International Journal of Multimedia Information Retrieval (2018) 7:147–155
https://doi.org/10.1007/s13735-018-0158-y

REGULAR PAPER

Improvement of image description using bidirectional LSTM

Vahid Chahkandi[1] · **Mohammad Javad Fadaeieslam**[1] · **Farzin Yaghmaee**[1]

Received: 11 December 2017 / Revised: 2 June 2018 / Accepted: 13 July 2018 / Published online: 19 July 2018

Abstract
As a high-level technique, automatic image description combines linguistic and visual information in order to extract an appropriate caption for an image. In this paper, we have proposed a method based on a recurrent neural network to synthesize descriptions in multimodal space. The innovation of this paper consists in generating sentences with variable length and novel structures. The Bi-LSTM network has been applied to achieve this purpose. This paper utilizes the inner product as common space, which reduces the computational cost and improves results. We have evaluated the performance of the proposed method on benchmark datasets: Flickr8K and Flickr30K. The results demonstrate that Bi-LSTM has better efficiency, as compared to the unidirectional model.

Keywords Image description · Bidirectional LSTM · Multimodal space · Region-based Convolutional Neural Networks (RCNN)

1 Introduction

Image captioning equipment, such as digital cameras and smartphones, provides large collections of image data. It is impossible to manually analyze such a huge number of images and extract semantic information from them. Automatic image description, which is one of the high-level techniques, combines linguistic and visual information [1–3]. In order to attain this goal, image-understanding methods are needed to detect objects and their relation. Moreover, NLP techniques must be used to generate grammatically well-formed sentences. Conventional methods use sentence templates or make use of the caption of the most similar image in the dataset. The weaknesses of these methods are as follows: low efficiency to generate sentences with variable length and disability to produce novel sentences. Recently, deep learning methods have been used to perform captioning in multimodal space [4, 5]. This paper proposes a method based on deep networks. Our innovation in this paper is the replacement of BRNN with Bi-LSTM, in order to overcome gradient vanishing and gradient exploding. Furthermore, we have reduced the number of parameters in comparison with the method presented by Wang et al. [6]. There are two Bi-LSTMs in [6]. One of those is utilized as a common semantic space. However, in our proposed method, we use the inner product as common space, which reduces the computational cost and improves the results.

The proposed method depicted in this paper consists of three main parts. The first part consists of a deep neural network, which detects objects in the Region of Interest (RoI) of images. The second one contains bidirectional LSTM (Bi-LSTM) to decode the sentences for the detected objects in the image. The Bi-LSTM enables us to consider sentences in two directions (the forward pass and the backward pass) and predict the appropriate relevant sentence according to the visual data. In unidirectional LSTM, next word prediction w_t, which maximizes $\log P(w_t | I, w_{1:t-1})$, is related to the words before the time t. Meanwhile, prediction in Bi-LSTM depends on the words which maximize both $\log P(w_t | I, w_{1:t-1})$ and $\log P(w_t | I, w_{t+1:T})$ before and after time t, respectively. The third part is a common semantic space, formulating an image-sentence score to define the score between one image and its corresponding sentence. We have evaluated the performance of the proposed method on benchmark datasets: Flickr8K and Flickr30K. The results demonstrate that Bi-LSTM has a better efficiency, as compared to the unidirectional model.

The remainder of this paper is organized as follows: Sect. 2 deals with the related works. Section 3 clarifies the proposed method with respect to unidirectional LSTM and

✉ Mohammad Javad Fadaeieslam
fadaei@semnan.ac.ir

[1] Department of Electrical and Computer Engineering, Semnan University, Semnan, Iran

Bi-LSTM, object detection by Region Convolutional Neural Network (RCNN), and model architecture. In Sect. 4, the training process of the neural network is explained. Experimental results are shown in Sect. 5. The conclusion is offered in the last section.

2 Related works

Moreover, image captioning and image description are two different concepts—in most papers they are used interchangeably. While image captioning includes non-visible image information, image description explains clearly visible information [7]. A comprehensive review of image description approaches includes three main categories: direct generation models from visual input, retrieval models from visual space, and retrieval models from multimodal space [7]. In the following subsections, each of these categories is explained in brief.

2.1 Direct generation models from visual input

These generation models have two main steps. In the first step, the objects and the relationship among them are detected by image processing techniques. The second step transforms the output of the previous step into words and phrases. These words are then combined by using techniques like the sentence template technique, the *n*-gram-based language model, and grammar rules to produce a sentence [8–10]. Elliot and Keller [11] presented Visual Dependency Representation (VDR) to illustrate the spatial relationship between the objects in dependency graph form. *n*-Gram language models are used in the method introduced in [12, 13]. A subset of Wikipedia is used to train the models. Another group of approaches in this category utilizes sentence templates. These templates are defined manually and have some blank spaces, which must be filled with some labels. Yang et al. [14] proposed a method in which they filled a sentence structure with a quadruplet (object, verb, preposition, and scene type). They used HMM to determine the most likely quadruplet, which would make up the main structure of the output sentence.

2.2 Retrieval models in visual space

These methods generate an image description by retrieving similar images. In the first step, some features are extracted from the input image, and in the second step, the images most similar to the input image are selected from the training set. Finally, the phrase fragments are combined, based on the given rules [15–17].

One of the first approaches using retrieval in visual space is IM2TEXT, which was proposed by Ordonez et al. [15]. They employ GIST and Tiny Image features to describe an input image. After detecting the types of objects in the description of a visually similar image, they reconsider the input image with the detectors and classifiers specific to these objects. The image features which are applied in [16] by Kuznetsova et al. are similar to IM2TEXT. They use integer linear programming method to generate final descriptors from extracted phrases. Gupta et al. [18] proposed another phrase-based model. They use RGB and HSV color histogram, Gabor, Haar, GIST, and SIFT descriptors to extract image features.

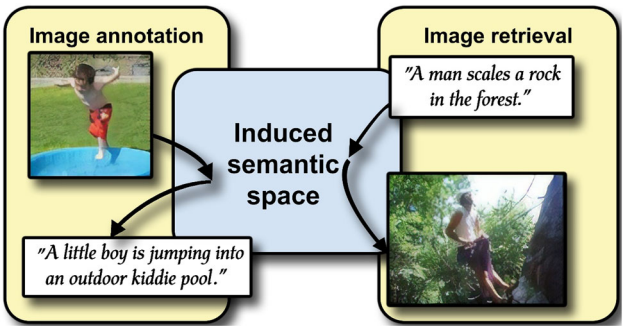


Fig. 1 Retrieval image description in multimodal space proposed by Hodosh et al. (*Source*: http://nlp.cs.illinois.edu/HockenmaierGroup/Framing_Image_Description/KCCA.html)

2.3 Retrieval models in multimodal space

Methods in this group have been developed recently. They consist of a learning process in a multimodal space. In 2010, in order to learn common semantic space, Farhadi et al. [19] used a triplet containing object, action, and scene. In their method, the representation was thus restricted to some predefined discrete slot fillers, which were set in a training step. Hodosh et al. [20] are among the pioneers in this category. They mapped images and sentences in the common space, as seen in Fig. 1. Their system can get each of the images or descriptions as an input and produce cross-modal retrieval. In comparison with [19], Hodosh et al. [20] used a kernelized Canonical Correlation Analysis (kCCA) to learn the common semantic space.

In 2014, Karpathy et al. [21] improved the earlier multimodal methods. In their paper, fragments of the image and the sentence were mapped into a joint space instead of mapping the whole image and the whole sentence together.

Socher et al. [21] proposed a method in which linguistic and visual information are first trained in their respective single modalities and then mapped into a joint space. The Dependency Tree Recursive Neural Network (DT-RNN) learns vector representations for sentences. DT-RNN is more invariant and robust to syntactic structure or changes in word order. To extract visual information, they apply a deep neural network consisting of nine layers. It takes pixels of

the query image as input. They use a max-margin objective function as a joint space.

Evaluation is a problem posed in description generation. Also either retrieval or ranking operations can be used to handle it, recent methods such as [22] apply both of them. But retrieval and ranking methods are limited by the available datasets with descriptions.

3 Proposed method

Our proposed method is an improved version of [22], developed by Karpathy and Fei-Fei. They present a visual-semantic model based on deep learning to establish a relationship between regions in the image and fragments of the sentence. In their method, they use the convolutional neural network to extract RoI and the bidirectional recurrent neural network (BRNN) to align sentences. In the last step, an objective structure is used to align the two modalities of the networks.

Replacement of BRNN with bidirectional LSTM (Bi-LSTM) is another contribution of this paper. There are some differences between BRNN and Bi-LSTM. Input gates in Bi-LSTM pass the words through the memory cells. If the words are inappropriate for the given image, they are blocked and the memory cells remember the previous status. The advantages of LSTM against RNN are the avoidance of dropout and gradient explosion. When gradient vanishes, a learning process may mislead to a wrong optimum. Gradient explosion may throw the learning algorithm off-track. The LSTM cells have memory gates and forget gates that make a more persistent memory, as compared to RNN. Therefore, they decrease vanishing and exploding gradient, and finally this would be effective in long-term dependencies. These advantages make our method more useful than RNN to predict words in sentences of different lengths. Therefore, our method attains better results, as compared to the previous ones.

In generating a sentence in unidirectional structures, such as RNN and LSTM, predicting the next word w_t is done by maximizing log $P(w_t|I, w_{1:t-1})$, in which I is visual context and $w_{1:t-1}$ represents previous words. Unidirectional structures use only past context, but bidirectional models add future context $w_{t+1:T}$ that can be maximized by log $P(w_t|I, w_{t+1:T})$. This difference between unidirectional and bidirectional structures enables the proposed method to exploit future and past word dependencies for better prediction.

The method proposed in this paper consists of three main modules: one RCNN to detect objects in the input image, Bi-LSTM to represent words in the sentence, and a common semantic space to calculate the score for each image sentence. The details of these modules are explained in the following subsections.

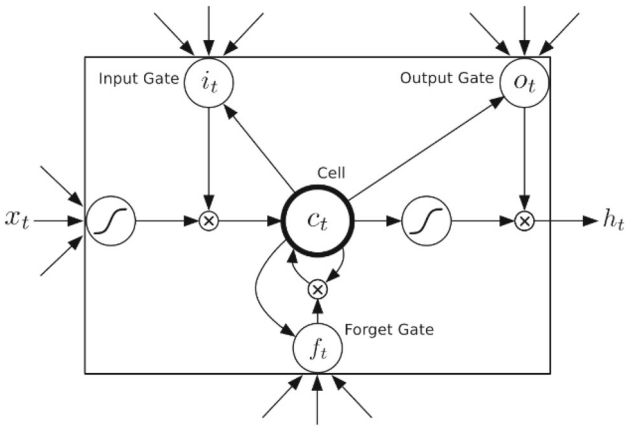


Fig. 2 LSTM cell

3.1 Object detection by using RCNN

We use the method presented in [23] in order to detect objects. Each object in the image is extracted by RCNN. Karpathy and her co-author use the top 19 detected bounding boxes in the whole image and compute their features as follows:

$$v = W_{\mathrm{m}}\left[\mathrm{CNN}_{\theta_{\mathrm{c}}}(I_{\mathrm{b}})\right] + b_{\mathrm{m}} \tag{1}$$

where $\mathrm{CNN}(I_{\mathrm{b}})$ transforms all pixels in the bounding box I_{b} and transfers them into a fully connected layer before the classifier. θ_c has about 60 million CNN parameters. The matrix W_{m} is used to extract SVM scores from the visual features of each bounding. Every image is converted to a set of vectors v_{i}.

3.2 Long short-term memory

In this paper, we used LSTM cells, which are a specific type of RNN. As can be seen in Fig. 2, there are four components in an LSTM cell: a memory cell and three gates (input, forget, and output).

The LSTM cell includes three different input sources: current input x_t, hidden state of all previous states h_{t-1}, and the previous cell-memory state c_{t-1} [6]. The precise form of LSTM at time step t is as follows:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \tag{2}$$
$$f_t = \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + b_f\right) \tag{3}$$
$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \tag{4}$$
$$g_t = \phi(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \tag{5}$$
$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \tag{6}$$
$$h_t = o_t \odot \phi(c_t) \tag{7}$$

where b is a bias vector, W indicates a trained weights matrix, ϕ represents the hyperbolic tangent, and σ is a sigmoid acti-

vation function. $\odot$ shows the product function [6]. The LSTM hidden output $h_t = \{h_{tk}\}_{k=0}^{K}$, $h_t \in R^K$ will be applied to predict the next word using W_s and b_s parameters:

$$\mathcal{F}(p_{ti}; W_s, b_s) = \frac{\exp(W_s h_{ti} + b_s)}{\sum_{j=1}^{K} \exp\left(W_s h_{tj} + b_s\right)} \tag{8}$$

where p_{ti} is the probability distribution of the predicted word.

3.3 Bidirectional LSTM

We use bidirectional LSTM, which is implemented with two separate unidirectional LSTMs to exploit both the past and the future information (similar to [6]). One computes the forward hidden sequence $\vec{h}$, and the other computes the backward hidden sequence $\overleftarrow{h}$. The forward LSTM starts at $t = 1$, and backward starts at time $t = T$.

Similar to Sect. 3.1, LSTM converts words of a sentence into vectors. Instead of using recurrent neural networks (such as RNN, Bi-LSTM), words can be mapped directly into common space. However, this approach misses the information obtained from the order of words. The maximum length of sentences is another important problem, which is handled by Bi-LSTM.

Wang et al. [6] introduced a deep Bi-LSTM, named Bi-F-LSTM. They added multilayer perceptron as an intermediate transition between LSTM layers. This deep LSTM has fewer parameters, as compared to the other deep LSTM. Both Bi-LSTM and Bi-F-LSTM are applied in our proposed method.

3.4 Architecture model

The overall structure of the proposed model is shown in Fig. 3. As mentioned, it has three main modules: RCNN, Bi-LSTM, and common space. Object detection is done by RCNN and Bi-LSTM encodes corresponding sentences. In the third module, an appropriate sentence is obtained for each input image using the inner product of v and p, which are the visual and textual vectors, respectively. Equation (9) shows the inner product of the kth image and the lth sentence:

$$S_{kl} = \sum_{t \in g_l} \sum_{i \in g_k} \max\left(0, v_i^{\mathrm{T}} p_t\right) \tag{9}$$

where g_k is the set of bounding boxes of image k and g_l is the set of parts of sentence l.

RCNN extracts objects from an input image and localizes each object using a bounding box. It then suggests a label with an SVM score for each detected object. The top 19 detected objects are selected for the next step [22]. In Eq. (9), v indicates these SVM scores.

In a training step, the Bi-LSTM module receives each input word as a vector x_i. The length of this vector is equal

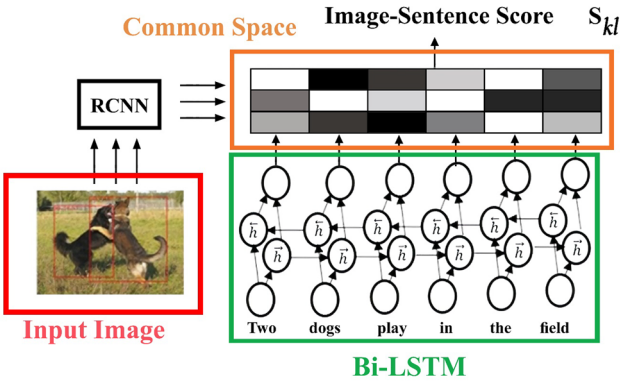


Fig. 3 The structure of the proposed model

to the vocabulary size. All entries of this vector are zero except an entry related to the input word, which is one. In this module, vector x_i and the learned parameter matrix $W \in \mathbb{R}^{d \times K}$ are multiplied together in order to prepare requisites for the next time. The parameter d is the dimension of the Bi-LSTM output, and K represents the size of the vocabulary.

In a test step, Bi-LSTM is used to predict the word w_t, using the image feature vector v, $p(w_t|w_{1:t-1}, v)$ in forward order and $p(w_t|w_{t-1:T}, v)$ in backward order:

$$P(w_{1:T}|v) = \max\left(\sum_{t=1}^{T} \vec{p}(w_t|v), \sum_{t=1}^{T} \overleftarrow{p}(w_t|v)\right) \tag{10}$$

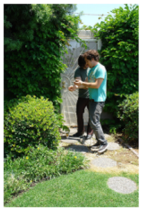
$$\vec{p}(w_t|v) = \prod_{t=1}^{T} p(w_t|w_1, w_2, \ldots, w_{t-1}, v) \tag{11}$$

$$\overleftarrow{p}(w_t|v) = \prod_{t=1}^{T} p(w_t|w_{t+1}, w_{t+2}, \ldots, w_T, v) \tag{12}$$

The Bi-LSTM module of the proposed method and its training parameter are similar to [6]. Instead of the entire image (I), it receives SVM scores of the bounding box (v). To optimize the model, we use Stochastic Gradient Descent with batch sizes, each consisting of 100 image-sentence pairs. The momentum is set to 0.9. The cross-validation is used to obtain learning rate and weight decay.

Wang et al. [6] utilized Bi-LSTM to embed visual and textual vectors in joint space. In our proposed method, the inner product is used as common space. Therefore, the number of parameters is reduced in comparison with the method presented by Wang et al.

We have used the inner product as a similarity metric. There are other methods for similarity measurement, such as Euclidean distance and Pearson correlation coefficient. However, we selected inner product because of its simplicity and low computational cost.



1. Two young guys with shaggy hair look at their hands while hanging out in the yard.
2. Two men in green shirts are standing in a yard.
3. A man in a blue shirt standing in a garden.
4. Two friends enjoy time spent together.
5. Two young, while males are outside near many bushes.

Fig. 4 One of the images from the Flickr30K dataset and its five annotated sentences

4 Experimental results

4.1 Datasets and evaluation metrics

Flickr8K and Flickr30K datasets are utilized in our experiments. They contain 8091 and 31,784 images, respectively. All the images are annotated with five descriptive sentences.

Figure 4 shows one of the images from the Flickr30K dataset with five descriptive sentences. The proposed model is evaluated in the process of generating a descriptive sentence and an image sentence. BLEU-N is applied to generate descriptive sentences, and the BLEU-N ($N = 1, 2, 3, 4$) scores are calculated as follows:

$$B_N = \min\left(1, e^{1-\frac{r}{c}}\right) \cdot e^{\frac{1}{N}\sum_{n=1}^{N} \log p_n} \tag{13}$$

where c is the length of the generated sentence, r is the length of the reference sentence, and p_n represents the modified n-gram precisions. In image-sentence retrieval, the results are based on S_{kl} score, and we adapt R@K ($K = 1, 5, 10$) and *Medr* as evaluation metrics. R@K is the recall rate R at top K candidates, and *Medr* is the median rank of the first retrieved ground-truth image and sentence [6].

Table 1 Performance comparison on BLEU-N

Models	Flickr8K				Flickr30K			
	B-1	B-2	B-3	B-4	B-1	B-2	B-3	B-4
Deep visual-semantic [22]	57.9	38.3	24.5	16.0	57.3	36.9	24.0	15.7
m-RNN (AlexNet) [10]	56.5	38.6	25.6	17.0	54.0	36.0	23.0	15.0
m-RNN (VGGNet) [10]	–	–	–	–	60.0	41.0	28.0	19.0
Hard-attention [24]	**67**	45.7	31.4	21.3	**66.9**	43.9	29.6	19.9
NIC [25]	63	41	27.2	–	66.3	42.3	27.7	18.3
Deep Bi-LSTM [6]	65.5	46.8	32	21.5	62.1	42.6	28.1	19.3
Our model (with Bi-LSTM)	66.7	**47.0**	**34.1**	**21.7**	63.1	**44.0**	**29.7**	**20.0**
Our model (with Bi-F-LSTM)	64.2	44.0	30.6	20.1	59.9	41.2	27.3	18.1

The values in bold are the best results

Table 2 Performance comparison on R@K (the higher the better) and *Medr* (the lower the better)

Datasets	Models	Image to sentence				Sentence to image			
		R@1	R@5	R@10	*Medr*	R@1	R@5	R@10	*Medr*
Flickr8K	Deep visual-semantic [22]	16.5	40.6	54.2	7.6	11.8	32.1	44.7	12.4
	m-RNN (AlexNet) [10]	14.5	37.2	48.5	11.0	11.5	31.0	42.4	15
	NIC [25]	20	–	60	6	19	–	**64**	5
	Mind's Eye [26]	17.2	42.5	57.4	7	15.4	40.6	50.1	8
	Deep Bi-LSTM [6]	29.3	58.2	69.6	3	19.7	47	60.6	5
	Our model (with Bi-LSTM)	**30.2**	**59.5**	**70.1**	**2.5**	**20.7**	**48.2**	61	**4.6**
	Our model (with Bi-F-LSTM)	21.1	44.9	57.5	6	15.7	38.6	51	9
Flickr30K	Deep visual-semantic [22]	22.2	48.2	61.4	4.8	15.2	37.7	50.5	9.2
	m-RNN (AlexNet) [10]	18.4	40.2	50.9	10	12.6	31.2	41.5	16
	NIC [25]	17	–	56	7	17	–	57	8
	Mind's Eye [26]	18.5	45.7	58.1	7	16.6	42.5	**58.9**	8
	Deep Bi-LSTM [6]	28.1	53.1	64.2	4	19.6	43.8	55.8	7
	Our model (with Bi-LSTM)	**29.2**	**54.0**	**64.9**	**3.8**	**20.8**	**44.5**	56.7	**6.7**
	Our model (with Bi-F-LSTM)	21.1	46.0	55.9	7	16.4	37.1	48.2	11

The values in bold are the best results

Fig. 5 Examples of image descriptions obtained from the proposed method

4.2 Results on generated image description

Table 1 shows the comparison results in terms of BLEU-N. A better descriptive sentence has a higher BLEU score. Bi-LSTM and fully connected Bi-LSTM (Bi-F-LSTM) are both applied in the proposed method. The proposed method attains better results, as compared with the existing methods.

4.3 Results on image-sentence retrieval

The suitability of assigning a sentence to an image is called image-sentence retrieval. The results are presented by image to sentence and sentence to image alignments. Table 2 shows the results on two benchmark datasets. As can be seen, the

a man in white shirt is sitting at a table with a glass of wine

a man in a blue shirt is surfing on a wave

a man in a blue shirt is sitting on a stool

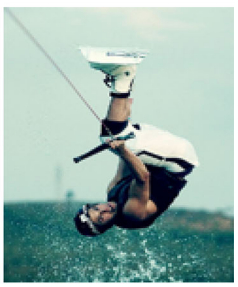
a young girl in a wetsuit is paddling a boat in the water

a man in a black shirt is playing a guitar

A group of people are standing in front of a building

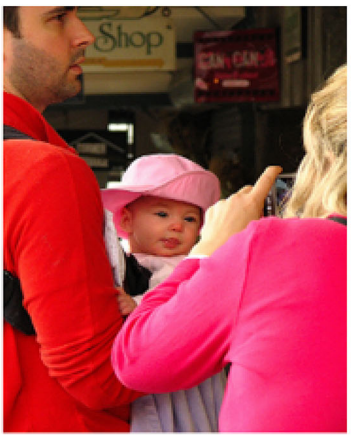
a group of people are sitting in front of a store

a dog is running through a field



a man in a wetsuit is surfing a wave

a woman in a white shirt is sitting at a table with a plate of food in her hand



a young girl in a red shirt is holding a red and white umbrella

a young girl is eating a meal

Fig. 6 Additional examples of image descriptions obtained from the proposed method

method proposed in this paper shows a very competitive performance on evaluated datasets.

4.4 Qualitative results

Using Bi-LSTM, the proposed model generates sensible descriptions for images. Some results of the proposed method are shown in Figs. 5 and 6. Qualitative assessment of the generated captions can be considered from various aspects. From the object detection point of view, the experimental results demonstrate a good performance of the proposed method. There are some mistakes in detecting an object feature (e.g., the color of a person's shirt). In some cases, despite the misidentification of object features, the relationship between objects is detected properly. There are some descriptions in which the relation between objects is recognized well enough, but the objects and their features are not identified correctly (for example, "a young girl in a pink shirt is holding a baby" in Fig. 5).

5 Conclusion

In this paper, we have proposed a method based on deep learning in order to synthesize description in multimodal space. The main innovation of this method is that of generating sentences of variable lengths and novel structures. The Bi-LSTM network is applied to attain this purpose. The proposed method consists of three main modules: The first module consists in a deep neural network, which detects objects in the RoI of images. The second one contains Bi-LSTM to decode the sentences for the detected objects in the image. Bi-LSTM has a mechanism to deal with the gradient vanishing/exploding problem. Although LSTM uses only past context, bidirectional LSTM uses both past and future contexts. This difference between unidirectional and bidirectional structures enables the proposed method to exploit future and past word dependencies for better prediction. The third module is a common semantic space formulating an image-sentence score to assign it to the image and its corresponding sentence. This paper utilizes the inner product as common space, which reduces the computational cost and improves results. We have evaluated the performance of the proposed method on benchmark datasets: Flickr8K and Flickr30K.

References

1. Coyne B, Sproat R (2001) Wordseye: an automatic text-to-scene conversion system. In: SIGGRAPH'01
2. Das P, Xu C, Doell RF, Corso JJ (2013) A thousand frames in just a few words: lingual description of videos through latent topic and sparse object stitching. In: CVPR
3. Krishnamoorthy N, Malkarnenkar G, Mooney RJ, Saenko K, Guadarrama S (2013) Generating natural-language video descriptions using text-mined knowledge. In: AAAI, vol 1
4. Karpathy A, Joulin A, Li F-F (2014) Deep fragment embeddings for bidirectional image sentence mapping. In: Advances in neural information processing systems
5. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems
6. Wang C, Yang H, Bartz C, Meinel C (2016) Image captioning with deep bidirectional LSTMs. In: Proceedings of the 2016 ACM on multimedia conference. ACM, Oct 2016, pp 988–997
7. Bernardi R, Cakici R, Elliott D, Erdem A, Erdem E, Cinbis NI, Keller F, Muscat A, Plank B (2016) Automatic description generation from images: a survey of models, datasets, and evaluation measures. J Artif Intell Res (JAIR) 55:409–442
8. Mitchell M, Han X, Dodge J, Mensch A, Goyal A, Berg A, Yamaguchi K, Berg T, Stratos K, Daumé H III (2012) Midge: generating image descriptions from computer vision detections. In: Proceedings of the 13th conference of the European chapter of the association for computational linguistics. Association for computational linguistics
9. Kuznetsova P, Ordonez V, Berg TL, Choi Y (2014) TREETALK: composition and compression of trees for image descriptions. In: Conference on empirical methods in natural language processing
10. Mao J, Xu W, Yang Y, Wang J, Huang Z, Yuille A (2015) Deep captioning with multimodal recurrent neural networks (m-RNN). In: International conference on learning representations
11. Elliott D, Keller F (2013) Image description using visual dependency representations. In: Proceedings of the 2013 conference on empirical methods in natural language processing
12. Kulkarni G, Premraj V, Ordonez V, Dhar S, Li S, Choi Y, Berg AC, Berg TL (2013) Babytalk: understanding and generating simple image descriptions. IEEE Trans Pattern Anal Mach Intell 35(12):2891–2903
13. Li S, Kulkarni G, Berg TL, Berg AC, Choi Y (2011) Composing simple image descriptions using web-scale n-grams. In: Proceedings of the fifteenth conference on computational natural language learning. Association for computational linguistics
14. Yang Y, Teo CL, Daumé H III Aloimonos Y (2011) Corpus-guided sentence generation of natural images. In: Proceedings of the conference on empirical methods in natural language processing. Association for computational linguistics
15. Ordonez V, Kulkarni G, Berg TL (2011) Im2text: describing images using 1 million captioned photographs. In: Advances in neural information processing systems
16. Kuznetsova P, Ordonez V, Berg AC, Berg TL, Choi Y (2012) Collective generation of natural image descriptions. In: Proceedings of the 50th annual meeting of the association for computational linguistics: long papers, vol 1. Association for computational linguistics
17. Patterson G, Xu C, Su H, Hays J (2014) The sun attribute database: beyond categories for deeper scene understanding. Int J Comput Vis 108(1–2):59–81
18. Gupta A, Verma Y, Jawahar CV (2012) Choosing linguistics over vision to describe images. In: AAAI
19. Farhadi A, Hejrati M, Sadeghi MA, Young P, Rashtchian C, Hockenmaier J, Forsyth D (2010) Every picture tells a story: generating sentences for images. In: ECCV
20. Hodosh M, Young P, Hockenmaier J (2013) Framing image description as a ranking task: data, models and evaluation metrics. J Artif Intell Res 47:853–899
21. Socher R, Karpathy A, Le QV, Manning CD, Ng AY (2014) Grounded compositional semantics for finding and describing images with sentences. Trans Assoc Comput Linguist 2:207–218

22. Karpathy A, Li F-F (2015) Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition
23. Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 580–587
24. Xu K, Ba J, Kiros R, Courville A, Salakhutdinov R, Zemel R, Bengio Y (2015) Show, attend and tell: neural image caption generation with visual attention. In: ICML
25. Vinyals O, Toshev A, Bengio S, Erhan D (2015) Show and tell: a neural image caption generator. In: CVPR, pp 3156–3164
26. Chen X, Zitnick CL (2015) Mind's eye: a recurrent visual representation for image caption generation. In: CVPR, pp 2422–2431