

Contextual and selective attention networks for image captioning

Jing WANG¹, Yehao LI², Yingwei PAN², Ting YAO², Jinhui TANG^{1*} & Tao MEI²¹*School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China;*²*JD Explore Academy, Beijing 100101, China*

Received 16 September 2020/Revised 27 March 2022/Accepted 3 June 2022/Published online 18 November 2022

Abstract The steady momentum of innovations has convincingly demonstrated the high capability of attention mechanisms for the sequence to sequence learning. Nevertheless, the computation of attention across a sequence is often independent in either hard or soft mode, thereby resulting in undesired effects such as repeated modeling. In this paper, we introduce a new design to holistically explore the interdependencies between attention histories and locally emphasize the strong focus of each attention on image captioning. Specifically, we present a contextual and selective attention network (namely CoSA-Net) that novelly memorizes contextual attention and brings out the principal components from each attention. Technically, CoSA-Net writes/updates the attended image region features into memory and reads from memory when measuring attention in the next time step to leverage contextual knowledge. Only the regions with the top- k highest attention scores are selected, and each region feature is individually employed to compute an output distribution. The final output is an attention-weighted mixture of all k distributions. In turn, the attention is then upgraded by the posterior distribution conditioned on the output. Our CoSA-Net is appealing given that it is pluggable to the sentence decoder in any neural captioning model. Extensive experiments on the COCO image captioning dataset demonstrate the superiority of CoSA-Net. More remarkably, integrating CoSA-Net to a one-layer long short-term memory (LSTM) decoder increases CIDEr-D performance from 125.2% to 128.5% on the COCO Karpathy test split. When further endowing a two-layer LSTM decoder with CoSA-Net, the CIDEr-D score is boosted to 129.5%.

Keywords image captioning, hybrid attention, contextual attention**Citation** Wang J, Li Y H, Pan Y W, et al. Contextual and selective attention networks for image captioning. *Sci China Inf Sci*, 2022, 65(12): 222103, <https://doi.org/10.1007/s11432-020-3523-6>

1 Introduction

Vision and language are two fundamental capabilities of human intelligence. Their interactions support the unique human capacity to discuss what is seen or imagined in a picture given a natural-language description. The recent development of deep learning has successfully pushed the limits of vision and language. Image captioning, as one of the “hottest” topics in this area over the past five years, is for automatically generating a descriptive utterance (usually a sentence) that describes an image content. The typical framework of neural captioning models is essentially an encoder-decoder structure. An image is first encoded into one feature vector or a set of region features via a convolutional neural network (CNN) or region-based CNN (R-CNN), and a decoder of recurrent neural network (RNN) is employed to generate a natural sentence.

In the literature, a series of innovations has been proposed to boost image captioning. One representative research direction is to leverage variants of attention mechanisms [1–4], which generally specify the spatial regions most informative for each output word. Figure 1(a) illustrates the most standard attention-based decoder.

Despite obtaining performance improvement by these techniques in terms of quantitative scores, the measure of attention in each time step is often independent, and the connections across attention are

* Corresponding author (email: jinhuitang@njust.edu.cn)

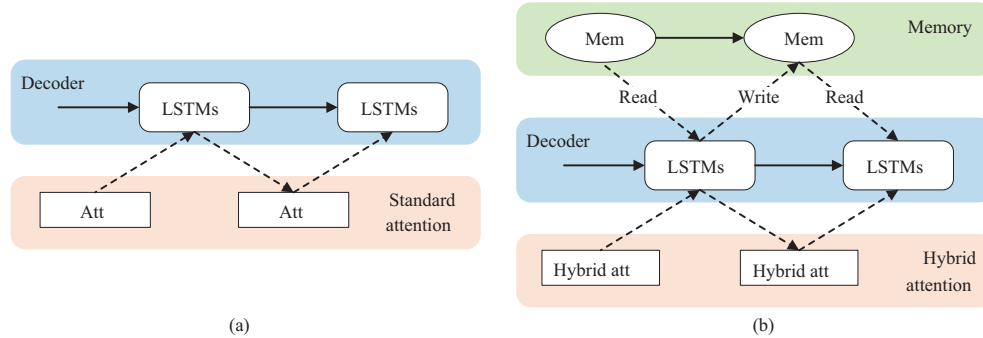


Figure 1 (Color online) Comparison between (a) a standard attention-based decoder and (b) the decoder in our CoSA-Net. (a) The standard attention-based decoder independently measures attention in each time step. (b) The decoder in our CoSA-Net memorizes the attention history through the design of memory and further develops a hybrid attention mechanism to boost image captioning.

seldom explored. As such, sequential evolution is not yet fully encoded into the attention, and the results may suffer from the problems of repeated words or incomplete sentence generation. We propose to mitigate this issue from the viewpoint of memorizing the attention history and capitalizing on such contextual knowledge to compute the next attention. Moreover, we endow the attention mechanism with more power by seeking a hybrid of easily trainable soft attention and more accurate but nondifferentiable hard attention, as shown in Figure 1(b). In our case, we perform hard attention on each of the selected inputs with the top- k highest attention values and linearly fuse the k output distributions with attention to predict the word. The attention vector is further improved with the posterior distribution on the output word at the end of each time step.

By consolidating the exploitation of contextual knowledge and the selection of a strong focus in attention modeling, we present a new contextual and selective attention network (CoSA-Net) to enhance image captioning. Specifically, faster R-CNN is first exploited to detect a set of image regions whose features are written into a static memory. Meanwhile, a dynamic memory manages the history of the sequential attention along the time steps, which the long short-term memory (LSTM) decoder reads from to measure the current attention on the static memory. At the end of each time step, the dynamic memory is updated under the guidance of the decoder state. On the basis of the attention vector, we only employ the image regions with the top- k highest attention values and couple each region individually to predict the output distribution. All k distributions are linearly averaged with the attention scores to infer the final prediction, conditioned on which we upgrade the attention by the posterior distribution. This process iterates as the sentence generation proceeds. Please also note that our CoSA-Net applies to any decoder structure, e.g., LSTM or Transformer.

The main contribution of this work is the proposal of CoSA-Net to improve the attention mechanism for captioning. The solution also leads to an elegant view of how to integrate the sequential context into attention estimation and how to devise an attention scheme of mixing soft and hard modes, which are problems not yet fully studied. Our CoSA-Net could be considered a common attention refiner and is readily pluggable into any neural captioning model. The remaining sections are organized as follows. Section 2 describes the related works. Section 3 reviews the standard soft and hard attention-based networks, and Section 4 presents our CoSA-Net. Section 5 provides the experimental results for the image captioning task, followed by the conclusion in Section 6.

2 Related work

2.1 Image captioning

Recent studies for image captioning mainly take the standard paradigm of encoder-decoder, which is usually implemented with a CNN plus RNN framework [1, 4–25]. *Show and tell* [7] is an early masterpiece that capitalizes on LSTM to conduct sequence learning conditioned on the image feature derived from CNN. Xu et al. [4] took one step further to explore the attention mechanisms, which enable the model to learn the alignments between captions and visual objects from scratch. Later on, semantic attributes are incorporated into image captioning as an additional input to the decoder [9, 15]. The captioning

performance is further boosted with the proposal of the self-critical training strategy [14]. By applying the self-critical training strategy and taking the sampling operation, the discrepancy between training and inference can be mitigated. Object level region features are exploited in the Up-Down model, in which Anderson et al. [1] detected a set of image regions with faster R-CNN and then extracted features. This work is further extended by Yao et al. [16] via modeling the relations between the objects and injecting the relations into the two-layer LSTM decoder to enhance caption generation. Ref. [26] presented a recurrent fusion network to exploit the interactions among multiple image encoders, aiming to produce more informative intermediate features for the decoder and thus enhance image captioning. Recently, by integrating the inductive bias of language generation into the encoder-decoder structure, Ref. [27] further bridged the gap between visual content and natural sentence with scene graph. Wang et al. [28] imitated the way that humans write captions with a recall mechanism in the cross-entropy optimization phase and a recalled-word reward in the CIDEr optimization phase. Considering that it could be easier to modify existing captions than to generate new ones, Ref. [29] explored fixing the details of an existing caption with a Copy-LSTM. Most recently, the captioning performance greatly benefits from the vision-language pretraining models which can provide better visual features [30–33]. This makes a new way for improving image captioning from a different angle.

2.2 Attention-based methods for image captioning

Inspired by sequence learning tasks such as machine translation, the attention mechanism, which is popular in various tasks [34–37], has brought significant improvements for image captioning [1–4, 38–44]. Given several region features of an image, the attention mechanism is to assign different importance scores to each of the regions. The decoder is thus exposed to the information which is more related to the decoding state and is able to produce more accurate descriptions. There are two main principles to incorporate the attention: soft attention and hard attention. Soft attention is the most popular attention mechanism for image captioning methods, which is firstly employed by Xu et al. [4]. In soft attention, all the region features are linearly averaged to form a context vector that will be fed into the decoder. Xu et al. [4] also explored merely sampling one region feature according to the probabilities (i.e., importance scores), which is known as the hard attention mechanism. Due to the sampling operation, hard attention is non-differentiable and REINFORCE [45] is required to learn such attention. Later on, Anderson et al. [1] incorporated the features extracted from detection models and devised a two-layer based LSTM model, where the hidden state from the first LSTM is exploited to derive the attended feature that is in turn fed into the second LSTM. Qin et al. [2] further extended [1] by leveraging the attended feature from the previous time step to compute attention scores and predicting forward to make better use of future information. By proposing a spatio-temporal memory attention mechanism, Ji et al. [20] leveraged the spatial-temporal relationship for image captioning. Moreover, Huang et al. [46] brought the idea of self-attention from the transformer [3] into the image captioning task and enhanced the attention by applying a novel attention on attention module to the encoder and decoder.

Despite leading to performance improvement, the attentions in different time steps are often considered independently and the connections across attention are seldom explored in the existing studies. In contrast, we devise a novel memory network to exploit the contextual knowledge by memorizing the attention history, conditioned on which the next attention is computed. As such, the sequential context is integrated into attention measurement. In addition, a hybrid attention mechanism is proposed to fuse soft and hard attention to enhance the attention vector, which is further upgraded with the posterior distribution on the output word.

2.3 Memory-augmented neural networks

With the capability of effectively managing sequential data, the memory network [47–49] has been a focus for recent years. In short, it stores the historical hidden states with a memory matrix, then reads and updates the memory matrix along the process. To read the memory, the attention mechanism is often adopted, according to which the memory is selectively read out. Based on [48], Sukhbaatar et al. [50] extended the memory network and trained it in an end-to-end manner. Such an architecture requires less supervision in training and thus is more general. Later on, the memory-augmented neural networks are successfully applied to several application tasks, such as neural machine translation [51, 52], textual/visual question answering [53, 54], knowledge tracking [55], sequential recommendation [56], and object tracking [57].

Inspired by the above memory-augmented neural networks, we propose to exploit the contextual knowledge by memorizing the attention history with a static-dynamic memory network in this paper, which is still a problem not yet fully studied.

3 Standard attention-based networks

Soft attention. We firstly take a brief review of the standard soft attention-based networks (SANs) [4], which is widely adopted in a series of image captioning techniques under the encoder-decoder structure. In general, given an input image I , the goal of image captioning is to generate a descriptive sentence, defined as $S = \{y_1, y_2, \dots, y_n\}$ (y_i denotes the i -th word in the sentence). For image encoder, SANs typically take CNN or R-CNN to extract a set of region features $\mathbf{v}_{1:M} = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_M\} \in \mathbb{R}^{M \times D_v}$. For a sentence decoder, LSTM is often adopted to produce words conditioned on the region features.

Formally, at each time step t , SANs learn where to look according to the decoding state $\mathbf{h}_{t-1} \in \mathbb{R}^{D_h}$ and the region features $\mathbf{v}_{1:M}$. The distribution of attention α_t over all regions can be computed as

$$\alpha_t = \text{softmax}(f_{\text{att}}(\mathbf{h}_{t-1}, \mathbf{v}_{1:M})), \quad (1)$$

where f_{att} is the function that scores how much \mathbf{h}_{t-1} attends to $\mathbf{v}_{1:M}$. The joint image representation is the weighted average of the region features: $\hat{\mathbf{v}}_t = \sum_i \alpha_{t,i} \mathbf{v}_i$, where $\alpha_{t,i}$ is the i -th element in α_t . Then, the concatenation of the attended feature $\hat{\mathbf{v}}_t$ and the embedding of the input word $\mathbf{y}_{t-1} \in \mathbb{R}^{D_y}$ is fed into LSTM to obtain \mathbf{h}_t , which is taken as the resulting state to predict the next word y_t . Hence, the output word probability for y_t is

$$P(y_t | y_{1:t-1}, \mathbf{v}_{1:M}) = P(y_t | y_{1:t-1}, \hat{\mathbf{v}}_t). \quad (2)$$

Hard attention. Instead of attending to all regions $\{v_1, v_2, \dots, v_M\}$, the hard attention mechanism [4] aligns output distribution with exactly one sampled region $v_{\tilde{m}}$. The region $v_{\tilde{m}}$ is obtained according to the attention score $\alpha_{t,\tilde{m}}$, which is sampled from the attention distribution α_t :

$$\alpha_{t,\tilde{m}} \sim \text{Multinoulli}(\alpha_t). \quad (3)$$

Because of the sampling operation, non-differentiable training is required to teach the network to choose that state and the gradient is subject to high variance. To reduce the variance, Xu et al. [4] combined REINFORCE with hard attention.

In summary, the soft attention mechanism assigns attention weights to all input states, which is end-to-end differentiable and easy to implement. In contrast, the hard attention mechanism only chooses one input state to infer the output distribution. Though a non-differentiable training strategy is needed for optimization, the hard attention mechanism is found to be more accurate than soft attention in [58].

4 CoSA-Net for image captioning

We present a new CoSA-Net to facilitate image captioning by the exploitation of contextual knowledge and the selection of strong focus in attention modeling. Figure 2 depicts the overview of the proposed CoSA-Net.

4.1 Overview

The whole architecture has three main components: an image encoder, a static-dynamic memory module (SDM) and an LSTM decoder with the hybrid attention mechanism (HAM). First of all, in image encoder, CoSA-Net extracts a set of region features $\mathbf{v}_{1:M}$ with faster R-CNN. The region features are then written into SDM as the static memory \mathbf{m}^s on one hand, and fed into LSTM decoder with the form of mean-pooled feature $\bar{\mathbf{v}} = \frac{1}{M} \sum_i \mathbf{v}_i$ on the other hand. In addition to the static memory \mathbf{m}^s , SDM also manages a dynamic memory \mathbf{m}^d which is updated along with the decoding process, aiming to memorize the attention history and thus to capitalize on such contextual knowledge to compute the next attention. Specifically, at each time step, LSTM decoder reads from the dynamic memory \mathbf{m}^d to measure the current attention vector α_t over the static memory \mathbf{m}^s . The dynamic memory \mathbf{m}^d is then updated under the guidance of the decoder state \mathbf{h}_t at the end of each time step.

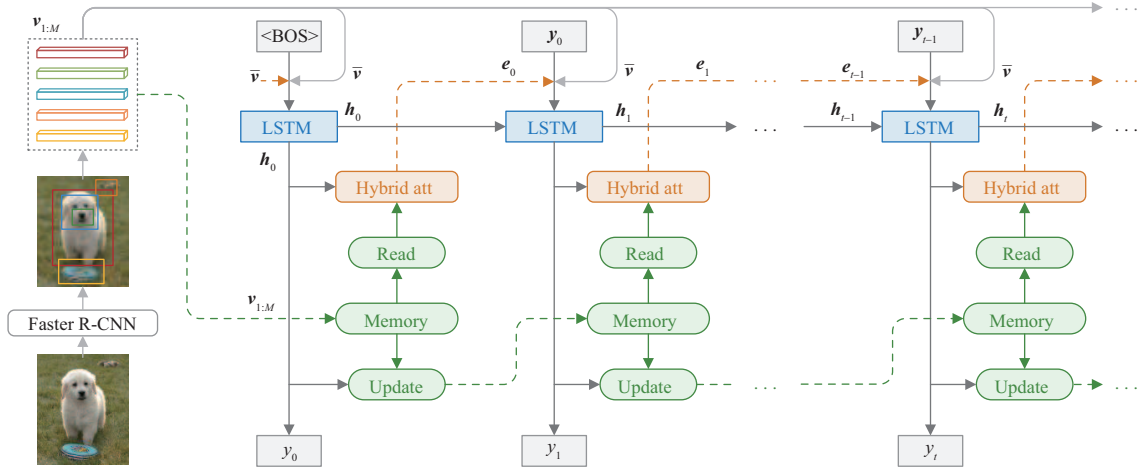


Figure 2 (Color online) Overview of CoSA-Net for image captioning. Faster R-CNN is first exploited to detect a set of image regions. The region features are then written into a static-dynamic memory module (SDM) as the static memory and fed into the LSTM decoder in the form of a mean-pooled feature \bar{v} . The LSTM decoder then reads from the dynamic memory to compute the current attention vector and further update the dynamic memory under the guidance of the decoder state. To predict the output distribution, the decoder adopts the hybrid attention mechanism (HAM) which is a hybrid of soft attention and hard attention. Finally, the attention vector is upgraded by the posterior distribution conditioned on the output word to refine the attended feature in the next time step.

Taking inspiration from easily trainable soft attention and the more accurate but non-differentiable hard attention, we design the hybrid attention mechanism to further strengthen the attention modeling with a hybrid of soft and hard attention. Such design not only seeks more accurate attention by coupling input states individually to the output as in hard attention, but also benefits from the end-to-end differentiability of soft attention. Note that different from the primary hard attention which often suffers from the high variance of Monte-Carlo sampling gradients, our HAM approximates the hard attention in an easily trainable way by aggregating the output distributions of top- k input states with the highest attention weights. In particular, we select the regions with top- k highest attention scores in α_t to model hard attention. Each selected region is individually leveraged to compute an output distribution. All the k output distributions are then linearly fused with original attention weights from α_t to predict the next word y_t . At the end of each time step, the attention vector α_t is further upgraded to the posterior attention β_t by the posterior distribution conditioned on the output word y_t to refine the attended feature at the next time step.

Next, we introduce the two core modules in our CoSA-Net, i.e., SDM and HAM, in detail. Recall that our CoSA-Net is applicable to any decoder structure; here we first present the two modules in the context of a basic decoder with one-layer LSTM for simplicity.

4.2 Static-dynamic memory

SDM contains two kinds of memories: a static memory and a dynamic memory. The fixed static memory \mathbf{m}^s stores the region features $\mathbf{v}_{1:M}$, whereas the dynamic one \mathbf{m}^d manages the history of the sequential attention along decoding time steps. Both of them are initialized by $\mathbf{v}_{1:M}$. The m -th memory slot in $\mathbf{m}^s/\mathbf{m}^d$ corresponds to the m -th region.

Read from SDM. At each time step, LSTM decoder reads the two memories in SDM to measure attention by exploiting the attention history. Details of attention measurement from memory can be referred to Subsection 4.3.

Update/Write into SDM. At the end of each time step, we update the dynamic memory and write the memory back to SDM under the guidance of the current decoder state \mathbf{h}_t . Figure 3 details the update. As such, the history of the sequential attention is memorized in time. Inspired by the gate units in memory networks [47, 51, 59], we adopt a forget gate $\mathbf{g}_f \in \mathbb{R}^{D_h}$ and an add gate $\mathbf{g}_a \in \mathbb{R}^{D_h}$ to determine which parts of the previous memory \mathbf{m}_{t-1}^d should be forgotten and what information from the current state \mathbf{h}_t should be added, respectively. Specifically, \mathbf{g}_f and \mathbf{g}_a are calculated as

$$\mathbf{g}_f = \text{sigmoid}(\mathbf{W}_u^f \mathbf{h}_t) \quad \text{and} \quad \mathbf{g}_a = \text{sigmoid}(\mathbf{W}_u^a \mathbf{h}_t), \quad (4)$$

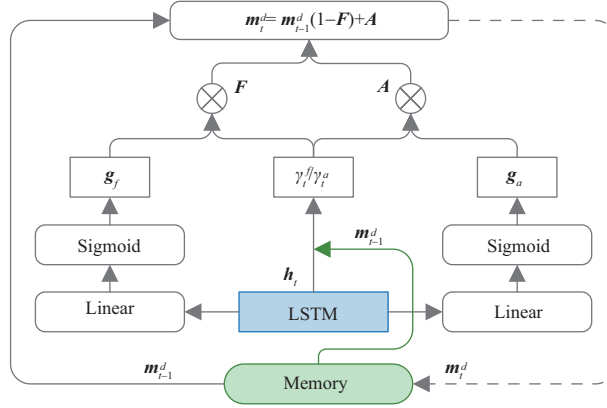


Figure 3 (Color online) The update of the dynamic memory.

where $\mathbf{W}_u^f, \mathbf{W}_u^a$ are transformation matrices and $\mathbf{W}_u^f \in \mathbb{R}^{D_h \times D_h}, \mathbf{W}_u^a \in \mathbb{R}^{D_h \times D_h}$. Next, we measure the similarity between each memory slot and current decoder state \mathbf{h}_t , which leads to normalized similarity vectors $\gamma_t^f, \gamma_t^a \in \mathbb{R}^M$ over all memory slots as

$$\begin{aligned} r_{t,i}^f &= \mathbf{w}_1 [\tanh(\mathbf{W}_1^h \mathbf{h}_t + \mathbf{W}_1^m \mathbf{m}_{t-1,i}^d)], & \gamma_t^f &= \text{softmax}(\mathbf{r}_t^f), \\ r_{t,i}^a &= \mathbf{w}_2 [\tanh(\mathbf{W}_2^h \mathbf{h}_t + \mathbf{W}_2^m \mathbf{m}_{t-1,i}^d)], & \gamma_t^a &= \text{softmax}(\mathbf{r}_t^a), \end{aligned} \quad (5)$$

where $r_{t,i}^f$ is the i -th element of \mathbf{r}_t^f and $r_{t,i}^a$ is the i -th element of \mathbf{r}_t^a , $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^{1 \times D_u}$, $\mathbf{W}_1^h, \mathbf{W}_2^h \in \mathbb{R}^{D_u \times D_h}$, $\mathbf{W}_1^m, \mathbf{W}_2^m \in \mathbb{R}^{D_u \times D_m}$. The forgetting content $\mathbf{F} \in \mathbb{R}^{M \times D_h}$ and the adding information $\mathbf{A} \in \mathbb{R}^{M \times D_h}$ are then obtained by applying the two gates to the normalized similarity vectors γ_t^f, γ_t^a :

$$\mathbf{F} = \gamma_t^f \mathbf{g}_f^T \quad \text{and} \quad \mathbf{A} = \gamma_t^a \mathbf{g}_a^T. \quad (6)$$

The dynamic memory is finally updated as

$$\mathbf{m}_t^d = \mathbf{m}_{t-1}^d \odot (1 - \mathbf{F}) + \mathbf{A}, \quad (7)$$

where \odot denotes element-wise multiplication.

4.3 Decoder with the hybrid attention mechanism

To facilitate attention estimation in decoder, we devise an HAM to additionally emphasize the focus on the regions with top- k highest attention scores by mixing soft and hard attention. Such design not only seeks more accurate attention through the hard attention over the selected regions with the highest attention scores, but also makes HAM end-to-end differentiable and easy to implement. After that, the learned attention vector is further enhanced with the posterior distribution.

Figure 4 illustrates the pipeline of our hybrid attention mechanism. Such a mechanism firstly employs soft attention to calculate an attention vector α_t that measures the importance of each slot in the static memory \mathbf{m}^s , which is then utilized to produce context output distributions P^s and P^h over vocabulary at soft and hard attention mode, respectively. The two derived distributions are subsequently fused with the original output distribution P conditioned on the hidden state \mathbf{h}_t to predict the next word y_t . Finally, α_t is refined to the posterior attention β_t depending on the context posterior distribution of y_t from the hybrid attention mechanism at hard attention mode.

Formally, at time step t , LSTM outputs state \mathbf{h}_t as

$$\mathbf{h}_t = \text{LSTM}(\mathbf{h}_{t-1}, \mathbf{x}_t) \quad \text{and} \quad \mathbf{x}_t = [\mathbf{e}_{t-1}, \bar{\mathbf{v}}, \mathbf{y}_{t-1}], \quad (8)$$

where the input contextual information \mathbf{x}_t is obtained by concatenating the input word \mathbf{y}_{t-1} , the mean-pooled image feature $\bar{\mathbf{v}}$, and the attended feature $\mathbf{e}_{t-1} = \sum_i \beta_{t-1,i} \mathbf{m}_i^s$ with the posterior attention β_{t-1} from the previous time step.

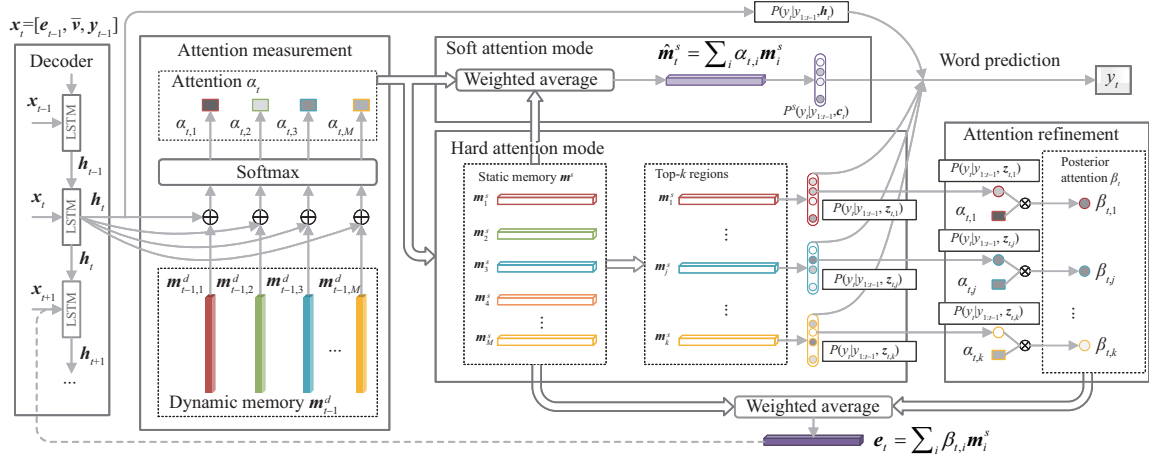


Figure 4 (Color online) Pipeline of the hybrid attention mechanism. At time step t , depending on hidden state \mathbf{h}_t , the LSTM decoder reads from the dynamic memory \mathbf{m}_{t-1}^d in the SDM and computes the attention vector α_t . Next, soft attention and hard attention are performed conditioned on α_t . For the soft attention mode, we first obtain the attended image feature, based on which we produce the soft context information and further obtain the soft context output distribution P^s . For the hard attention mode, we select the image regions with the top- k highest attention values in α_t . After that, we collect the hard context information for each selected region and predict k distributions over the vocabulary. Depending on these output distributions, the final prediction of the next word y_t is performed. Finally, we leverage the posterior distribution of y_t to enhance the attention vector α_t as the posterior attention β_t . The posterior attention is then applied to all regional features to obtain the attended image feature \mathbf{e}_t , which guides the decoding phase in the next time step.

Hybrid attention. Depending on the current hidden state \mathbf{h}_t , the LSTM decoder reads from the dynamic memory \mathbf{m}_{t-1}^d in the SDM, and learns how much to attend to the static memory \mathbf{m}^s . The attention value $\alpha_{t,i}$ for the i -th region in the vector α_t is thus computed as

$$\alpha_{t,i} = \frac{\exp(a_{t,i})}{\sum_{j=1}^M \exp(a_{t,j})}, \quad a_{t,i} = \mathbf{w}_a^T \tanh(\mathbf{W}_a^h \mathbf{h}_t + \mathbf{W}_a^m \mathbf{m}_{t-1}^d), \quad (9)$$

where \mathbf{W}_a^h , \mathbf{W}_a^m are transformation matrices and $\mathbf{W}_a^h \in \mathbb{R}^{D_a \times D_h}$, $\mathbf{W}_a^m \in \mathbb{R}^{D_a \times D_m}$, $\mathbf{w}_a \in \mathbb{R}^{D_a}$.

Conditioned on the attention vector α_t , both soft attention and hard attention are performed, which we refer to as soft and hard attention mode in the hybrid attention mechanism, respectively. In the soft attention mode, we firstly obtain the attended image feature $\hat{\mathbf{m}}_t^s = \sum_i \alpha_{t,i} \mathbf{m}_i^s$ given the attention vector α_t derived from soft attention overall image regions as mentioned above. Next, we treat the concatenation of the attended image feature $\hat{\mathbf{m}}_t^s$ and the LSTM hidden state \mathbf{h}_t as the soft context information, which is formulated as

$$\mathbf{c}_t = \sigma([\hat{\mathbf{m}}_t^s, \mathbf{h}_t]), \quad (10)$$

where σ is a gated linear unit (GLU) [60]. The soft context information \mathbf{c}_t is further leveraged to produce the soft context output distribution P^s :

$$P^s(y_t|y_{1:t-1}, \mathbf{c}_t) = \text{softmax}(f(\mathbf{c}_t)), \quad (11)$$

where $f(\cdot)$ is a linear layer. In hard attention mode, we select regions with top- k highest attention values in α_t . The hard attention mechanism is then performed over the selected k regions. More specifically, we collect hard context information $\mathbf{z}_{t,j}$ for each selected region by concatenating region feature \mathbf{m}_j^s with hidden state \mathbf{h}_t :

$$\mathbf{z}_{t,j} = \sigma([\mathbf{m}_j^s, \mathbf{h}_t]). \quad (12)$$

As such, the hard context information $\mathbf{z}_{t,j}$ for each selected region is further leveraged to predict the distribution over vocabulary individually:

$$P(y_t|y_{1:t-1}, \mathbf{z}_{t,j}) = \text{softmax}(f(\mathbf{z}_{t,j})). \quad (13)$$

All k distributions of the selected regions are then aggregated with the normalized attention weights from α_t , and the hard context output distribution P^h is computed by

$$P^h(y_t|y_{1:t-1}, \mathbf{Z}_t) = \sum_{j=1}^k \alpha'_{t,j} P(y_t|y_{1:t-1}, \mathbf{z}_{t,j}), \quad (14)$$

where $\alpha'_{t,j} = \frac{\alpha_{t,j}}{\sum_r \alpha_{t,r}}$.

Word prediction. The final prediction of the next word y_t is performed depending on the combination of the soft context output distribution $P^s(y_t|y_{1:t-1}, \mathbf{c}_t)$, the hard context output distribution $P^h(y_t|y_{1:t-1}, \mathbf{Z}_t)$, and the distribution $P(y_t|y_{1:t-1}, \mathbf{h}_t)$ conditioned on \mathbf{h}_t :

$$P(y_t|y_{1:t-1}, \mathbf{h}_t, \mathbf{c}_t, \mathbf{Z}_t) = [P^s(y_t|y_{1:t-1}, \mathbf{c}_t) + P^h(y_t|y_{1:t-1}, \mathbf{Z}_t) + P(y_t|y_{1:t-1}, \mathbf{h}_t)]/3. \quad (15)$$

Attention refinement. Taking the inspiration from posterior attention models [61], we leverage the posterior distribution of the predicted word y_t to enhance the attention vector α_t as the posterior attention β_t . The spirit behind follows the philosophy that the posterior attention, which is obtained conditioned on the current output, is more closely associated with the real output and thus is more accurate. Specifically, the posterior attention β_t is computed as

$$\beta_{t,i} = \frac{p(y_t|y_{1:t-1}, \mathbf{z}_{t,i})\alpha_{t,i}}{\sum_s p(y_t|y_{1:t-1}, \mathbf{z}_{t,s})\alpha_{t,s}}, \quad (16)$$

where $p(y_t|y_{1:t-1}, \mathbf{z}_{t,s})$ denotes the predicted probability for y_t from the distribution $P(y_t|y_{1:t-1}, \mathbf{z}_{t,s})$ at hard attention mode. For the remaining $M - k$ regions, $\beta_{t,i}$ is set to be zero as a mask. In this way, the top- k attention values are refined by mixing the soft and hard attention. The posterior attention is then applied to all the region features to obtain the attended image feature $\mathbf{e}_t = \sum_i \beta_{t,i} \mathbf{m}_i^s$. The attended feature will be fed into the LSTM as part of the input contextual information at the next time step ($\mathbf{x}_{t+1} = [\mathbf{e}_t, \bar{\mathbf{v}}, \mathbf{y}_t]$, similar to (8)) and guide the decoding phase again. It is notable that when the current predicted word is not directly correlated to a visual feature, the predicted probabilities $p(y_t|y_{1:t-1}, \mathbf{z}_{t,s})$ of the word are observed to be extremely small. To alleviate the negative impact that may arise, we adopt a threshold ρ to determine whether to conduct refinement. If the summation of the k predicted probabilities is less than ρ , the attention will remain unchanged.

4.4 Up-Down with CoSA-Net

Since we design our CoSA-Net architecture to be a common attention refiner that strengthens attention modeling via the exploitation of contextual knowledge and the selection of strong focus regions, it is feasible to plug CoSA-Net into any decoder structure. We next present how to integrate our CoSA-Net into Up-Down [1] with respect to two-layer LSTM, as shown in Figure 5. Specifically, at time step t , the input of the first LSTM is

$$\mathbf{x}_t^1 = [\mathbf{h}_{t-1}^2, \mathbf{e}_{t-1}, \bar{\mathbf{v}}, \mathbf{y}_{t-1}], \quad (17)$$

where \mathbf{h}_{t-1}^2 is the previous output hidden state of the second LSTM. \mathbf{x}_t^1 is fed into the first LSTM to obtain the state \mathbf{h}_t^1 . After that, we compute the attention distribution α_t with the newly obtained \mathbf{h}_t^1 and the dynamic memory \mathbf{m}_{t-1}^d from the previous time. The attended feature $\hat{\mathbf{m}}_t^s = \sum_i \alpha_{t,i} \mathbf{m}_i^s$ and \mathbf{h}_t^1 are concatenated as the input of the second LSTM. The output hidden state \mathbf{h}_t^2 is thus given by

$$\mathbf{h}_t^2 = \text{LSTM}(\mathbf{h}_{t-1}^2, [\mathbf{h}_t^1, \hat{\mathbf{m}}_t^s]). \quad (18)$$

To predict the current word y_t , we still combine the soft context information \mathbf{c}_t and the hard context information $\mathbf{z}_{t,j}$ for the regions having the top- k attention values as

$$\mathbf{c}_t = \sigma([\hat{\mathbf{m}}_t^s, \mathbf{h}_t^2]) \quad \text{and} \quad \mathbf{z}_{t,j} = \sigma([\mathbf{m}_j^s, \mathbf{h}_t^2]). \quad (19)$$

The same combination strategy is adopted when computing the distribution over possible words as in Subsection 4.3. Given the current word y_t , the attention vector is refined with the posterior distribution to get a more accurate attention distribution β_t , which has a more direct correlation with the current output. Similarly, the refined attended feature \mathbf{e}_t is fed into the first LSTM at the next time step to effectively incorporate β_t into the following decoding phase.

In the meanwhile, we update dynamic memory conditioned on the hidden state from the second LSTM. Different from Subsection 4.2, we compute the two gates \mathbf{g}_f , \mathbf{g}_a and the weights γ_t^f, γ_t^a according to \mathbf{h}_t^2 :

$$\mathbf{g}_f = \text{sigmoid}(\mathbf{W}_u^f \mathbf{h}_t^2), \quad \mathbf{g}_a = \text{sigmoid}(\mathbf{W}_u^a \mathbf{h}_t^2), \quad (20)$$

$$\mathbf{r}_{t,i}^f = \mathbf{w}_1 [\tanh(\mathbf{W}_1^h \mathbf{h}_t^2 + \mathbf{W}_1^m \mathbf{m}_{t-1,i}^d)], \quad \gamma_t^f = \text{softmax}(\mathbf{r}_t^f), \quad (21)$$

$$\mathbf{r}_{t,i}^a = \mathbf{w}_2 [\tanh(\mathbf{W}_2^h \mathbf{h}_t^2 + \mathbf{W}_2^m \mathbf{m}_{t-1,i}^d)], \quad \gamma_t^a = \text{softmax}(\mathbf{r}_t^a), \quad (22)$$

where \mathbf{g}_f , \mathbf{g}_a and γ_t^f, γ_t^a are then leveraged to compute the forgetting content \mathbf{F} , the adding information \mathbf{A} , and finally the updated memory \mathbf{m}_t^d , which is then delivered to the next time step.

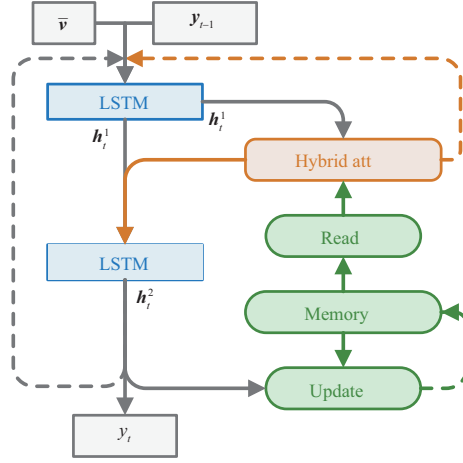


Figure 5 (Color online) Illustration of how to integrate our CoSA-Net into an Up-Down structure with respect to a two-layer LSTM.

5 Experiments

5.1 Datasets and settings

Dataset and evaluation metrics. We conduct experiments on the widely-adopted MS COCO dataset [62]. The dataset contains 123287 images in total, in which 82783 images are used for training and the rest are used for validation. Each image is annotated with at least 5 captions. In our experiments, we follow [1] and take the “Karpathy” split (5000 for validation, 5000 for testing and the rest for training) for a fair comparison. Following [63], we convert all the sentences into lowercase and establish a vocabulary with the words occurring more than 4 times. Standard evaluation metrics including METEOR [64], CIDEr-D [65], BLEU@N [66], ROUGE-L [67] and SPICE [68] are measured by using officially released codes¹⁾.

Compared methods. (1) LSTM [7] feeds image features extracted from the encoder into an LSTM decoder at the first time step and produces sentences. (2) SCST [14] proposes a self-critical training strategy which directly optimizes the captioning model with CIDEr-D score. (3) LSTM-A [15] explores semantic attributes and takes them as the additional input to the decoder. (4) Up-Down [1] extends region features to the object level and devises a two-layer based decoder. (5) RFNet [26] fuses multiple encoders with a novel recurrent fusion network to enhance the encoder features. (6) GCN-LSTM [16] upgrades the object-level features in [1] by exploring relationships between objects. (7) SGAE [27] leverages the scene graph representation and a shared dictionary to guide the decoding phase. (8) LBPf [2] extends [1] by utilizing future information for the current prediction. (9) Base₁-LSTM and Base₂-LSTM are the re-implementations of [1, 14] based on our experimental setting. Base₁-LSTM + CoSA-Net and Base₂-LSTM + CoSA-Net are our proposals by plugging the designed CoSA-Net into [1, 14].

Implementation details. The region features utilized in this paper are extracted from the pre-trained object detector on the visual genome as in [1]. Each image has 10–100 regions, which are represented as 2048-dimensional vectors, respectively. The vectors are embedded to the dimension of 1000, which is equal to the dimension of hidden states in LSTM and the embedding size of input words. The number k of the selected regions is set as 12 and we set the threshold ρ for attention refinement as 0.1. To alleviate the sparsity induced by the top- k operation, the posterior attention is augmented with the values from standard soft attention and the two distributions are combined with the same weight in our experiments as in [61]. We implement the proposed CoSA-Net based on PyTorch, with Adam [69] as the optimizer. The learning rate for training CoSA-Net under cross-entropy loss is initialized as 2×10^{-4} , with a mini-batch size of 10. After being trained with cross-entropy loss for 30 epochs, we select the model which achieves the best CIDEr-D score on a validation set as the initial model for self-critical learning. The learning rate is set as 2×10^{-5} and the model is optimized with the CIDEr-D score for another 30 epochs. In the inference stage, the beam search strategy is adopted and the beam size is set to 2.

1) <https://github.com/tylin/coco-caption>.

Table 1 Performance (%) of our CoSA-Net and other state-of-the-art methods on MS-COCO Karpathy test split, where C, M, S, B@4, and R are short for CIDEr-D, METEOR, SPICE, BLEU@4, and ROUGE-L scores

	Cross-entropy loss					CIDEr-D score optimization				
	C	M	S	B@4	R	C	M	S	B@4	R
LSTM [7]	94.0	25.2	–	29.6	52.6	106.3	25.5	–	31.9	54.3
SCST [14]	99.4	25.9	–	30.0	53.4	114.0	26.7	–	34.2	55.7
LSTM-A [15]	108.8	26.9	20.0	35.2	55.8	118.3	27.3	20.8	35.5	56.8
Up-Down [1]	113.5	27.0	20.3	36.2	56.4	120.1	27.7	21.4	36.3	56.9
RFNet [26]	116.3	27.9	20.8	37.0	57.3	125.7	28.3	21.7	37.9	58.3
GCN-LSTM _{spa} [16]	115.6	27.8	20.8	36.5	56.8	127.0	28.4	21.9	37.8	58.1
GCN-LSTM _{sem} [16]	116.3	27.9	20.9	36.8	57.0	127.6	28.5	22.0	38.2	58.3
SGAE [27]	–	–	–	–	–	127.7	28.4	22.1	38.4	58.6
LBPF [2]	116.4	28.1	21.2	37.4	57.5	127.6	28.5	22.0	38.3	58.4
Base ₁ -LSTM	114.0	27.6	20.8	36.3	56.8	125.2	28.3	21.8	37.6	58.1
Base ₁ -LSTM + CoSA-Net	116.9	28.1	21.2	36.7	57.2	128.5	28.8	22.4	38.5	58.6
Base ₂ -LSTM	115.1	28.0	21.1	36.5	57.1	127.6	28.5	22.0	38.5	58.5
Base ₂ -LSTM + CoSA-Net	117.3	28.3	21.3	37.1	57.5	129.5	29.0	22.5	39.0	58.7

Table 2 Performance (%) of our CoSA-Net and other state-of-the-art methods with model ensembles, where C, M, S, B@4, and R are short for CIDEr-D, METEOR, SPICE, BLEU@4, and ROUGE-L scores

	Cross-entropy loss					CIDEr-D score optimization				
	C	M	S	B@4	R	C	M	S	B@4	R
SCST [14]	106.5	26.7	–	32.8	55.1	117.5	27.1	–	35.4	56.6
RFNet [26]	116.3	27.9	20.8	37.0	57.3	125.7	28.3	21.7	37.9	58.3
GCN-LSTM [16]	117.1	28.1	21.1	37.1	57.2	128.7	28.6	22.1	38.3	58.5
SGAE [27]	–	–	–	–	–	129.1	28.4	22.2	39.0	58.9
Base ₁ -LSTM + CoSA-Net	118.3	28.4	21.4	37.4	57.6	130.7	29.0	22.6	39.6	59.1
Base ₂ -LSTM + CoSA-Net	119.0	28.4	21.5	37.3	57.6	131.0	29.1	22.7	39.7	59.2

5.2 Quantitative analysis

We compare with several state-of-the-art methods and summarize the performances in Table 1. Overall, CoSA-Net exhibits better performances than the non-attention approaches (LSTM and LSTM-A) and the attention-based models (SCST, Up-Down, RFNet, GCN-LSTM, SGAE, and LBPF) in terms of CIDEr-D, SPICE, and METEOR. With the optimization on cross-entropy loss, Base₁-LSTM + CoSA-Net and Base₂-LSTM + CoSA-Net lead to the absolute improvement over Base₁-LSTM and Base₂-LSTM by 2.9% and 2.2%, respectively, in CIDEr-D. The results basically indicate the advantage of exploiting the dependency among attention history and emphasizing the strong focus in each attention by seeking a hybrid of soft and hard attention. Furthermore, LSTM-A, which injects semantic attributes into decoders, outperforms LSTM with a large margin. Nevertheless, the attention-based models (SCST, Up-Down, RFNet, GCN-LSTM, SGAE, and LBPF) still yield better performances than LSTM-A. That verifies the impact of the attention mechanism. By further exploring the dependency among attention history and leveraging the hybrid attention mechanism, Base₁-LSTM + CoSA-Net and Base₂-LSTM + CoSA-Net are superior to SCST and Up-Down. In addition, when being optimized with CIDEr-D, the CIDEr-D score of Base₂-LSTM + CoSA-Net is boosted up to 129.5%. As expected, the results indicate that employing a self-critical training strategy can effectively alleviate the gap between training and inference. Similar to the observations on the optimization with cross-entropy loss, Base₁-LSTM + CoSA-Net and Base₂-LSTM + CoSA-Net outperform Base₁-LSTM and Base₂-LSTM by 3.3% and 1.9% in CIDEr-D, respectively, when being optimized with CIDEr-D.

We further conduct evaluations by ensembling multiple models with different parameter initializations in our CoSA-Net. Table 2 details the performance comparison between CoSA-Net and other state-of-the-art methods with model ensembles. As shown in Table 2, the ensembled CoSA-Net exhibits better performance against the other attention-based models (SCST, RFNet, GCN-LSTM, and SGAE) across all the metrics, which demonstrates the effectiveness of the memory module and the hybrid attention in our CoSA-Net.

Table 3 Leaderboard of state-of-the-art methods on the online MS-COCO test server, where B@N, M, R, and C are short for BLEU@N, METEOR, ROUGE-L, and CIDEr-D scores

	B@1		B@2		B@3		B@4		M		R		C	
	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
SCST [14]	78.1	93.7	61.9	86.0	47.0	75.9	35.2	64.5	27.0	35.5	56.3	70.7	114.7	116.7
LSTM-A [15]	78.7	93.7	62.7	86.7	47.6	76.5	35.6	65.2	27.0	35.4	56.4	70.5	116.0	118.0
Up-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
RFNet [26]	80.4	95.0	64.9	89.3	50.1	80.1	38.0	69.2	28.2	37.2	58.2	73.1	122.9	125.1
SGAE [27]	81.0	95.3	65.6	89.5	50.7	80.4	38.5	69.7	28.2	37.2	58.6	73.6	123.8	126.5
GCN-LSTM [16]	80.8	95.2	65.5	89.3	50.8	80.3	38.7	69.7	28.5	37.6	58.5	73.4	125.3	126.5
Base ₁ -LSTM + CoSA-Net	81.0	95.4	65.6	89.8	50.9	81.1	38.9	70.7	28.8	38.1	58.8	74.1	126.0	128.4
Base ₂ -LSTM + CoSA-Net	81.0	95.4	65.7	89.9	51.0	81.2	39.0	70.9	28.8	38.3	58.8	74.1	126.2	128.5

Table 4 Performance contribution of each component in CoSA

Method	HAM	AttRef	SDM	C
Base ₁ -LSTM				114.0
+ HAM	✓			115.9
+ HAM, AttRef	✓	✓		116.4
+ SDM			✓	115.3
Base ₁ -LSTM + CoSA-Net	✓	✓	✓	116.9

5.3 Online evaluation

To fully verify the effectiveness of the proposed method, we additionally evaluate our CoSA-Net on the online MS-COCO test server. Table 3 summarizes the performance leaderboard on the official testing set with 5 reference captions (c5) and 40 reference captions (c40). Compared to the published top-performing methods on the leaderboard, our CoSA-Net shows better performance across all evaluation metrics. The results again demonstrate the advantage of memorizing the contextual attention and focusing on the principal components from each attention for image captioning.

5.4 Ablation study

Effect of the individual component. In order to examine how each component in CoSA influences the overall performance, we conduct an ablation study in Table 4 by successively taking HAM, attention refinement (AttRef) and SDM into Base₁-LSTM. Compared to Base₁-LSTM which is only equipped with soft attention, the proposed HAM increases the performances by 1.9% in CIDEr-D. HAM is benefited from the subtle mix of soft and hard attention, and empowers Base₁-LSTM with the capability of globally attending to all the regions in the image and locally emphasizing the top- k most important regions simultaneously. Moreover, the use of AttRef boosts the CIDEr-D score from 115.9% to 116.4%. Such results indicate that the posterior distributions of the output word derived from HAM can effectively associate the attention vector with the real output, and thus produce a more accurate visual context to guide the decoder at the next time step. To verify the effectiveness of SDM, we also experiment by directly integrating Base₁-LSTM with SDM to strengthen the soft attention with attention history. The CIDEr-D score of Base₁-LSTM + SDM is 115.3%, which is higher than 114.0% of Base₁-LSTM. This validates the impact of memorizing and dynamically updating the sequential context of attention for enhancing caption generation. In addition, adopting HAM and SDM together finally improves the performance to 116.9% in CIDEr-D score.

Effect of the number of selected regions in HAM. Next, we investigate the effect of the number of selected regions with the highest attention values in HAM. Figure 6(a) details the results of Base₁-LSTM + CoSA-Net with different number k varying from 6 to 24. The best performance is achieved when k is 12. In particular, once the number of selected regions is larger than 12, the performance slightly decreases. We speculate that this may be the result of involving more invalid regions and that double proves the motivation of focusing on the top- k regions with the highest attention scores in HAM.

Effect of the threshold in attention refinement. Figure 6(b) details the effect of the threshold ρ in attention refinement. As shown in Figure 6(b), the highest CIDEr-D score is attained when ρ is 0.1. In the case that ρ is smaller than 0.1, the posterior attention is applied to refine the attended image feature

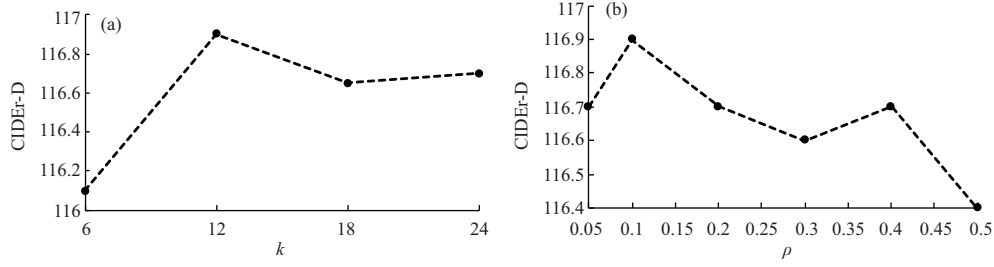


Figure 6 Effect of the number of selected regions with the highest attention values in the Hybrid Attention Mechanism (a) and the threshold in Attention Refinement (b).

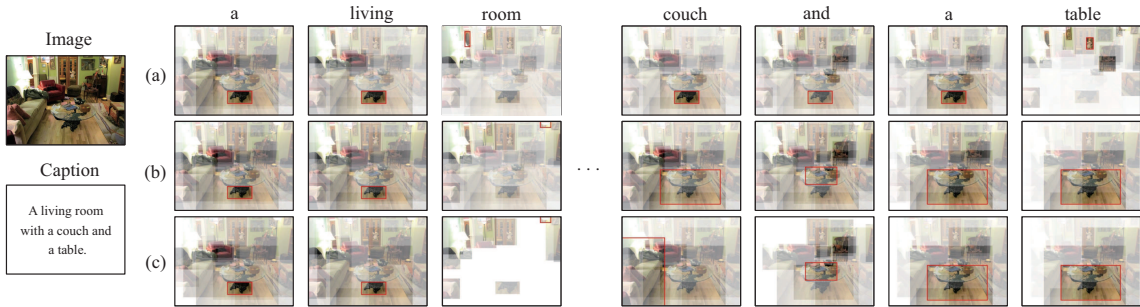


Figure 7 (Color online) The visualization of attention with the respect to time step for Base₁-LSTM (a) and Base₁-LSTM + CoSA-Net ((b) for attention α_t , (c) for the refined β_t). The image region with maximum attention weight in each decoding step is highlighted in red.

at each decoding step, even when the current predicted word is not directly correlated to visual feature (i.e., the summation of the attention values of the top- k regions is small). This inevitably results in the inferior performances. Moreover, when ρ is larger than 0.1, the performance is slightly dropped, since the additional process of attention refinement is easily omitted and the top- k attention values remain unchanged. The results basically verify the merit of exploiting the posterior distribution of the output word for attention refinement.

5.5 The visualization of attention

To demonstrate how attention transits among object regions during the caption generation, we visualize the attention learned by Base₁-LSTM, the attention α_t , and the posterior attention β_t in Base₁-LSTM + CoSA-Net in each time step from row (a) to row (c) in Figure 7, respectively. The transparency of each region box indicates the corresponding strength of focus, and the red bounding box represents the region with the highest attention score. Obviously, Base₁-LSTM fails to focus on either “couch” or “table” in row (a) by merely adopting soft attention while “table” is successfully attended to in row (b) by incorporating the sequential context into soft attention through the SDM in Base₁-LSTM + CoSA-Net. Moreover, after applying AttRef, “couch” and “table” are pinpointed by Base₁-LSTM + CoSA-Net in row (c), which indicates that upgrading the attention vector α_t to β_t conditioned on the posterior distribution of the output word can effectively enhance the alignments between visual objects and captions. Such results again verify the superiority of CoSA-Net.

5.6 Qualitative analysis

Figure 8 showcases several image examples with captions produced by Base₁-LSTM, Base₁-LSTM + CoSA-Net, and ground truth annotations. As illustrated in the exemplar results, the sentences from Base₁-LSTM + CoSA-Net are more coherent and accurate. For example, in the first row, compared to Base₁-LSTM, which ignores the “table” in the image, the sentence by Base₁-LSTM + CoSA-Net mentions “table” and depicts the image content more accurately by capitalizing on the attention history and attending to more regions. This result proves the reasonableness of exploiting the dependency among attention histories and the subtle mix of soft and hard attention to boost image captioning.

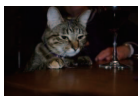







Image	Captions	GTs
	Base _{1-LSTM} : A cat sitting on top of a glass of wine. + CoSA-Net: A cat sitting on a table next to a glass of wine.	A cat leaning on top of a wooden table. A cat with its paws on a table near a glass of wine. A cat with its front paws on the table.
	Base _{1-LSTM} : A woman sitting at a table with a table. + CoSA-Net: A woman sitting at a table in a restaurant.	A woman with beautiful breast sitting at a table. A person sitting at a table with food and a drink. Young women gazing pensively left while sitting in a restaurant.
	Base _{1-LSTM} : A group of elephants standing next to each other. + CoSA-Net: A group of elephants standing next to a fence.	Two elephants standing behind a rope in an enclosure. A couple of elephants standing next to each other. Two large elephants waiting to enter their shelter.
	Base _{1-LSTM} : A group of zebras grazing in a field of grass. + CoSA-Net: A group of zebras grazing in the grass near a body of water.	Several zebras eat the green grass in the pasture. Two zebras and another animal grazing in the grass. Three zebra in the middle of a field with a body of water in the distance.
	Base _{1-LSTM} : A man and a woman sitting on a table with a banana. + CoSA-Net: A young boy wearing a bunch of bananas on his head.	A person wearing a banana headdress and necklace. A lady dressed in a blue and purple outfit wearing a hat made of fruit. A person wearing a hat made out of yellow bananas.
	Base _{1-LSTM} : A cat laying on top of a book. + CoSA-Net: A cat sleeping on top of a book.	A cat is laying next to a blue book. A cat is sleeping in front of a book. A cat sleeping next to a book on the floor.
	Base _{1-LSTM} : A man sitting on a chair in a living room. + CoSA-Net: A man sitting on a chair holding a wii game controller.	A man in a chair holding a Wii remote. A person that is playing a video game. A bearded man is sitting in a chair with a controller.
	Base _{1-LSTM} : A group of boats are sitting in the water. + CoSA-Net: A group of boats are sitting on the beach.	Several beached boats on the sand with orange balls hanging over the sides. A large number of boats that have been beached. The ships are all docked on the beach by the water.

Figure 8 (Color online) Examples for captions generated by Base_{1-LSTM}, Base_{1-LSTM} + CoSA-Net and ground truths (GTs).

6 Conclusion

We have presented CoSA-Net, which explores the sequential evolution of sentence generation via attention mechanisms. Particularly, we have studied the problem from the viewpoint of leveraging the attention history in context. To realize our idea, we have devised a static-dynamic memory module in which the image region features are written into a static memory, and a dynamic memory manages the attention history to compute the attention for the decoder in each time step. Moreover, a hybrid attention mechanism is presented by exploiting soft and hard attention, followed by an update on attention with the posterior distribution. Extensive experiments performed on the COCO image captioning dataset demonstrate the efficacy of CoSA-Net when integrating CoSA-Net into a one-layer and a two-layer LSTM decoder. More remarkably, CoSA-Net formulates attention in a new paradigm and shapes an encouraging attention structure for any decoder.

Acknowledgements This work was supported by National Key Research and Development Program of China (Grant No. 2018AAA0102002) and National Natural Science Foundation of China (Grant No. 61732007).

References

- Anderson P, He X, Buehler C, et al. Bottom-up and top-down attention for image captioning and VQA. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. 6077–6086
- Qin Y, Du J, Zhang Y, et al. Look back and predict forward in image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019. 8367–8375
- Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Proceedings of Advances in Neural Information Processing Systems, 2017. 5998–6008
- Xu K, Ba J, Kiros R, et al. Show, attend and tell: neural image caption generation with visual attention. In: Proceedings of the International Conference on Machine Learning, 2015. 2048–2057
- Mao J, Xu W, Yang Y, et al. Explain images with multimodal recurrent neural networks. 2014. ArXiv:1410.1090
- Donahue J, Hendricks L A, Guadarrama S, et al. Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015. 2625–2634

- 7 Vinyals O, Toshev A, Bengio S, et al. Show and tell: a neural image caption generator. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015. 3156–3164
- 8 Yang Z, Yuan Y, Wu Y, et al. Review networks for caption generation. In: Proceedings of Advances in Neural Information Processing Systems, 2016. 2361–2369
- 9 You Q, Jin H, Wang Z, et al. Image captioning with semantic attention. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016
- 10 Liu S, Zhu Z, Ye N, et al. Optimization of image description metrics using policy gradient methods. 2016. ArXiv:1612.00370
- 11 Karpathy A, Fei-Fei L. Deep visual-semantic alignments for generating image descriptions. *IEEE Trans Pattern Anal Mach Intell*, 2016, 39: 664–676
- 12 Fu K, Jin J, Cui R, et al. Aligning where to see and what to tell: image captioning with region-based attention and scene-specific contexts. *IEEE Trans Pattern Anal Mach Intell*, 2016, 39: 2321–2334
- 13 Wu Q, Shen C, Wang P, et al. Image captioning and visual question answering based on attributes and external knowledge. *IEEE Trans Pattern Anal Mach Intell*, 2017, 40: 1367–1381
- 14 Rennie S J, Marcheret E, Mroueh Y, et al. Self-critical sequence training for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. 7008–7024
- 15 Yao T, Pan Y, Li Y, et al. Boosting image captioning with attributes. In: Proceedings of the IEEE International Conference on Computer Vision, 2017. 4894–4902
- 16 Yao T, Pan Y, Li Y, et al. Exploring visual relationship for image captioning. In: Proceedings of the European Conference on Computer Vision, 2018. 684–699
- 17 Park C C, Kim B, Kim G. Towards personalized image captioning via multimodal memory networks. *IEEE Trans Pattern Anal Mach Intell*, 2018, 41: 999–1012
- 18 Zha Z J, Liu D, Zhang H, et al. Context-aware visual policy network for fine-grained image captioning. *IEEE Trans Pattern Anal Mach Intell*, 2022, 44: 710–722
- 19 Gao L, Li X, Song J, et al. Hierarchical LSTMs with adaptive attention for visual captioning. *IEEE Trans Pattern Anal Mach Intell*, 2020, 42: 1112–1131
- 20 Ji J, Xu C, Zhang X, et al. Spatio-temporal memory attention for image captioning. *IEEE Trans Image Process*, 2020, 29: 7615–7628
- 21 Liu S, Ren Z, Yuan J. SibNet: sibling convolutional encoder for video captioning. *IEEE Trans Pattern Anal Mach Intell*, 2021, 43: 3259–3272
- 22 Li Y, Yao T, Pan Y, et al. Contextual transformer networks for visual recognition. *IEEE Trans Pattern Anal Mach Intell*, 2022. doi: 10.1109/TPAMI.2022.3164083
- 23 Li Y, Pan Y, Yao T, et al. Comprehending and ordering semantics for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2022
- 24 Li Y, Pan Y, Chen J, et al. X-modaler: a versatile and high-performance codebase for cross-modal analytics. In: Proceedings of the ACM International Conference on Multimedia, 2021. 3799–3802
- 25 Yao T, Pan Y, Li Y, et al. Hierarchy parsing for image captioning. In: Proceedings of the IEEE International Conference on Computer Vision, 2019. 2621–2629
- 26 Jiang W, Ma L, Jiang Y G, et al. Recurrent fusion network for image captioning. In: Proceedings of the European Conference on Computer Vision, 2018. 499–515
- 27 Yang X, Tang K, Zhang H, et al. Auto-encoding scene graphs for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019. 10685–10694
- 28 Wang L, Bai Z, Zhang Y, et al. Show, recall, and tell: image captioning with recall mechanism. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2020. 12176–12183
- 29 Sammani F, Melas-Kyriazi L. Show, edit and tell: a framework for editing image captions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020. 4808–4816
- 30 Lu J, Batra D, Parikh D, et al. ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: Proceedings of Advances in Neural Information Processing Systems, 2019
- 31 Zhou L, Palangi H, Zhang L, et al. Unified vision-language pre-training for image captioning and VQA. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2020. 13041–13049
- 32 Li X, Yin X, Li C, et al. Oscar: object-semantics aligned pre-training for vision-language tasks. In: Proceedings of the European Conference on Computer Vision, 2020. 121–137
- 33 Zhang P, Li X, Hu X, et al. VinVL: revisiting visual representations in vision-language models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2021. 5579–5588
- 34 Chen J, Lian Z H, Wang Y Z, et al. Irregular scene text detection via attention guided border labeling. *Sci China Inf Sci*, 2019, 62: 220103
- 35 Ye Y Y, Zhang C, Hao X L. ARPNET: attention region proposal network for 3D object detection. *Sci China Inf Sci*, 2019, 62: 220104
- 36 He N J, Fang L Y, Plaza A. Hybrid first and second order attention Unet for building segmentation in remote sensing images. *Sci China Inf Sci*, 2020, 63: 140305
- 37 Li Z C, Tang J H. Semi-supervised local feature selection for data classification. *Sci China Inf Sci*, 2021, 64: 192108
- 38 Jin J, Fu K, Cui R, et al. Aligning where to see and what to tell: image caption with region-based attention and scene factorization. 2015. ArXiv:1506.06272
- 39 Lu J, Xiong C, Parikh D, et al. Knowing when to look: adaptive attention via a visual sentinel for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. 375–383
- 40 Pedersoli M, Lucas T, Schmid C, et al. Areas of attention for image captioning. In: Proceedings of the IEEE International Conference on Computer Vision, 2017. 1242–1250
- 41 Wang J, Pan Y, Yao T, et al. Convolutional auto-encoding of sentence topics for image paragraph generation. In: Proceedings of the International Joint Conference on Artificial Intelligence, 2019. 940–946
- 42 Pan Y, Yao T, Li Y, et al. X-Linear attention networks for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020. 10971–10980
- 43 Wang J, Tang J, Yang M, et al. Improving OCR-based image captioning by incorporating geometrical relationship. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2021. 1306–1315
- 44 Wang J, Tang J, Luo J. Multimodal attention with image text spatial relationship for OCR-based image captioning. In: Proceedings of the ACM International Conference on Multimedia, 2021. 4337–4345

- 45 Williams R J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach Learn*, 1992, 8: 229–256
- 46 Huang L, Wang W, Chen J, et al. Attention on attention for image captioning. In: *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 4634–4643
- 47 Graves A, Wayne G, Danihelka I. Neural Turing machines. 2014. ArXiv:1410.5401
- 48 Weston J, Chopra S, Bordes A. Memory networks. In: *Proceedings of the International Conference on Learning Representations*, 2015
- 49 Graves A, Wayne G, Reynolds M, et al. Hybrid computing using a neural network with dynamic external memory. *Nature*, 2016, 538: 471–476
- 50 Sukhbaatar S, Weston J, Fergus R, et al. End-to-end memory networks. In: *Proceedings of Advances in Neural Information Processing Systems*, 2015
- 51 Meng F, Tu Z, Cheng Y, et al. Neural machine translation with key-value memory-augmented attention. In: *Proceedings of the International Joint Conference on Artificial Intelligence*, 2018. 2574–2580
- 52 Meng F, Zhang J. DTMT: a novel deep transition architecture for neural machine translation. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019. 224–231
- 53 Kumar A, Irsoy O, Ondruska P, et al. Ask me anything: dynamic memory networks for natural language processing. In: *Proceedings of the International Conference on Machine Learning*, 2016. 1378–1387
- 54 Xiong C, Merity S, Socher R. Dynamic memory networks for visual and textual question answering. In: *Proceedings of the International Conference on Machine Learning*, 2016. 2397–2406
- 55 Zhang J, Shi X, King I, et al. Dynamic key-value memory networks for knowledge tracing. In: *Proceedings of the International Conference on World Wide Web*, 2017. 765–774
- 56 Chen X, Xu H, Zhang Y, et al. Sequential recommendation with user memory networks. In: *Proceedings of the ACM International Conference on Web Search and Data Mining*, 2018. 108–116
- 57 Yang T, Chan A B. Learning dynamic memory networks for object tracking. In: *Proceedings of the European Conference on Computer Vision*, 2018. 152–167
- 58 Shankar S, Garg S, Sarawagi S. Surprisingly easy hard-attention for sequence to sequence learning. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2018. 640–645
- 59 Collier M, Beel J. Implementing neural Turing machines. In: *Proceedings of the International Conference on Artificial Neural Networks*, 2018. 94–104
- 60 Dauphin Y N, Fan A, Auli M, et al. Language modeling with gated convolutional networks. In: *Proceedings of the International Conference on Machine Learning*, 2017. 933–941
- 61 Shankar S, Sarawagi S. Posterior attention models for sequence to sequence learning. In: *Proceedings of the International Conference on Learning Representations*, 2018
- 62 Chen X, Fang H, Lin T Y, et al. Microsoft COCO captions: data collection and evaluation server. 2015. ArXiv:1504.00325
- 63 Karpathy A, Fei-Fei L. Deep visual-semantic alignments for generating image descriptions. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 3128–3137
- 64 Banerjee S, Lavie A. Meteor: an automatic metric for MT evaluation with improved correlation with human judgments. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005. 65–72
- 65 Vedantam R, Zitnick C L, Parikh D. CIDEr: consensus-based image description evaluation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 4566–4575
- 66 Papineni K, Roukos S, Ward T, et al. BLEU: a method for automatic evaluation of machine translation. In: *Proceedings of the Association for Computational Linguistics*, 2002. 311–318
- 67 Lin C Y. ROUGE: a package for automatic evaluation of summaries. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics Workshop: Text Summarization Branches Out 2004*, 2004. 74–81
- 68 Anderson P, Fernando B, Johnson M, et al. Spice: semantic propositional image caption evaluation. In: *Proceedings of the European Conference on Computer Vision*, 2016. 382–398
- 69 Kingma D P, Ba J. Adam: a method for stochastic optimization. In: *Proceedings of the International Conference on Learning Representations*, 2015