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] proposed an architecture comprising of an image encoder, a sentence generator, a discriminator with adversarial loss and concept reward. [Laina et al., 2019] learned a joint semantic embedding space, where either images or sentences were transformed. A discriminator was then designed to judge where the embedding feature came from, image or sentence domain. Both of them resolved the task with adversarial training, while obeying the usage of GAN (Generative Adversarial Networks) in unsupervised mode [Lample et al., 2018; Donahue and Simonyan, 2019; Yang et al., 2018]. As it is widely known, current GAN methods based on ordinary recurrent models (e.g., LSTM) always employ RL heuristics and are quite sensitive to parameter initializations and hyper-parameter settings [Nie et al., 2019].

Orthogonal to above GAN-based models, in this paper, we propose a novel memory-based solution, named Recurrent Relational Memory Network ($R^2M$). The novelty of $R^2M$ lies in its exploitation on visual concepts and describing image via memory, serving as a concepts-to-sentence memory translator to learn the textual knowledge from discrete common concepts in diverse sentences, meanwhile being capable of making sentences with correctly semantic and grammar syntax rules.
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 important visual semantics $\tilde{v}^S$ or $\tilde{v}^I$ corresponding to respective $C$. The supervised loss $L_S$ on text corpus is optimized via cross-entropy loss $L_{XE}(S, C)$ and reconstruction loss $L_{rec}(v^S, \tilde{v}^S)$. Then unsupervised loss $L_I$ for images is optimized by semantic matching (triplet semantic ranking) loss $L_M(v^I, \tilde{v}^I)$ and reconstruction loss $L_{rec}(v^I, \tilde{v}^I)$.

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.
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 $V = \{v_i\}_{i=1}^{|V|}$ 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 $v = v^I$ or $v^S$ from $I$ or $S$.

### 2.2 R²M. Decoder

Details of decoder are illustrated in Fig.3. The effect of R²M. 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 $x_t = [v; w_{t-1}] \in \mathbb{R}^{2 \times d}$, 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: $v \rightarrow v, v \rightarrow w_{t-1}, w_{t-1} \rightarrow w_{t-1}$ and $w_{t-1} \rightarrow v$. They are performed by the dot-product of query and key transformers of $x_t$ as follows:

$$A_{v \leftrightarrow w_{t-1}} = \begin{bmatrix} v \rightarrow v, & w_{t-1} \rightarrow v \\ v \rightarrow w_{t-1}, & w_{t-1} \rightarrow w_{t-1} \end{bmatrix} \in \mathbb{R}^{2 \times 2} = \text{softmax}(x_t U_q \cdot (x_t U_k)^\top / \sqrt{\lambda_1}), \quad (1)$$

where $U_q, U_k \in \mathbb{R}^{d \times d_k}$ 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 attentions and values as follows:

$$\hat{x}_t = A_{v \leftrightarrow w_{t-1}} \cdot (\{v; w_{t-1}\} \cdot U_v) \in \mathbb{R}^{2 \times d_v}, \quad (2)$$

where $U_v \in \mathbb{R}^{d \times d_v}$ 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 $x'_t$ as follows:

$$x'_t = [x_t^h][h=1] = [x_t^1, \cdots, x_t^H] \in \mathbb{R}^{2 \times (H \cdot d_v),} \quad (3)$$

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

$$f_t = FC(x'_t) \in \mathbb{R}^d. \quad (4)$$

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

Observing $f_t$ at the $t$-th time step, RM recurrently learns a decoded memory variable $M^d_t$ 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 $M^d_{t-1}$, and fusion-aware feature $f_t$, where $M^d_0$ is initialized with zero-padding. Note that we merely focus on the memory variation itself. The query is related to $M^d_{t-1}$, key and value refer to $[M^d_{t-1}; f_t]$, implying that the joint effect of $[M^d_{t-1}; f_t]$ is learned under the guidance of $M^d_{t-1}$. In this part, the detailed dimensions of parameters are shown in Fig.4.

$$A_{f_t \rightarrow M^d_{t-1}} = [M^d_{t-1} \rightarrow M^d_{t-1}, f_t \rightarrow M^d_{t-1}]^h = \text{softmax}(M^d_{t-1}W^h_q \cdot ([M^d_{t-1}; f_t]W^h_k)^\top / \sqrt{\lambda_2}), \quad \text{query key} \quad (5)$$

$$M^d_t = [A_{f_t \rightarrow M^d_{t-1}} \cdot ([M^d_{t-1}; f_t]W^h_v)] \in \mathbb{R}^{d \times d_v}. \quad (6)$$

where $W^h_q, W^h_k, W^h_v \in \mathbb{R}^{d \times d_k}$ and $W^h_v \in \mathbb{R}^{d \times d_v}$ are learnable parameters, and $\lambda_2$ is the scaling factor to balance the attention distribution in RM.

**Module $\psi$** $M^d_t$ is then fed into two residual connection layers and one row-wise multi-layer perception (MLP) with layer normalization. Thus, we achieve a memory gain $M^d_t$.

**Relational Gate** To model the temporal dependencies along 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 $\tilde{M}^d_t$ and original memory $M^d_{t-1}$, respectively. Both $g_{i,t}$ and $g_{f,t}$ are affected by $f_t$ and $M^d_{t-1}$.

\[
g_{i,t} = \sigma(W_{i} \cdot f_t + U_{i} \cdot \tanh(M^d_{t-1}) + b_{i})
\]

\[
g_{f,t} = \sigma(W_{f} \cdot f_t + U_{f} \cdot \tanh(M^d_{t-1}) + b_{f})
\]

\[
M^d_t = g_{i,t} \odot \tanh(\tilde{M}^d_t) + g_{f,t} \odot M^d_{t-1},
\]

where $\odot$ and $\sigma$ denote dot product and sigmoid functions.

Based on the updated memory $M^d_t$, RM outputs the word $w_t$:

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

where $W_d$ is a learnable matrix that maps $M^d_t$ 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 ensure that R²M can talk about correct contents. As inspired, we reversely reconstruct the concept semantics, i.e., rebuilding the crucial concept semantics from the generated sentence. We adopt the memory unit RM to compose the R²M. Reconstructor. Note that learnable parameters of RM in R²M. Decoder and Reconstructor are completely different.

If we define the RM operation in R²M. Decoder as a function $M^d_t = \text{RM}(M^d_{t-1}, f_t)$ involving Eqs. 5~7, the R²M. Reconstructor operation is formulated as follows:

\[
M^r_t = \text{RM}(M^r_{t-1}, M^d_t), t \in \{0, \cdots, \text{len}\},
\]

where $M^r_t$ indicates a reconstructed memory at time $t$, $M^r_0$ is initialized with zero-padding, and $\text{len}$ is the length of the generated caption $C$. The last output of R²M. Reconstructor is treated as the reconstructed vector of concepts, denoted as $\tilde{v}^r$ 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 $R^2M.\text{Decoder}$ by maximizing log-likelihood of the generated sentences with original corpus sentences:

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

For $R^2M.\text{Reconstructor}$, 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:

\[
L_S = L_{XE} + \beta L^S_{\text{rec}},
\]

where $L^S_{\text{rec}} = ||v^S - \tilde{v}^S||_2^2$, $\beta$ is a hyper-parameter, and $||.||_2^2$ 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 $L_M$, 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 $L_M$ as follows:

\[
L_M = [m - S(I, C_I) + S(I', C_I)]_+ + [m - S(I, C_I) + S(I, C')]_+ - [m - S(I', C_I) + S(I', C')]_+, \text{s.t.}, I' = \text{argmax}_{l \neq I} S(l, C_I), C' = \text{argmax}_{l \neq I} S(I, C_l),
\]

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 $\tilde{v}^r$ is the reconstructed vector by the RM unit. For computational efficiency, we search the negatives $I'$ and $C'$ within each mini-batch instead of the entire training set.

Besides, the image reconstruction loss $L^I_{\text{rec}}$ is utilized to train the model. The full objective on images is:

\[
L_I = L^I_M + \gamma L^I_{\text{rec}},
\]

where $L^I_{\text{rec}} = ||v^I - \tilde{v}^I||_2^2$ 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.
Table 1: Performance comparison with the state-of-the-art methods. The best performance is marked with bold face.

| Dataset          | Method                      | B-1 | B-2 | B-3 | B-4 | METEOR | ROUGE | CIDEr | SPICE |
|------------------|-----------------------------|-----|-----|-----|-----|--------|-------|-------|-------|
| MSCOCO→Shutterstock | UC-GAN [Feng et al., 2019]  | 41.0| 22.5| 11.2| 5.6 | 12.4   | 28.7  | 28.6  | 8.1   |
|                  | $R^2M$                      |     |     |     |     |        |       |       |       |
| Flickr30k→MSCOCO  | SME-GAN [Laina et al., 2019] | 44.0| 25.4| 12.7| 6.4 | 13.0   | 31.3  | 31.3  | 9.1   |
|                  | $R^2M$                      |     |     |     |     |        |       |       |       |
| MSCOCO→GCC       | SME-GAN [Laina et al., 2019] | 51.2| 29.5| 15.4| 8.3 | 14.0   | 35.0  | 35.0  | 9.6   |

Table 2: Ablation studies of $R^2M$ with different memory settings. The best performance is marked with bold face. (1) In “D w/o FM”, $f_1$ is calculated by a linear layer on the concatenation of $v$ and $w$; (2) “D w/o Memory in RM” replaces the RM operation by LSTM in both $R^2M$.Decoder and Reconstructor; (3) “D w/o Memory in RM” and (4) “R w/o Memory in RM” replaces RM by LSTM in respective $R^2M$.Decoder and Reconstructor.

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, including tokens <start>, <end>, and <UNK>. For experimental setting, we filter out visual concepts form images with the detected score $\geq 0.3$. Both the sizes of LSTM and RM memory are set to $N = 1$ and $d = 512$. The parameters of multi-head self-attention are $H = 2$, $d_k = d_v = 256$, and $d_q = d_t = 256$. The margin in Eq. 12 is $m = 0.2$. Adam optimizer is adopted with batch size of 256. For three datasets, hyper-parameters ($\beta$, $\gamma$) are set to (1, 1), (1, 1), (0.2, 0.2). We train the model with a loss $L_{XE}$ under learning rate $10^{-4}$, while fine-tune it with the joint loss $L_I$. After that, $L_M^I$ is used to train with a learning rate $10^{-5}$. Finally, we jointly train the model with $L_I$. In the test process, we use the beam search tactic [Anderson et al., 2017] with width of 3.

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 [Laina 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 28.6/8.1 to (29.0/9.1), 9.9/7.5, 18.1/8.3 and 22.7/7.4 (29.3/9.6). The promising improvements demonstrate the consistency of superior performances. With the released code of UC-GAN [Feng et al., 2019] on the MSCOCO→Shutterstock dataset, here is an efficiency comparison: $R^2M$ vs. UC-GAN $\approx 35$ min vs. $34$ hours. $R^2M$ also enjoys higher computational efficiency.

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 DkR 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 $L_{rec}^S$, $L_{rec}^I$ and $L_{rec}^M$, 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 $L_M$, 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 a more interpretable visualization of attention weights in FM & RM. Qualitative Results

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 [Laina 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 28.6/8.1 to (29.0/9.1), 9.9/7.5, 18.1/8.3 and 22.7/7.4 (29.3/9.6). The promising improvements demonstrate the consistency of superior performances. With the released code of UC-GAN [Feng et al., 2019] on the MSCOCO→Shutterstock dataset, here is an efficiency comparison: $R^2M$ vs. UC-GAN $\approx 35$ min vs. $34$ hours. $R^2M$ also enjoys higher computational efficiency.

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 DkR 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 $L_{rec}^S$, $L_{rec}^I$ and $L_{rec}^M$, 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 $L_M$, 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> 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 “portrait” as “portrait” is the first generated concept and deserves more attention. Then, we discuss the
interpretation of RM. Taking previous word wearing\textsubscript{(t=9)} as an example, it affects the generation of sunglasses\textsubscript{(t=10)} more influentially than the memory $M'_t$. However, at $t=11$, under the previous cue sunglasses\textsubscript{(t=10)}, the model infers a relational conjunction and\textsubscript{(t=11)} by mainly recalling $M'_t$. The same situation holds at the last time umbrella\textsubscript{(t=14)}, there is no relevance cues to be found from $M'_t$, the model decides to terminate the entire generation process.

**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 $L_{\text{XE}}$ and $L_{\text{rec}}$, $R^2M$ translates discrete concepts to a sentence, still containing phone. With further unsupervised training on images over $L_M$ and $L_{\text{rec}}$, $R^2M$ 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. $R^2M$ 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 $R^2M$ 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, $R^2M$ acquires the knowledge to infer a new context-independent concept beach; however, it is irrelevant. After unsupervised visual alignment learning, the caption finally outputs a new word field instead of beach. $R^2M$ 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. SPL on texts: man with a surfboard on the beach. UPL on images: fresh carrot on a wooden background. GT: A person at the beach with a surfboard.

**4 Conclusion**

This paper proposes a novel recurrent relational memory network ($R^2M$) for unsupervised image captioning with low cost of supervision. $R^2M$ 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 $R^2M$ 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

[Anderson et al., 2017] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Guided open vocabulary image captioning with constrained beam search. In EMNLP, page 936–945, 2017.

[Antol et al., 2015] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In ICCV, pages 2425–2433, 2015.

[Donahue and Simonyan, 2019] Jeff Donahue and Karen Simonyan. Large scale adversarial representation learning. In NeurIPS, pages 10541–10551, 2019.

[Fan et al., 2019] Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. Heterogeneous memory enhanced multimodal attention model for video question answering. In CVPR, pages 1999–2007, 2019.

[Feng et al., 2019] Yang Feng, Lin Ma, Wei Liu, and Jiebo Luo. Unsupervised image captioning. In CVPR, pages 4125–4134, 2019.

[Fu et al., 2019] Canniao Fu, Wenjie Pei, Qiong Cao, Chaopeng Zhang, Yong Zhao, Xiaoyong Shen, and Yu-Wing Tai. Non-local recurrent neural memory for supervised sequence modeling. In ICCV, pages 6311–6320, 2019.

[Gu et al., 2019] Jiuxiang Gu, Shafiq Joty, Jianfei Cai, Handong Zhao, Xu Yang, and Gang Wang. Unpaired image captioning via scene graph alignments. In ICCV, pages 10323–10332, 2019.

[Guo et al., 2019] Longteng Guo, Jing Liu, Peng Yao, Jiawei Li, and Hanqing Lu. Mcap: Multi-style image captioning with unpaired stylized text. In CVPR, pages 4204–4213, 2019.

[Huang and Wang, 2019] Yan Huang and Liang Wang. Acmn: Aligned cross-modal memory for few-shot image and sentence matching. In ICCV, pages 5774–5783, 2019.

[Huang et al., 2017] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zhiguo Wojna, Yang Song, Sergio Guadarrama, et al. Speed/accuracy trade-offs for modern convolutional object detectors. In CVPR, pages 7310–7311, 2017.

[Huang et al., 2019] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In ICCV, pages 4634–4643, 2019.

[Krasin et al., 2017] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, et al. Openimages: A public dataset for large-scale multi-label and multi-class image classification. Dataset available from https://github.com/openimages, 2:3, 2017.

[Kuznetsova et al., 2018] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. arXiv preprint arXiv:1811.00982, 2018.

[Laina et al., 2019] Ira Laina, Christian Rupprecht, and Nassir Navab. Towards unsupervised image captioning with shared multimodal embeddings. In ICCV, pages 7414–7424, 2019.

[Lample et al., 2018] Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc’Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. In ICLR, 2018.

[Lin et al., 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In ECCV, pages 740–755, 2014.

[Mathews et al., 2016] Alexander Patrick Mathews, Lexing Xie, and Xuming He. Senticap: Generating image descriptions with sentiments. In AAAI, pages 3574–3580, 2016.

[Nie et al., 2019] Weili Nie, Nina Narodytska, and Ankit Patel. Relgan: Relational generative adversarial networks for text generation. In ICLR, 2019.

[Pei et al., 2017] Wenjie Pei, Jiuyuan Zhang, Xiangrong Wang, Lei Ke, Xiaoyong Shen, and Yu-Wing Tai. Memory-attended recurrent network for video captioning. In CVPR, pages 8347–8356, 2019.

[Santoro et al., 2018] Adam Santoro, Ryan Faulkner, David Raposo, Jack Rae, Mike Chrzanowski, Theophane Weber, Daan Wierstra, Oriol Vinyals, Razvan Pascanu, and Timothy Lillicrap. Relational recurrent neural networks. In NeurIPS, pages 7299–7310, 2018.
[Sharma et al., 2018] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In ACL, pages 2556–2565, 2018.

[Szegedy et al., 2017] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In AAAI, pages 4278–4284, 2017.

[Vaswani et al., 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In NeurIPS, pages 5998–6008, 2017.

[Yang et al., 2018] Zhen Yang, Wei Chen, Feng Wang, and Bo Xu. Unsupervised neural machine translation with weight sharing. In ACL, pages 46–55, 2018.

[Yao et al., 2019] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Hierarchy parsing for image captioning. In ICCV, pages 2621–2629, 2019.

[Young et al., 2014] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. TACL, 2:67–78, 2014.