



Neural Image Caption Generation with Weighted Training and Reference

Guiguang Ding¹ · Minghai Chen¹ · Sicheng Zhao¹ · Hui Chen¹ · Jungong Han² · Qiang Liu¹

Received: 17 March 2018 / Accepted: 16 July 2018
© The Author(s) 2018

Abstract

Image captioning, which aims to automatically generate a sentence description for an image, has attracted much research attention in cognitive computing. The task is rather challenging, since it requires cognitively combining the techniques from both computer vision and natural language processing domains. Existing CNN-RNN framework-based methods suffer from two main problems: in the training phase, all the words of captions are treated equally without considering the importance of different words; in the caption generation phase, the semantic objects or scenes might be misrecognized. In our paper, we propose a method based on the encoder-decoder framework, named Reference based Long Short Term Memory (R-LSTM), aiming to lead the model to generate a more descriptive sentence for the given image by introducing reference information. Specifically, we assign different weights to the words according to the correlation between words and images during the training phase. We additionally maximize the consensus score between the captions generated by the captioning model and the reference information from the neighboring images of the target image, which can reduce the misrecognition problem. We have conducted extensive experiments and comparisons on the benchmark datasets MS COCO and Flickr30k. The results show that the proposed approach can outperform the state-of-the-art approaches on all metrics, especially achieving a 10.37% improvement in terms of CIDEr on MS COCO. By analyzing the quality of the generated captions, we come to a conclusion that through the introduction of reference information, our model can learn the key information of images and generate more trivial and relevant words for images.

Keywords Image captioning · Reference · Long short-term memory · Encoder-decoder

Introduction

Benefiting from the significant advances of large-scale labeled datasets, such as ImageNet [7] and deep learning,

This work was supported by the National Natural Science Foundation of China (Nos. 61571269, 61701273), and the Project Funded by China Postdoctoral Science Foundation (Nos. 2018T110100, 2017M610897).

✉ Sicheng Zhao
schzhao@gmail.com

Jungong Han
jungonghan77@gmail.com

Guiguang Ding
dinggg@tsinghua.edu.cn

¹ School of Software, Tsinghua University, Beijing 100084, People's Republic of China

² Computing and Communications, Lancaster University, Lancaster, UK

especially deep convolutional neural networks (CNN) [24, 33]; the problems of image classification [73] and object recognition [56, 72] have been studied thoroughly. As a result, computers even outperform humans at these tasks [60]. Recently, automatically generating a sentence description for an image, has attracted much research attention in artificial intelligence. This problem, known as image captioning, plays an important role in computer vision, i.e., enabling computers to understand images, which can be exploited in wide applications, such as video tracking [28–32], cross-view retrieval [10, 39], sentiment analysis [36, 52], childhood education [54], and visual impairment rehabilitation [11]. However, image captioning is a challenging task due to the coverage of both computer vision and natural language processing technologies. Apart from the need for identifying the objects contained in an image [37], the generator should also be able to analyze their states, understand the relationship among them, and express the semantic information in natural language [59].

The early efforts on image captioning mainly adopted the template-based methods, which require recognizing the various elements, such as objects as well as their attributes and relationships in the first phase. These elements are then organized into sentences based on either templates [15, 25, 35, 62] or pre-defined language models [13, 26, 27, 46], which normally end up with rigid and limited descriptions. As a typical transfer-based method, nearest neighbor (NN) is employed to retrieve a description from the corpus for a given image [9]. Although this method cannot generate any novel sentence, it suggests that NN can indeed provide valuable information.

Inspired by recent advances in machine translation [6, 41, 55, 57], neural network-based methods have been widely applied in image captioning tasks [12, 14, 23, 43, 59] and achieved great success. These methods are primarily based on the encoder-decoder pipeline, consisting of two basic steps. First, visual features are extracted using CNN to encode the image into a fixed length embedding vector. Second, recurrent neural network (RNN), especially long short-term memory (LSTM) [18] is adopted as the decoder to generate the sentence description by maximizing the likelihood of a sentence given the visual features. Thanks to the feature representation capability of CNN and the temporal modeling of RNN, the neural network-based methods are more flexible, which can generate new sentences coherently. On the other hand, motivated by the attention [58] mechanisms, which have been proven to be effective in visual scene analysis [2, 47], different attention mechanisms are proposed for image captioning, such as region-based attention [21], visual attention [61], semantic attention [65], global-local attention [34], and spatial and channel-wise attention [3].

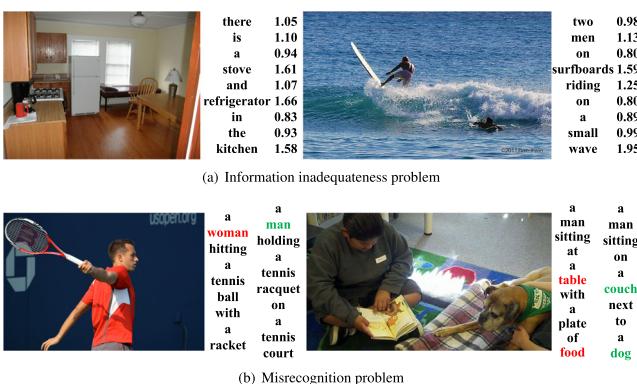


Fig. 1 Motivation of the proposed model. **a** Traditional methods treat all the words (on the right of each image) equally without considering the relative importance. On the right of the words are the assigned weights based on the overall occurrences. **b** The main subjects or scenes are misrecognized using the traditional methods (red), which can be corrected with the help of consensus score between the neighboring references and the target image (green)

Despite achieving the state-of-the-art performance, existing CNN-RNN framework-based methods suffer from two main problems, as illustrated in Fig. 1:

- *Information inadequateness problem.* These methods treat different words of a caption equally, which makes distinguishing the important parts of the caption difficult.
 - *Misrecognition problem.* The main subjects or scenes might be misrecognized using the traditional methods.

Obviously, in an image description, the words are not equally important. Take the first image of Fig. 1a as an example, the words “surfboards” and “wave” should be the most important as they constitute the main content of the image; “men” is the subject and “riding” is the status of the subject, which are less important; “two” “on” “a” “small” are relatively uninformative. Furthermore, once the subjects or scenes are misrecognized, the generation error would accumulate and cannot be easily corrected. The subsequent words in the caption may be disturbed by these irrelevant text context. Motivated by these observations, we propose to make use of the visual features and the labeled captions of the training images as references to address the above mentioned problems. The references are incorporated in both the training phase and the generation phase of the LSTM framework, which constitute the novel R-LSTM model. In the training phase, we endow the words in a caption with different weights in terms of their relevance to the target image, part of speech, and corresponding synonyms. A word with higher relevance score indicates high importance to describe the image, and thus a larger weight value is assigned to it when calculating the loss. In this way, the model could learn more in-depth information of the caption, such as what the principal objects are, which attributes are important to them and how they relate to each other. In the generation phase, the NNs of the input image are employed as references by jointly combining the consensus score [8] and the likelihood of generating sentence. The information provided by the NNs could help reduce the misrecognition from beginning, and better match the habit of human cognition.

We evaluate the proposed R-LSTM model on the MS COCO and Flickr30k datasets. The comparative results demonstrate the significant superiority of R-LSTM over the state-of-the-art approaches. We also report the performance of our method on the MS COCO Image Captioning Challenge. Comparing with all the latest approaches, we obtain comparable performances on all the 14 metrics.

The main contributions of this paper are threefold:

1. We propose to use the training images as *references* and design a novel model, named Reference based Long Short Term Memory (R-LSTM), for image captioning.

2. In the training phase, we assign unequal weights to different words according to the overall occurrences, part of speech, and corresponding synonyms. Such a training biased by the weights can better learn the in-depth information of the captions, which helps to address information inadequateness problem.
3. In the caption generation phase, we define a novel objective function by combining the consensus score and the traditional log likelihood to exploit the reference information from neighbor images of the target image. Reference-based generation can help to address the misrecognition problem and make the descriptions more natural sounding.

One preliminary conference version on R-LSTM for image captioning was first introduced in our previous work [4]. The enhancements in this paper as compared to the conference version lie in the following three aspects. First, we perform a more comprehensive review of related work. Second, we extend the weighted training by combining the overall occurrences, part of speech and corresponding synonyms. Third, we conduct more comparative experiments and enrich the analysis of results.

The rest of this paper is structured as follows. Section “[Related Work](#)” briefly reviews related work on image captioning. Section “[System Overview](#)” gives an overview of the proposed model. Detailed algorithms, including weighted training and reference-based generation are described in “[Weighted Training](#)” and “[Generation Using Reference](#),” respectively. Experimental results and analysis are presented in “[Experiments](#),” followed by a conclusion and the summary of future works in “[Conclusion](#).”

Related Work

Generally, the existing image captioning algorithms can be divided into three categories based on the way of sentence generation [20]: template-based methods, transfer-based methods, and neural network-based methods.

The *template-based methods* either use templates or design a language model, which fill in slots of a template based on co-occurrence relations gained from the corpus [15], conditional random field [25], or web-scale n -gram data [35]. More complicated models have also been used to generate relatively flexible sentences. Mitchell et al. [46] exploited syntactic trees to create a data-driven model. Visual dependency representation is proposed to extract relationships among the objects [13]. The template-based methods are simple and intuitive, but are heavily hand-designed and unexpressive, which are not flexible enough to generate meaningful sentences.

The *transfer-based methods* are based on the retrieval approaches, which directly transfer the descriptions of the retrieved images to the query image. Some approaches [16, 19] took the input image as a query and selected a description in a joint image-sentence embedding space. Kuznetsova et al. [26, 27] retrieved images that are similar to the input image, extracted segments from their captions, and organized these segments into a sentence. Devlin et al. [9] simply found similar images and calculated the consensus score [8] of the corresponding captions to select the one with the highest score. The generated sentences by the transfer-based methods are often with correct grammar. However, these methods may misrecognize the visual content and cannot generate novel phrases or sentences, and thus are limited in image captioning. Notwithstanding, they indicate that we can take advantage of the images similar to the input image. This idea can be applied in other approaches, such as re-ranking candidate descriptions generated by other models [44] and emotion distribution prediction [70, 71]. We also undertake this idea in our generation process.

The *neural network-based methods* come from the recent advantages in machine translation [6, 55, 57], with the use of RNN. Mao et al. [43] proposed a multimodal layer to connect a deep CNN for images and a deep RNN for sentences, allowing the model to generate the next word given the input word and the image. Inspired by the encoder-decoder model [6] in machine translation, Vinyals et al. [59] used a deep CNN to encode the image instead of a RNN for sentences, and then used LSTM [18], a more powerful RNN, to decode the image vector to a sentence. Many works follow this idea, and apply attention mechanisms in the encoder. Xu et al. [61] extracted features from a convolutional layer rather than the fully connected layer. With each feature representing a fixed-size region of the image, the model can learn to change the focusing locations. Jin et al. [21] employed a pre-trained CNN for object detection to analyze the hierarchically segmented image, and then ran attention-based decoder on these visual elements. Combining the whole image feature with the words obtained from the image by attribute detectors can also drive the attention model [65]. Li et al. [34] proposed a global-local attention mechanism by integrating local representation at object-level with global representation at image level.

More recently, reinforcement learning has been integrated in the encoder-decoder framework. To address the deficiencies of exposure bias and a loss, which does not operate at the sequence level in traditional encoder-decoder frameworks, Ranzato et al. [51] proposed a novel sequence level training algorithm, named Mixed Incremental Cross-Entropy Reinforce (MIXER), that directly optimizes the

metric used at test time. Liu et al. [40] proposed a novel training procedure for image captioning models based on policy gradient methods, which directly optimize for the metrics of interest, rather than just maximizing likelihood of human generated captions. Self-critical sequence training (SCST) [53] is proposed for image captioning by utilizing the output of its own test-time inference algorithm to normalize the rewards it experiences. It is noticed that directly optimizing the CIDEr metric with SCST and greedy decoding at test time is highly effective.

Please note that there are other captioning tasks that are related to our research, such as dense captioning [22] and video captioning [1, 48].

Similarly, our work follows the encoder-decoder model. But different from [59], the words in a caption are weighted in the training phase according to their relevance to the corresponding image, which well balances the model with the importance of a word to the caption. In the generation phase, we take advantage of the consensus score [8] to improve the quality of the sentences. Different from [44], which simply used the consensus score to re-rank the final candidate descriptions, we use this score in the whole generation process, which means that the decoder takes the neighbors' information of the input image into account. With the likelihood of a sentence combined, we propose a better evaluation function than just maximizing the likelihood.

System Overview

Our goal is to generate a description sentence for an image. Suppose we have N training images I_1, I_2, \dots, I_N , which also denote related visual features. For image $I_n (n = 1, 2, \dots, N)$, we have M_n correct description sentences $S_{n1}, S_{n2}, \dots, S_{nM_n}$. Our task aims to maximize

the likelihood of the correct descriptions given the training images by the following:

$$\theta^* = \arg \max_{\theta} \sum_{n=1}^N \sum_{m=1}^{M_n} \mathcal{L}(S_{nm} | I_n; \theta), \quad (1)$$

where θ are the parameters of our model and $\mathcal{L}()$ is a pre-defined likelihood function. In the next section, we will firstly describe the conventional likelihood function for image captioning used in previous works [42, 64] (see Eq. 4), and then we will introduce the proposed likelihood objective function (see Eq. 5).

After training, we can generate a sentence for a test image J by the following:

$$R = \arg \max_{R'} \mathcal{O}(R' | J; \theta^*), \quad (2)$$

where $\mathcal{O}()$ is a pre-defined objective function. This objective function aims to generate the best sentence for the given image J . Usually, the log likelihood function is previously employed to replace $\mathcal{O}()$. However, the conventional log likelihood function roughly selects the sentence with the highest probability learned by the model and may cause the misrecognition issue. In this paper, we introduce the supervision of the reference sentence and reformulate this objective function, aiming to resolve the misrecognition issue in the generation stage. The details are provided in “[Generation Using Reference](#).”

The overview of the proposed image captioning method is illustrated in Fig. 2, which consists of two stages: weighted training and reference-based generation. For both stages, the deep ResNet-101 model [17] is employed as the encoder to extract CNN features of the target image and the training images. During the weighted training stage, the weight attached to each word in the training captions is calculated firstly. Then, the LSTM model is trained using the weighted words and CNN features of the training images

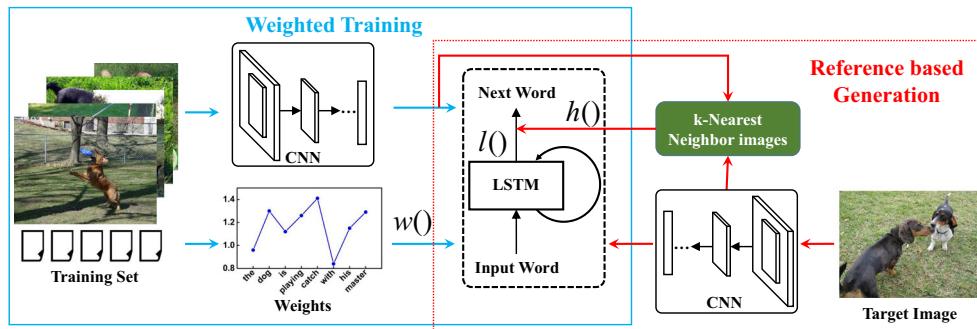


Fig. 2 Overview of the proposed R-LSTM model, including two parts: weighted training (in blue rectangle) and reference-based generation (in red rectangle). Each part is an encoder-decoder model, using CNN to encode the image and LSTM to decode the sentence. The functions

$w()$ and $h()$ indicate that the reference information is used to weigh the input words when training and improve the output sentences when generating, respectively. $l()$ is the log likelihood

under the proposed weighted likelihood objective. In the reference-based generation stage, the trained LSTM plays as a decoder role, which takes the CNN features of the target image as input and generates the description words one by one. Besides, in the generation stage, we jointly consider the likelihood and the consensus score as the evaluation function in beam search. Details can be referred to “Weighted Training” and “Generation Using Reference.”

Weighted Training

For similarity, we use I and S to replace I_n and S_{nm} in Eq. 1, respectively. Suppose $S = \{s_0, s_1, s_2, \dots, s_T, s_{T+1}\}$, where $\{s_1, s_2, \dots, s_T\}$ is the original labeled words, s_0 is a special start word, s_{T+1} is a special stop word and T is the length of this particular sentence, which depends on I . At time t , the likelihood of word s_t is decided by the input image I and previous words s_0, s_1, \dots, s_{t-1} :

$$p(s_t|I, s_0, s_1, \dots, s_{t-1}). \quad (3)$$

The joint log likelihood of description S , namely the objective likelihood function $\mathcal{L}()$ in Eq. 1 of the NIC model [59], is calculated by the chain rule:

$$\mathcal{L}(S|I) = \log p(S|I) = \sum_{t=1}^{T+1} \log p(s_t|I, s_0, s_1, \dots, s_{t-1}), \quad (4)$$

where the dependency on θ is dropped for convenience.

As stated in “Introduction” and illustrated in Fig. 1, different words are not equally meaningful and important. It is reasonable that the subject and its corresponding status express more information than the articles and prepositions. Unlike the NIC model [59], we take into consideration the words’ importance by assigning different weights to the words, which enables the model to be concentrated on the main information of the captions. Logically, we assign higher weights to the words, which correspond to important elements, such as the main subject, its status, and the environment, etc. Suppose the weight of word s_t to image I is $w(s_t, I)$, then our model is trained to maximize the weighted log likelihood:

$$\mathcal{L}(S, I) = \sum_{t=1}^{T+1} w(s_t, I) \log p(s_t|I, s_0, s_1, \dots, s_{t-1}). \quad (5)$$

Note that in the training phase, the words s_0, s_1, \dots, s_t are given by the labeled caption. So their weights could be calculated as a preprocessing step.

There are different ways to calculate the weights of different words. We propose three schemes based on the overall occurrences, part of speech, and corresponding synonyms, respectively. The overall occurrence-based scheme

follows the tag ranking approach [38] by calculating the weight of word s_i to image I as follows:

$$w_1(s_i, I) = \frac{\beta p(s_i|I)}{p(s_i)}, \quad i = 1, 2, \dots, N, \quad (6)$$

where β is a parameter to ensure the average of all the weights is 1, and $p(s_i|I)$ denotes the likelihood of s_i in the captions of image I . The reason for dividing $p(s_i|I)$ by $p(s_i)$ is that a frequent word, such as “a” and “the,” is not informative although it may appear in most descriptions.

Based on Bayes’ rule, we have as follows:

$$w_1(s_i, I) = \frac{\beta P(I|s_i)P(s_i)}{P(I)P(s_i)} = \frac{\beta P(I|s_i)}{P(I)}. \quad (7)$$

Since $P(I)$ is determined given image I , we can redefine (7) as follows:

$$w_1(s_i, I) \doteq \beta P(I|s_i). \quad (8)$$

Based on kernel density estimation (KDE) [50],

$$w_1(s_i, I) = \beta P(I|s_i) = \frac{\beta}{|G_{s_i}|} \sum_{I_j \in G_{s_i}} K_\sigma(I - I_j), \quad (9)$$

where G_{s_i} denotes the set of images whose captions contain word s_i , and the Gaussian kernel function K_σ is defined as follows:

$$K_\sigma(I - I_j) = \exp\left(-\frac{\|I - I_j\|^2}{\sigma^2}\right), \quad (10)$$

where the radius parameter σ is set as the average distance of each two images in the training set, and the image vectors are extracted from a deep CNN. Therefore, in a set of images containing a same description word, if an image is very similar to others, it is natural to infer that the word is very relevant to the image, and thus will be assigned with a high weight in the image’s captions. Otherwise, if an image does not look like other images, which means that the word is not important or is even noise to the image, the word will be given a low weight. Equation 10 is meaningful in two aspects: it measures the importance of different words in a same caption (Fig. 3) and the importance of a word to different images (Fig. 4).

The part-of-speech-based strategy is conducted on the basis of the overall occurrences. We observe that the nouns and verbs in the captions are relatively more important and convey more information than others. The contributions of prepositions, conjunctions, and qualifiers to the whole caption sentence are relatively small. Therefore, we enlarge the weights of all the nouns and verbs, and reduce the weights of all the prepositions, conjunctions, and qualifiers by the following:

$$w_2(s_i, I) = \mu_{s_i} * w_1(s_i, I), \quad (11)$$

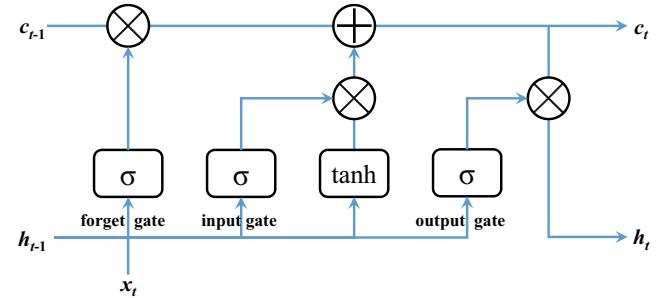
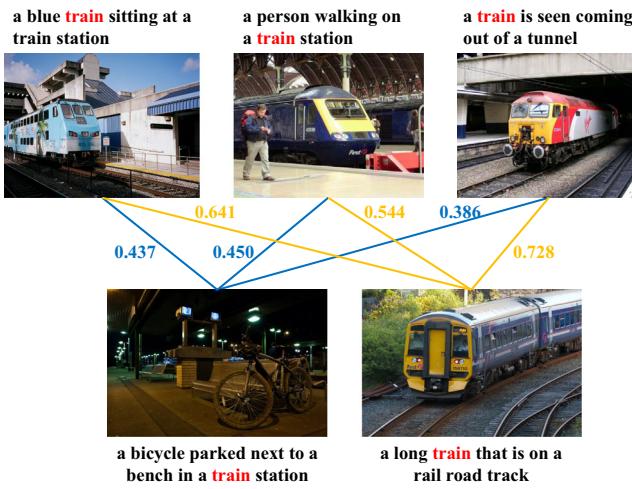
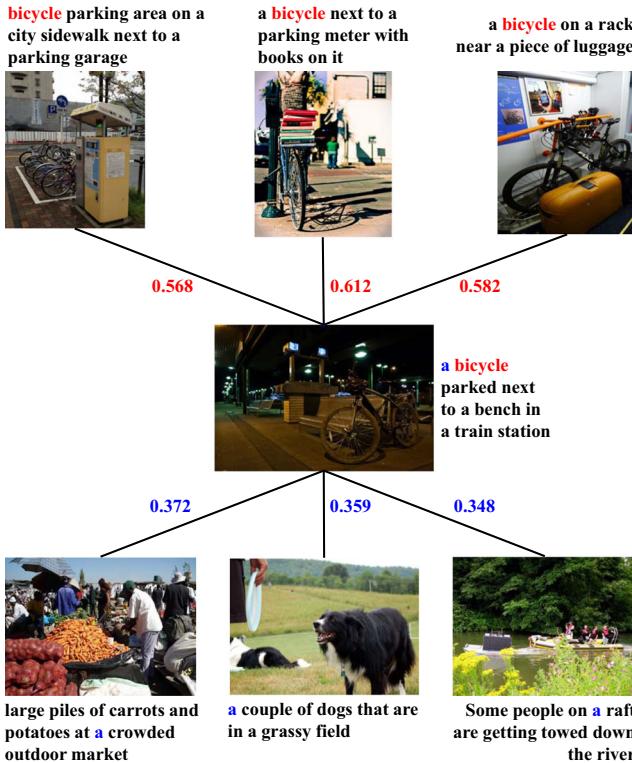


Fig. 5 A typical LSTM unit, consisting of forget gate, input gate, and output gate

where μ_{s_i} is the parameter that controls the relative importance, the value of which depends on the different part of speech as follows:

$$\mu_{s_i} = \begin{cases} \mu_1, & \text{if } \text{pos}(s_i) = \text{noun}, \\ \mu_2, & \text{if } \text{pos}(s_i) = \text{verb}, \\ \mu_3, & \text{if } \text{pos}(s_i) = \text{qualifier}, \\ \mu_4, & \text{if } \text{pos}(s_i) = \text{preposition/conjunction}, \end{cases} \quad (12)$$

where $\text{pos}(s_i)$ is the part of speech of word s_i . The values of different part of speeches have a great impact on learning the primary parts of sentence. Therefore, we assign different weights to different part of speeches according to the information they have. Generally, nouns and verbs are much more informative to the image than qualifiers or prepositions/conjunctions, and thus make more contributions to the quality of the sentence. Motivated by this observation, we endow more values to the noun and the verb to direct the model to learn the informative words better. While for qualifiers, prepositions, and conjunctions, we put less weights to them. In the experiment, μ_1, μ_2, μ_3 , and μ_4 are set to be 1.1, 1.05, 0.9, and 0.8, respectively.

The third scheme is based on the synonyms. We observe that many different words have similar meanings, such as role, character, and function. Jointly, modeling them together to calculate the weights may complement each other. We propose a weighting strategy based on words’ similarity by considering the semantic information. For word s_i , suppose the synonym set (‘synset’) of its k th meaning is ss_{ik} by WordNet,¹ we compute the weight of the synset as follows:

$$\widehat{w}_3(ss_{ik}, I) = \frac{\beta}{|G_{ss_{ik}}|} \sum_{I_j \in G_{ss_{ik}}} K_\sigma(I - I_j), \quad (13)$$

where $G_{ss_{ik}}$ denotes the set of images whose captions contain “synset” ss_{ik} . Since a word belongs to several different

¹<https://wordnet.princeton.edu>

“synsets,” we use the maximum value as the final weight:

$$w_3(s_i, I) = \max_{k \in C_{s_i}} \hat{w}_3(\text{ss}_{ik}, I), \quad (14)$$

where C_{s_i} is the set of different semantic meanings of s_i .

After obtaining the encoded CNN features of the training images and the weights attached to each word in the training captions, we can train the LSTM model, as shown in Fig. 5, by optimizing (1) with likelihood function defined in Eq. 5. The detailed operations of a LSTM unit are as follows:

$$\begin{aligned} x_{-1} &= I, \quad x_t = W_e s_t (t = 0, 1, \dots, T - 1), \\ f_t &= \sigma(W_{fx}x_t + W_{fh}h_{t-1} + b_f), \\ i_t &= \sigma(W_{ix}x_t + W_{ih}h_{t-1} + b_i), \\ o_t &= \sigma(W_{ox}x_t + W_{oh}h_{t-1} + b_o), \\ c_t &= f_t \otimes c_{t-1} \oplus i_t \otimes \tanh(W_{cx}x_t + W_{ch}h_{t-1} + b_c), \\ h_t &= o_t \otimes c_t, \\ p_{t+1} &= \text{Softmax}(h_t), \end{aligned} \quad (15)$$

where \otimes , \oplus , $\tanh(\cdot)$, and $\sigma(\cdot)$ are the product with a gate value, sum operation, hyperbolic tangent function, and sigmoid function, respectively. W_{*x} , W_{*h} , W_e , and b_{*} are the parameters learned by our model and shared in all steps.

Generation Using Reference

After training, the model can generate a description $R = \{r_0, r_1, r_2, \dots, r_M, r_{M+1}\}$ (r_0 and r_{M+1} are special start word and stop word, respectively) for test image J by optimizing the following objective function:

$$\mathcal{O}(R, J) = (1 - \alpha)l(R, J) + \alpha h(R, J), \quad (16)$$

where $h(R, J)$ is the consensus score of sentence R , and $l(R, J)$ is the log likelihood:

$$l(R, J) = \log p(R|J) = \sum_{t=1}^{M+1} \log p(r_t|J, r_0, r_1, \dots, r_{t-1}), \quad (17)$$

which is obtained by Eq. 15. When $\alpha = 1$, the objective function $\mathcal{O}(R, J)$ turns to the one in the NIC model [59].

The idea of consensus score comes from transfer-based methods, which indicate that the descriptions of similar images are very helpful in image captioning. Some existing transfer-based methods directly use the captions of the similar images as the description of the input image. For example, Devlin et al. [9] simply utilized the k -nearest neighbor model. First, retrieve k nearest neighbors of the input image and obtain the set of their captions $C = \{c_1, c_2, \dots, c_{5k}\}$ (five captions for each image). Second, calculate the n -gram overlap F score for every two captions in C . The consensus score of c_i is defined as the mean of its top m F scores. Finally, select the caption with the highest consensus score as the description of the input image.

Similar to [9], we calculate the consensus score $h(R, J)$ for image J and the generated sentence R (including incomplete ones that are being generated by the decoder) as follows:

$$h(R, J) = \frac{1}{|C_J|} \sum_{c \in C_J} \text{sim}(R, c), \quad (18)$$

where C_J is the caption set of the k -nearest neighbor images of image J , and $\text{sim}(\cdot, \cdot)$ is the function to calculate the similarity between two sentences (we use BLEU-4 [49] in experiments).

Since $l(R, J)$ is much larger than $h(R, J)$ in terms of absolute value, we normalize them before linear weighting:

$$\begin{aligned} l'(R, J) &= \frac{l(R, J) - \min_{c \in H} l(c, J)}{\max_{c \in H} l(c, J) - \min_{c \in H} l(c, J)}, \\ h'(R, J) &= \frac{h(R, J) - \min_{c \in H} h(c, J)}{\max_{c \in H} h(c, J) - \min_{c \in H} h(c, J)}, \end{aligned} \quad (19)$$

where H is the set of generated candidate descriptions. The final evaluation function is as follows:

$$\mathcal{O}(R, J) = (1 - \alpha)l'(R, J) + \alpha h'(R, J), \quad 0 \leq \alpha \leq 1. \quad (20)$$

Different from training, in the generation phase, the labeled captions are no longer available, and the input word at time t is the output word r_{t-1} . Besides, as our dictionary size is large, which is up to about 10000 words after filtering out infrequent ones on the MS COCO dataset, the searching space is too large for an exhaustive enumeration. Therefore, we implement the beam search as an approximation. At each time step, we keep a set of K (called “beam size”) best sentences from K^2 candidates according to Eq. 20. When a sentence is completed (the next word generated by the decoder is the stop word, or the sentence reaches the maximum length), it will be moved to the final pool, which also has the size of K and is maintained according to Eq. 20.

Experiments

To evaluate the effectiveness of the proposed method, we carry out extensive experiments on the Flickr30k dataset [66] and MS COCO dataset [5].

Experimental Settings

Datasets Flickr30k dataset [66] contains 31,783 images, while the more challenging MS COCO dataset [5] consists of 123,287 images. Each image is labeled with at least five captions by different Amazon Mechanical Turk (AMT) workers. Since there is no standardized split on both datasets, we follow the publicly available split² as in [19, 23, 61, 65] on Flickr30k dataset and in [23, 61, 65] on MS COCO for fair comparison. That is, 1000 images are used for validation, 1000 for testing and the rest for training in

²<https://github.com/karpathy/neuraltalk>

Flickr30; 5000 images are selected for validation, 5000 for testing, and the rest for training in MS COCO.

Evaluation Metrics Following the evaluation API provided by the MS COCO server, we report the results on different metrics, including BLEU-1, 2, 3, 4, METEOR, ROUGE-L and CIDEr. BLEU is based on the n -gram precision. METEOR is based on the harmonic mean of uni-gram precision and recall, which weighs recall higher than precision. Different from BLEU, METEOR seeks correlation at the corpus level. ROUGE-L is used to measure the common subsequence with maximum length between the target and source sentences. CIDEr is designed to evaluate image descriptions using human consensus. Higher values represent better performances for all these metrics.

Implementation Details The proposed R-LSTM model is implemented based on the NIC model [59]. The sentences are preprocessed following the publicly available code³. Unless specified otherwise, the beam size K used in the beam search is set to 10, similar to [21], while parameter α is set to 0.7 for Flickr30k and 0.4 for MS COCO. The LSTM cell size is 512 and the number of layers is 1. The image feature is extracted from the last 2048-dimensional fully connected layer of the ResNet-101 CNN model [17].

Results on Weighted Word Training

Some of the weighted words are shown in Fig. 6. Take the second image in the first row for example, considering the overall occurrences, the weights of the main subjects “girls” and “pizzas” are the largest, followed by the modifiers “little” and “pepperoni.” The part of speech strategy enlarges the weights of the nouns “girls” “pepperoni” and “pizzas” to emphasize the importance of these words. The synonyms-based method adjusts the weights based on the semantic meanings of corresponding synonyms. The last column combines the part-of-speech- and synonym-based methods together. We can conclude that after weighting, the main contents in the image are emphasized.

The performances of the LSTM networks trained before and after weighting the words are shown in the first and second rows in Tables 1 and 2 on Flickr30k and MO COCO datasets, respectively. The best performances are emphasized in italic. We can see that compared with the original NIC model [59], the performance is improved by all kinds of weighting schemes. On the CIDEr metric, the enhanced weighting versions “part of speech,” “synonyms,” and “part of speech + synonyms” outperform the previous “occurrences” version [4]. The combined weighting method generally performs better than others on MS COCO

dataset, while the “synonyms” method achieves the best performance on four metrics on Flickr30k dataset. Unless otherwise specified, we report the results of the combined weighting method in the following experiments.

Results on Reference-based Generation

To compare the performance contribution of weighted training and reference-based generation, we also conduct experiments on R-LSTM with no weighted training involved. The results on Flickr30k dataset and MS COCO dataset are shown in the last row of Tables 1 and 2, respectively. The best results of each column are highlighted in bold. It is clear that the reference-based generation achieves the best results on all the metrics with the significant performance gains. In view of this comparison, we can conclude that the reference-based generation contributes more than weighted training in the proposed R-LSTM model.

On Parameter α

The parameter α in Eq. 20 is crucial in our methods, which determines to what extent the generator depends on references. The black lines in Figs. 7a and 8a show how the quality of generated captions (on CIDEr) varies with respect to α on Flickr30k and MO COCO datasets, respectively. We can see that for both datasets with the increase of α , the performance firstly becomes better and then drops, which demonstrates that referring neighboring images can improve the performance but relying too much on references will lead to poor performance. The best α is 0.7 and 0.4 for Flickr30k and MO COCO datasets, respectively. We can conclude that the best α depends on the dataset.

In the generation phase, the sentence length is increasing. Since a sentence certainly becomes more informative when it has more words, it may not be a good idea to keep the same weight of the references. We try to change α in different generation stages. For similarity, in the early stage, we set $\alpha = \alpha_1$, and in the final pool stage, we set $\alpha = \alpha_2$. To adjust α_2 , we conduct experiment with $\alpha_1 = 0.7$ and $\alpha_1 = 0.4$ fixed for Flickr30k and MO COCO datasets, respectively. As shown in the blue lines of Figs. 7b and 8b, we can obtain better a performance by varying α_2 from 0.0 to 0.4, and the performance tends to decrease when $\alpha_2 > 0.4$. We can conclude that there exists the best α_2 for the specified α_1 . We repeat this process for $\alpha_1 = 0, 0.1, 0.2, \dots, 1$, and they all perform better by adjusting α_2 (the red lines in Figs. 7a and 8a). We believe that further changing α in more details (i.e., in each generation step) may achieve better performance, which is our future work. In the following experiments, we report the results when $\alpha_1 = 0.7$, $\alpha_2 = 0.4$ and $\alpha_1 = 0.4$, $\alpha_2 = 0.4$

³<https://github.com/karpathy/neuraltalk>

	a 0.90 0.81 0.90 0.81 wooden 0.97 1.07 0.97 1.07 box 0.96 1.06 0.98 1.07 with 0.91 0.73 0.91 0.73 compartments 1.05 1.16 1.02 1.12 filled 0.98 1.02 0.98 1.02 with 0.91 0.73 0.91 0.73 eating 0.85 0.89 0.84 0.90 utensils 1.07 1.18 1.06 1.16		two 1.04 1.04 1.04 1.04 adorable 1.12 1.12 1.10 1.10 little 1.15 1.15 1.15 1.15 girls 1.20 1.32 1.19 1.31 making 1.14 1.19 1.14 1.19 small 1.02 1.02 1.04 1.04 pepperoni 1.15 1.27 1.15 1.27 pizzas 1.25 1.37 1.22 1.35		a 1.06 0.95 1.06 0.95 horse 1.38 1.52 1.39 1.53 grazing 1.41 1.56 1.42 1.56 on 1.05 0.84 1.06 0.85 grass 1.28 1.41 1.29 1.42 on 1.05 0.84 1.06 0.85 a 1.06 0.95 1.06 0.95 farm 1.41 1.56 1.41 1.56 near 1.12 0.90 1.12 0.90 a 1.06 0.95 1.06 0.95 fence 1.28 1.41 1.29 1.42
	a 0.95 0.86 0.95 0.86 cat 1.17 1.28 1.17 1.28 sitting 1.00 1.05 1.00 1.05 on 0.95 0.76 0.95 0.76 top 0.98 1.08 0.98 1.08 of 0.96 0.77 0.96 0.77 a 0.95 0.86 0.95 0.86 book 1.19 1.30 1.20 1.32 shelf 1.18 1.30 1.17 1.29 filled 0.99 1.04 0.99 1.04 with 0.96 0.77 0.96 0.77 books 1.28 1.41 1.20 1.32		a 1.11 1.00 1.11 1.00 wooden 1.18 1.29 1.18 1.29 box 1.15 1.26 1.15 1.26 with 1.11 0.89 1.11 0.89 a 1.11 1.00 1.11 1.00 bird 1.17 1.29 1.15 1.27 in 1.11 0.89 1.11 0.89 a 1.11 1.00 1.11 1.00 tree 1.13 1.25 1.11 1.22 painted 1.19 1.25 1.20 1.26 on 1.10 0.88 1.11 0.88 it 1.15 1.15 1.14 1.14		a 1.05 0.94 1.05 0.94 bunch 1.04 1.15 1.07 1.18 of 1.05 0.84 1.05 0.84 cars 1.21 1.33 1.20 1.32 are 1.03 1.08 1.05 1.10 passing 1.15 1.26 1.15 1.26 by 1.07 0.86 1.08 0.86 a 1.05 0.94 1.05 0.94 couple 1.03 1.14 1.04 1.15 meters 1.40 1.54 1.38 1.52

Fig. 6 Results of weighted words. On the right of each image are the original caption words and assigned weights by the overall occurrences, part of speech, corresponding synonyms, and the combination of the latter two, respectively

Table 1 Performance comparison (%) of different weighting methods and reference-based generation on Flickr30k dataset after fine tuning the CNN encoder

	BL-1	BL-2	BL-3	BL-4	METEOR	ROU-L	CIDEr
Google NIC [59]	66.3	42.3	27.7	18.3	-	-	-
Occurrences	62.4	43.2	29.4	19.9	18.1	43.3	38.6
Part of speech	61.6	42.8	29.4	19.9	17.3	42.6	37.6
Synonyms	62.9	43.9	30.0	20.4	17.8	43.1	37.2
Part of speech							
+ synonyms	61.7	42.9	29.1	19.4	17.8	42.4	38.9
Reference-based generation	67.3	47.8	32.6	22.1	19.5	45.9	44.4

Table 2 Performance comparison (%) of different weighting methods and reference-based generation on MS COCO dataset after finetuning the CNN encoder

	BL-1	BL-2	BL-3	BL-4	METEOR	ROU-L	CIDEr
Google NIC [59]	66.6	46.1	32.9	24.6	-	-	-
Occurrences	71.1	54.2	40.5	30.5	25	52.5	95.7
Part of speech	71.7	54.7	41.1	31	25.1	52.9	95.8
Synonyms	71	54.1	40.6	30.7	25.0	52.6	95.4
Part of speech							
+ synonyms	71.4	54.6	41.1	31.2	25.2	53.0	96.1
Reference based generation	76.8	60.5	45.8	34.2	26.1	55.5	105.5

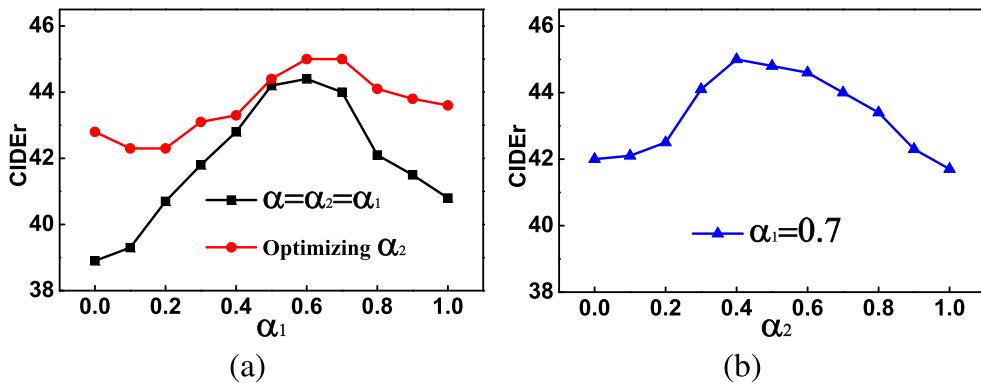


Fig. 7 The influence of α in the proposed generator on Flickr30k dataset. **a** The black and red lines are the influence of parameter α (i.e., $\alpha_2 = \alpha_1 = \alpha$) and α_1 (when optimizing α_2), respectively. **b** The blue line is the performance of the generator with different α_2 when $\alpha_1 = 0.7$

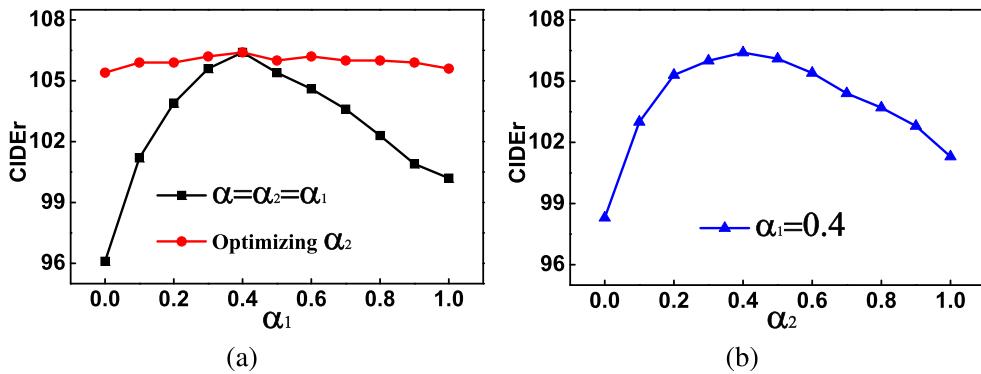


Fig. 8 The influence of α in the proposed generator on MS COCO dataset. **a** The black and red lines are the influence of parameter α (i.e., $\alpha_2 = \alpha_1 = \alpha$) and α_1 (when optimizing α_2), respectively. **b** The blue line is the performance of the generator with different α_2 when $\alpha_1 = 0.4$

for Flickr30k and MO COCO datasets respectively, unless otherwise specified.

We take the ninth image in Fig. 9, for example, to understand the beam search process of Eq. 20, i.e., the significance of $\alpha_1 \neq 0$. We can see that the subject “sheep” is misrecognized as “cattle” without using the consensus score, whose beam search process is illustrated in Fig. 10a. At the beginning, the model is waving between “sheep,” “cattle,” and ‘animal.’ As $\alpha_1 = 0$, the model cannot utilize the neighbor images to correct the mistake. When $t = 12$, there is no “sheep” in the candidate sentences. Regardless of the value of α_2 , this mistake cannot be corrected. However, when $\alpha_1 \neq 0$, this situation is avoidable with the help of

references, as shown in Fig. 10b, from which we can see that when $t = 8$, all the candidate sentences contain the correct subject “sheep.”

On Beam Search Size K

In order to analyze the effect of the beam search size K in the testing stage, we illustrate the performances on CIDEr of the best α_1 and α_2 as in “On Parameter α ” with the beam size in the range of {1, 2, 3, 5, 10, 20} in Fig. 11. We can see that the performances are like the “ \wedge ” shapes on both datasets when beam size K varies from 1 to 20. The best K is 10 for both datasets, which is the adopted beam size

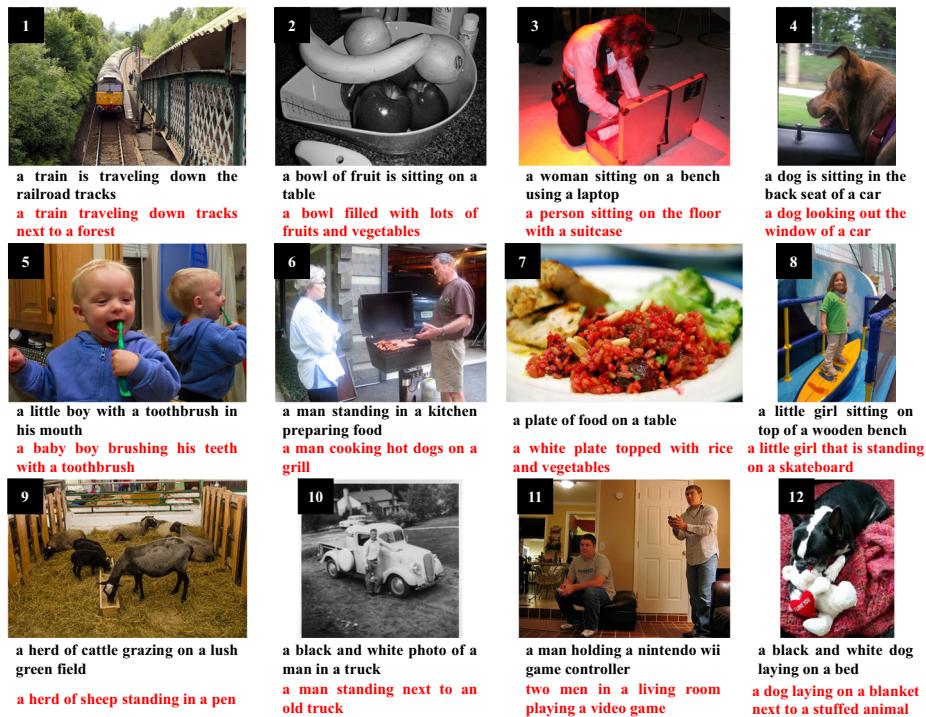


Fig. 9 Examples of generated captions by Google NIC [59] (in black) and the proposed R-LSTM model (in red)

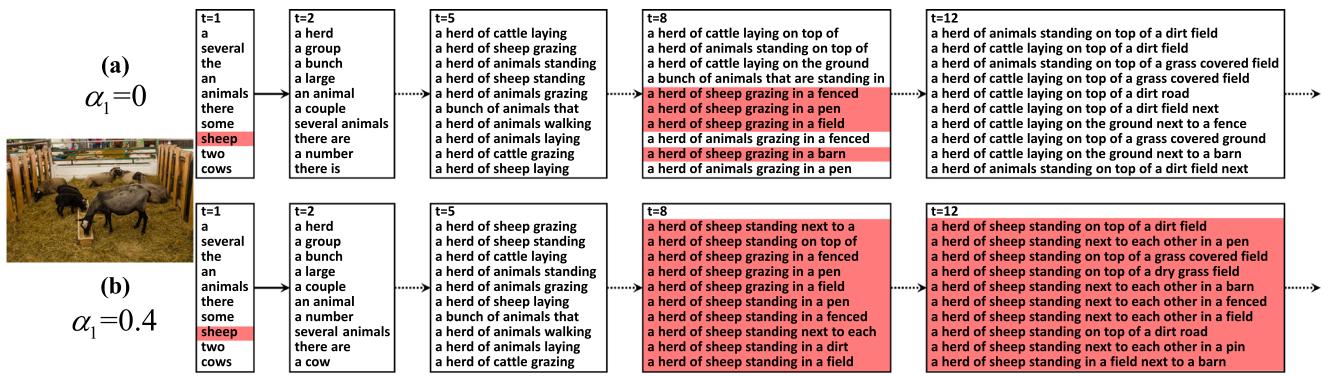


Fig. 10 The beam search process of the given image ranked by Eq. 20 when **a** $\alpha_1 = 0$, **b** $\alpha_1 = 0.4$ after weighted training. The red lines are the generating captions correctly recognizing the subject “sheep”

in experiment when comparing with other methods. We can conclude that a larger would not necessarily mean better performances.

Comparison with the State-of-the-arts

Tables 3 and 4 show the comparison of the completed model and several state-of-the-art methods on Flickr30k and MO COCO datasets respectively, where “-” represents unknown scores. From Table 3, we can see that the proposed method outperforms the state-of-the-art methods on all the metrics except for BLEU-4. It is clear from Table 4 that our approach performs the best on all the metrics, respectively achieving 5.52%, 8.45%, 9.83%, 10.26%, 5.60%, 4.50%, and 10.37% improvements compared with previous best results. These comparisons demonstrate the effectiveness of the proposed R-LSTM model for image captioning.

We also test our approach on the online MS COCO server (a sort of competition). The results compared with the latest methods are reported in Table 5. Despite keen competition, we are still one of the top ten methods in terms of the overall performance. It is noted that those methods,

which outperform our method, utilize either more complicated REINFORCE to maximize the likelihood [40, 53] or time-consuming attribute learning [64] and adaptive attention [42]. In principle, our idea of using weighted training and reference can also be applied to the frameworks, such as reinforcement learning and attribute learning, which will be one of our future works. We want to emphasize that our method performs the best when compared with those published papers [59, 63, 65] adopting the structure of CNN-RNN. In addition, compared with our previous conference version (*THU-MIG** (*ours*)) [4], the enhanced algorithm achieves superior performances on almost all metrics.

To verify that the proposed approach can significantly improve the image captioning model, we carry out the *T* test experiments on both MS COCO and Flickr30k datasets. We choose the NIC as the control group, because our approach focuses on optimizing the training objective function and generation process, sharing the same architecture as Google NIC [59]. The results are shown in Table 6. We can see that, the *p* values of all metrics on both datasets are all smaller than 5%, which demonstrates the significant improvement of the proposed approach.

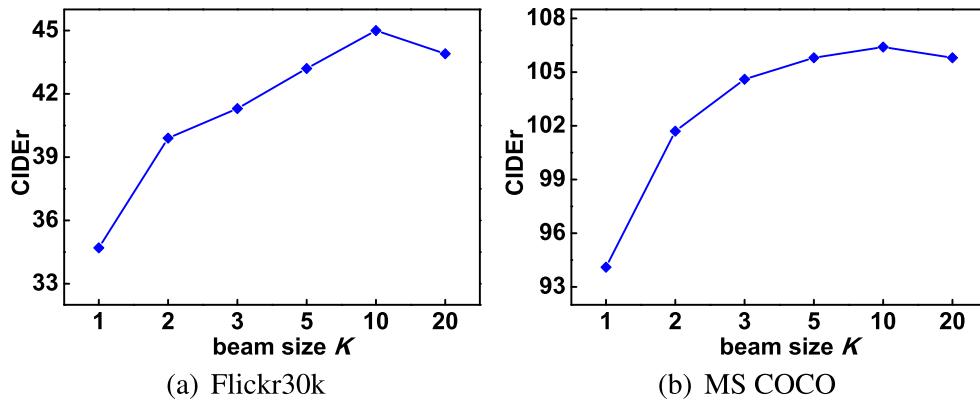


Fig. 11 The influence of beam search size K in the R-LSTM model

Table 3 Performance (%) of the proposed model compared with several state-of-the-art methods on Flickr30k dataset

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
Google NIC [59]	66.3	42.3	27.7	18.3	-	-	-
m-RNN [44]	60.0	41.0	28.0	19.0	-	-	-
LRCN [12]	58.7	39.1	25.1	16.5	-	-	-
Toronto [61]	66.9	43.9	29.6	19.9	18.5	-	-
ATT [65]	64.7	46.0	32.4	23.0	18.9	-	-
SCA-CNN [3]	66.2	46.8	32.5	22.3	19.5	-	-
R-LSTM (ours)	67.7	48.0	32.6	22.1	19.5	45.7	45.0

Table 4 Performance (%) of the proposed model compared with several state-of-the-art methods on MS COCO dataset

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
Google NIC [59]	66.6	45.1	30.4	20.3	-	-	-
m-RNN [44]	67.0	49.0	35.0	25.0	-	-	-
LRCN [12]	66.9	48.9	34.9	24.9	-	-	-
Toronto [61]	71.8	50.4	35.7	25.0	23.0	-	-
ATT [65]	70.9	53.7	40.2	30.4	24.3	-	-
USC [21]	69.7	51.9	38.1	28.2	23.5	50.9	83.8
SCA-CNN [3]	71.9	54.8	41.1	31.1	25.0	-	-
GLA-BEAM3 [34]	72.5	55.6	41.7	31.2	24.9	53.3	96.4
R-LSTM (ours)	76.5	60.3	45.8	34.4	26.4	55.7	106.4

Table 5 Evaluation results (%) of the latest captioning methods on dataset c5 and c40 on the online MS COCO server (<http://mscoco.org/dataset/#captions-leaderboard>) by December 8, 2016

	BLEU-1		BLEU-2		BLEU-3		BLEU-4		METEOR		ROUGE-L		CIDEr	
	c5	c40	c5	c40	c5	c40								
Watson multimodal [53]	77.3	92.9	60.9	85.6	46.1	75.1	34.4	63.6	26.8	35.3	55.9	70.4	112.3	114.6
MSM@MSRA [64]	75.1	92.6	58.8	85.1	44.9	75.1	34.3	64.6	26.6	36.1	55.2	70.9	104.9	105.3
G-RMI(PG-SPIDER-TAG) [40]	75.1	91.8	59.1	84.1	44.5	73.8	33.1	62.4	25.5	34.0	55.1	69.5	104.2	103.2
MetaMind/VT_GT [42]	74.8	92.0	58.4	84.5	44.4	74.4	33.6	63.7	26.4	35.9	55.0	70.5	104.2	105.9
ATT-IMG (MSM@MSRA)	75.2	92.6	59.0	85.2	44.9	75.2	34.0	64.5	26.2	35.6	55.1	70.7	102.3	103.6
G-RMI (PG-BCMR) [40]	75.4	91.6	59.1	84.2	44.5	73.8	33.2	62.4	25.7	33.9	55.0	69.4	101.3	107.1
DONOT_FAIL AGAIN	73.4	91.2	56.4	82.9	42.5	72.4	32.0	61.2	26.2	35.5	54.2	69.8	101.0	102.6
DLTC@MSR	74.0	91.7	57.5	83.9	43.6	73.9	33.1	63.1	25.7	34.8	54.3	69.6	100.3	101.3
THU_MIG (Ours)	76.5	91.7	59.3	83.9	44.2	73.1	32.3	61.4	25.3	33.6	54.2	68.5	99.0	101.3
Postech.CV	74.3	91.5	57.5	83.2	43.1	72.2	32.1	60.7	25.5	34.1	53.9	68.6	98.7	100.1
Feng	74.3	91.7	57.8	84.0	43.4	73.5	32.3	62.1	25.5	34.3	54.0	69.1	98.6	100.2
THU_MIG * (Ours) [4]	75.1	91.3	58.3	83.3	43.6	72.7	32.3	61.6	25.1	33.6	54.1	68.8	96.9	98.8
Reviewnet [63]	72.0	90.0	55.0	81.2	41.4	70.5	31.3	59.7	25.6	34.7	53.3	68.6	96.5	96.9
ATT [65]	73.1	90.0	56.5	81.5	42.4	70.9	31.6	59.9	25.0	33.5	53.5	68.2	94.3	95.8
Google NIC [59]	71.3	89.5	54.2	80.2	40.7	69.4	30.9	58.7	25.4	34.6	53.0	68.2	94.3	94.6

Table 6 Results of *T* test (*p* value) compared with Google NIC [59] on MS COCO and Flickr30k datasets

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
MS COCO	9.72e-6	1.99e-6	3.73e-6	2.32e-6	2.82e-5	2.54e-5	3.92e-7
Flickr30k	8.42e-6	2.69e-6	2.89e-6	6.53e-5	1.43e-5	9.80e-6	2.66e-5

Case Study

Some examples of the generated sentences are illustrated in Fig. 9. The captions in red show how the proposed R-LSTM improves the generation quality as compared to Google NIC [59]: misrecognition is fixed in image 3 (bench->floor, laptop->suitcase), image 6 (kitchen->grill), image 8 (bench->skateboard), image 9 (cattle->sheep); more semantic details are given in image 1 (next to a forest), image 7 (rice and vegetables), image 10 (next to), and image 12 (blanket and a stuffed animal); better match the habit of human cognition in image 2 (sitting on a table vs. filled with lots of), image 4 (sitting in the back seat vs. looking out of the window), image 5 (with a toothbrush in his mouth vs. brushing his teeth with a toothbrush), and image 11 (when holding a nintendo wii game controller, the people are actually playing a video game).

Conclusion

In this paper, we have presented a reference-based LSTM model, where the central idea is to use the training images as references to improve the quality of generated captions. In the training phase, the words are weighted in terms of their relevance to the image, including the overall occurrences, part of speech and corresponding synonyms, which drives the model to focus on the key information of the captions. In the generation phase, we proposed a novel evaluation function by combining the likelihood with the consensus score, which could fix misrecognition and make the generated sentences more natural sounding. Extensive experiments conducted on the MS COCO and Flickr30k datasets corroborated the superiority of the proposed R-LSTM over the state-of-the-art approaches for image captioning. In further studies, we plan to incorporate the attention mechanisms [21, 34, 61, 65] into the reference model and try other weighting strategies. How to generate stylized image captions with emotion [67, 68] and sentiment [45] and extend it to personalized settings [69] is also worth studying. In addition, combining the reference information with reinforcement learning [53] may further improve the image captioning performance.

Compliance with Ethical Standards

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

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

1. Alayrac JB, Bojanowski P, Agrawal N, Sivic J, Laptev I, Lacoste-Julien S. Unsupervised learning from narrated instruction videos. In: IEEE Conference on computer vision and pattern recognition, pp 4575–4583. 2016.
2. Borji A, Itti L. State-of-the-art in visual attention modeling. IEEE Trans Pattern Anal Mach Intell. 2013;35(1):185–207.
3. Chen L, Zhang H, Xiao J, Nie L, Shao J, Chua TS. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In: IEEE Conference on computer vision and pattern recognition. 2017.
4. Chen M, Ding G, Zhao S, Chen H, Liu Q, Han J. Reference based LSTM for image captioning. In: AAAI Conference on artificial intelligence, pp 3981–3987. 2017.
5. Chen X, Fang H, Lin TY, Vedantam R, Gupta S, Dollár P, Zitnick CL. Microsoft COCO captions: data collection and evaluation server. arXiv:[1504.00325](https://arxiv.org/abs/1504.00325). 2015.
6. Cho K, Van Merriënboer B, Gülcühre Ç, Bahdanau D, Bougares F, Schwenk H, Bengio Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Conference on empirical methods on natural language processing, pp 1724–1734. 2014.
7. Deng J, Dong W, Socher R, Li L, Li K, Fei-Fei L. Imagenet: a large-scale hierarchical image database. In: IEEE Conference on computer vision and pattern recognition, pp 248–255. 2009.
8. Devlin J, Cheng H, Fang H, Gupta S, Deng L, He X, Zweig G, Mitchell M. Language models for image captioning: the quirks and what works. In: Annual meeting of the association for computational linguistics, pp 100–105. 2015.
9. Devlin J, Gupta S, Girshick R, Mitchell M, Zitnick CL. Exploring nearest neighbor approaches for image captioning. arXiv:[1505.04467](https://arxiv.org/abs/1505.04467). 2015.
10. Ding G, Guo Y, Zhou J, Gao Y. Large-scale cross-modality search via collective matrix factorization hashing. IEEE Trans Image Process. 2016;25(11):5427–40.
11. Dodds A. Rehabilitating blind and visually impaired people: a psychological approach. Berlin: Springer; 2013.

12. Donahue J, Anne Hendricks L, Guadarrama S, Rohrbach M, Venugopalan S, Saenko K, Darrell T. Long-term recurrent convolutional networks for visual recognition and description. In: IEEE Conference on computer vision and pattern recognition, pp 2625–2634. 2015.
13. Elliott D, Keller F. Image description using visual dependency representations. In: Conference on empirical methods on natural language processing, pp 1292–1302. 2013.
14. Fang H, Gupta S, Iandola F, Srivastava RK, Deng L, Dollár P, Gao J, He X, Mitchell M, Platt JC, et al. From captions to visual concepts and back. In: IEEE Conference on computer vision and pattern recognition, pp 1473–1482. 2015.
15. Farhadi A, Hejrati M, Sadeghi MA, Young P, Rashtchian C, Hockenmaier J, Forsyth D. Every picture tells a story: generating sentences from images. In: European conference on computer vision, pp 15–29. 2010.
16. Gong Y, Wang L, Hodosh M, Hockenmaier J, Lazebnik S. Improving image-sentence embeddings using large weakly annotated photo collections. In: European conference on computer vision, pp 529–545. 2014.
17. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: IEEE conference on computer vision and pattern recognition, pp 770–778. 2016.
18. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput.* 1997;9(8):1735–80.
19. Hodosh M, Young P, Hockenmaier J. Framing image description as a ranking task: data, models and evaluation metrics. *J Artif Intell Res.* 2013;47:853–99.
20. Jia X, Gavves E, Fernando B, Tuytelaars T. Guiding the long-short term memory model for image caption generation. In: IEEE international conference on computer vision, pp 2407–2415. 2015.
21. Jin J, Fu K, Cui R, Sha F, Zhang C. Aligning where to see and what to tell: image caption with region-based attention and scene factorization. arXiv:1506.06272. 2015.
22. Johnson J, Karpathy A, Fei-Fei L. Densecap: Fully convolutional localization networks for dense captioning. In: IEEE Conference on computer vision and pattern recognition, pp 4565–4574. 2016.
23. Karpathy A, Li FF. Deep visual-semantic alignments for generating image descriptions. In: IEEE conference on computer vision and pattern recognition, pp 3128–3137. 2015.
24. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp 1097–1105. 2012.
25. Kulkarni G, Premraj V, Dhar S, Li S, Choi Y, Berg A, Berg T. Baby talk: understanding and generating simple image descriptions. In: IEEE conference on computer vision and pattern recognition, pp 1601–1608. 2011.
26. Kuznetsova P, Ordonez V, Berg A, Berg T, Choi Y. Collective generation of natural image descriptions. In: Annual meeting of the association for computational linguistics, pp 359–368. 2012.
27. Kuznetsova P, Ordonez V, Berg T, Choi Y. Treetalk: composition and compression of trees for image descriptions. *Trans Assoc Comput Linguist.* 2014;2(10):351–62.
28. Lan X, Ma A, Yuen PC, Chellappa R. Joint sparse representation and robust feature-level fusion for multi-cue visual tracking. *IEEE Trans Image Process.* 2015;24(12):5826.
29. Lan X, Ma AJ, Yuen PC. Multi-cue visual tracking using robust feature-level fusion based on joint sparse representation. In: Computer vision and pattern recognition, pp 1194–1201. 2014.
30. Lan X, Yuen PC, Chellappa R. Robust mil-based feature template learning for object tracking. 2017.
31. Lan X, Zhang S, Yuen PC. Robust joint discriminative feature learning for visual tracking. In: International joint conference on artificial intelligence, pp 3403–3410. 2016.
32. Lan X, Zhang S, Yuen PC, Chellappa R. Learning common and feature-specific patterns: a novel multiple-sparse-representation-based tracker. *IEEE Trans Image Process.* 2018;27(4):2022–37.
33. Li J, Zhang Z, He H. Hierarchical convolutional neural networks for EEG-based emotion recognition. *Cogn Comput.* 2018;10(2):368–80.
34. Li L, Tang S, Deng L, Zhang YZ, Qi T. Image caption with global-local attention. In: AAAI conference on artificial intelligence, pp 4133–4139. 2017.
35. Li S, Kulkarni G, Berg T, Berg A, Choi Y. Composing simple image descriptions using web-scale n-grams. In: The SIGNLL conference on computational natural language learning, pp 220–228. 2011.
36. Li Y, Pan Q, Yang T, Wang S, Tang J, Cambria E. Learning word representations for sentiment analysis. *Cogn Comput.* 2017;9(6):843–51.
37. Lin Z, Ding G, Hu M, Lin Y, Ge SS. Image tag completion via dual-view linear sparse reconstructions. *Comput Vis Image Underst.* 2014;124:42–60.
38. Liu D, Hua XS, Yang L, Wang M, Zhang HJ. Tag ranking. In: International world wide web conference, pp 351–360. 2009.
39. Liu L, Yu M, Shao L. Learning short binary codes for large-scale image retrieval. *IEEE Trans Image Process.* 2017;26(3):1289–99.
40. Liu S, Zhu Z, Ye N, Guadarrama S, Murphy K. Optimization of image description metrics using policy gradient methods. arXiv:1612.00370. 2016.
41. Liu Y, Vong C, Wong P. Extreme learning machine for huge hypotheses re-ranking in statistical machine translation. *Cogn Comput.* 2017;9(2):285–94.
42. Lu J, Xiong C, Parikh D, Socher R. Knowing when to look: adaptive attention via a visual sentinel for image captioning. arXiv:1612.01887. 2016.
43. Mao J, Xu W, Yang Y, Wang J, Yuille AL. Explain images with multimodal recurrent neural networks. arXiv:1410.1090. 2014.
44. Mao J, Xu W, Yang Y, Wang J, Yuille AL. Deep captioning with multimodal recurrent neural networks (m-rnn). In: International conference on learning representations. 2015.
45. Mathews AP, Xie L, He X. Senticap: Generating image descriptions with sentiments. In: AAAI conference on artificial intelligence, pp 3574–3580. 2016.
46. Mitchell M, Han X, Dodge J, Mensch A, Goyal A, Berg A, Yamaguchi K, Berg T, Stratos K, Daumé III H. Midge: generating image descriptions from computer vision detections. In: Conference of the European chapter of the association for computational linguistics, pp 747–756. 2012.
47. Mnih V, Heess N, Graves A, et al. Recurrent models of visual attention. In: Advances in neural information processing systems, pp 2204–2212. 2014.
48. Pan Y, Mei T, Yao T, Li H, Rui Y. Jointly modeling embedding and translation to bridge video and language. In: IEEE conference on computer vision and pattern recognition, pp 4594–4602. 2016.
49. Papineni K, Roukos S, Ward T, Zhu WJ. Bleu: a method for automatic evaluation of machine translation. In: Annual meeting of the association for computational linguistics, pp 311–318. 2002.
50. Parzen E. On estimation of a probability density function and mode. *Ann Math Stat.* 1962;33(3):1065–76.
51. Ranzato M, Chopra S, Auli M, Zaremba W. Sequence level training with recurrent neural networks. arXiv:1511.06732. 2015.
52. Reforgiato Recupero D, Presutti V, Consoli S, Gangemi A, Nuzzolese AG. Sentilo: Frame-based sentiment analysis. *Cogn Comput.* 2015;7(2):211–25.
53. Rennie SJ, Marcheret E, Mroueh Y, Ross J, Goel V. Self-critical sequence training for image captioning. arXiv:1612.00563. 2016.
54. Roopnarine J, Johnson JE. Approaches to early childhood education. Upper Saddle River: Merrill/Prentice Hall; 2013.

55. Schwenk H. Continuous space translation models for phrase-based statistical machine translation. In: International conference on computational linguistics, pp 1071–1080. 2012.
56. Spratling MW. A hierarchical predictive coding model of object recognition in natural images. *Cogn Comput*. 2017;9(2):151–67.
57. Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. In: Advances in neural information processing systems, pp 3104–3112. 2014.
58. Taylor JG, Cutsuridis V. Saliency, attention, active visual search, and picture scanning. *Cogn Comput*. 2011;3(1):1–3. <https://doi.org/10.1007/s12559-011-9096-1>.
59. Vinyals O, Toshev A, Bengio S, Erhan D. Show and tell: a neural image caption generator. In: IEEE conference on computer vision and pattern recognition, pp 3156–3164. 2015.
60. Wu R, Yan S, Shan Y, Dang Q, Sun G. Deep image: scaling up image recognition. arXiv:1501.02876.7.8. 2015.
61. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, Zemel R, Bengio Y. Show, attend and tell: neural image caption generation with visual attention. In: International conference on machine learning, pp 2048–2057. 2015.
62. Yang Y, Teo CL, Daumé III H, Aloimonos Y. Corpus-guided sentence generation of natural images. In: Conference on empirical methods on natural language processing, pp 444–454. 2011.
63. Yang Z, Yuan Y, Wu Y, Cohen WW, Salakhutdinov RR. Review networks for caption generation. In: Advances in neural information processing systems, pp 2361–2369. 2016.
64. Yao T, Pan Y, Li Y, Qiu Z, Mei T. Boosting image captioning with attributes. arXiv:1611.01646. 2016.
65. You Q, Jin H, Wang Z, Fang C, Luo J. Image captioning with semantic attention. In: IEEE conference on computer vision and pattern recognition, pp 4651–4659. 2016.
66. Young P, Lai A, Hodosh M, Hockenmaier J. From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions. *Trans Assoc Comput Linguist*. 2014;2:67–78.
67. Zhao S, Gao Y, Ding G, Han J. Approximating discrete probability distribution of image emotions by multi-modal features fusion. In: International joint conference on artificial intelligence. 2017.
68. Zhao S, Gao Y, Jiang X, Yao H, Chua TS, Sun X. Exploring principles-of-art features for image emotion recognition. In: ACM international conference on multimedia, pp 47–56. 2014.
69. Zhao S, Yao H, Gao Y, Ding G, Chua TS. Predicting personalized image emotion perceptions in social networks. *IEEE Transactions on Affective Computing*. 2017.
70. Zhao S, Yao H, Gao Y, Ji R, Ding G. Continuous probability distribution prediction of image emotions via multi-task shared sparse regression. *IEEE Trans Multimed*. 2017;19(3):632–45.
71. Zhao S, Yao H, Jiang X, Sun X. Predicting discrete probability distribution of image emotions. In: IEEE international conference on image processing, pp 2459–2463. 2015.
72. Zheng A, Xu M, Luo B, Zhou Z, Li C. CLASS: collaborative low-rank and sparse separation for moving object detection. *Cogn Comput*. 2017;9(2):180–93.
73. Zhong G, Yan S, Huang K, Cai Y, Dong J. Reducing and stretching deep convolutional activation features for accurate image classification. *Cogn Comput*. 2018;10(1):179–86.