}$ and $L_{rec}^S$, $R^2 M$ translates discrete concepts to a sentence, still containing $phone$. With further unsupervised training on images over $L_{*M}$ and $L_{rec}^I$, $R^2 M$ automatically eliminates the wrong concept. By contrast, exemplified in Fig.6 (b), $clothing$ is a correctly identified concept, but irrelevant to salient visual regions of the image. $R^2 M$ eliminates the redundant visual concepts yet. Moreover, there are new learned concepts $beach$ and an adjective $young$ from all the joint SPL and UPL semantic learning. To strengthen the intuition that $R^2 M$ can extrapolate beyond the concepts in the images, we offer another example in Fig.6 (c). Both the new words $beach$ and $field$ are undetected visual concepts. Following the textual cues learning from text corpus, $R^2 M$ acquires the knowledge to infer a new concept-independent concept $beach$; however, it is irrelevant. After unsupervised visual alignment learning, the caption finally outputs a new word $field$ instead of $beach$. $R^2 M$ is effective to infer promising descriptions about images without annotated captions.

We also extend the experiments with new corpora with different language styles, such as VQA-v2 [Antol et al., 2015] and SentiCap [Mathews et al., 2016], involving the questions about the visual content and sentiment captions. For our experiments, 1,105,904 questions provided by VQA-v2 and 4,892 positive captions of SentiCap are respectively trained as text corpus. As shown in Fig.7, $R^2 M$ also excels at questioning and describing images with positive emotion.

**4 Conclusion**

This paper proposes a novel recurrent relational memory network ($R^2 M$) for unsupervised image captioning with low cost of supervision. $R^2 M$ is a lightweight network, characterizing self-attention and a relational gate to design the fusion and recurrent memory for long-term semantic generation. Experimental results show that the $R^2 M$ surpasses the state-of-the-arts on three benchmark datasets.

**Acknowledgments**

This work is supported by the National Key Research and Development Program of China under grant 2018YFB0804205, and the National Natural Science Foundation of China (NSFC) under grants 61806035, U1936217, 61725203, 61732008, and 61876058.
References


