Comprehensive Comparative Study on Several Image Captioning Techniques Based on Deep Learning Algorithm

Chitrapriya Ningthoujam and Tejbanta S. Chingtham

Abstract Image captioning is evolving as an interesting area of research that involves generating a caption or describing the content in the image automatically. The idea behind image captioning is to make the computer perceive a given image like a human mind leading to automatic description. Image captioning is a challenging task that involves capturing semantically correct information and expressing in a simple sentence. A large number of methods have been proposed in the recent past, and we aim to do a comprehensive survey in the different deep learning algorithms used in image captioning based on the method framework.

Keywords Captioning · Annotation · Image · Deep learning · Encoder · Decoder · Neural image generator · LSTM · Language method

1 Introduction

In recent years, image captioning has received huge attention. It involves observing the contents in an image and then describing it. It has a broad application area with a wide range of scenarios. Areas of research in Natural Language Processing (NLP) and also in Computer Vision (CV) fields are achieving immense advancements; larger datasets have been made available while generating text of images and videos leading to implementation of deep neural network-based methods acquiring more and more accurate results on image captioning. It involves the task of capturing an image, analyzing the video contents, recognizing the most important features of the image, and then generating the textual description based on it. Deep learning algorithms have shown better results in handling many complex and challenges of an image captioning task [1]. The image processing can be categories into three different approaches based on: retrieval, text, and novel. Retrieval-based approach caption an image from a collection of already existing captions [2]. In template based, captions are generated based on the templates which identify a set of visual notions

C. Ningthoujam (✉) · T. S. Chingtham
Department of Computer Science and Engineering, Sikkim Manipal Institute of Technology, Sikkim Manipal University, Majitar, Sikkim, India

H. K. D. Sarma et al. (eds.), *Contemporary Issues in Communication, Cloud and Big Data Analytics*, Lecture Notes in Networks and Systems 281,
https://doi.org/10.1007/978-981-16-4244-9_18

first then connected through the sentence template to compose a sentence used by [3]. Novel based on the other hand, generates captions of an image from both visual spaces as well as from multimodal space.

This paper starts with the discussion of different image captioning methods categories into two different frameworks in Sect. 1.1. Section 1.1.1 discusses the encoder-decoder framework along with five different methods under it. References [4–8] methods are based on encoder-decoder architecture to generate a caption. Similarly, Sect. 1.1.2 discusses the compositional architecture-based image captioning and five other different methods another same. References [9–13] methods are based on the second type of framework where captions are generated by extracting components from relevant captions and later combined for describing the image. In Sect. 1.2 summarizes the various image captioning methods based on deep learning method on two different frameworks.

1.1 Image Captioning Methods

Among the various methods based on deep learning model, this paper has considered the framework used to build a model that can generate a caption or describe a given image trained and tested on some of the benchmark datasets. The architecture considered are: encoder-decoder-based framework and compositional-based framework.

1.1.1 Encoder-Decoder Framework

(a) ***Encoder-Decoder pipeline***: The main idea of this method is to translate a sentence from one language into another language by supplying an input as an image and the output as a sentence illustrated in Fig. 1 [4]. This method has been adopted from the neural translation concept as given by [14].

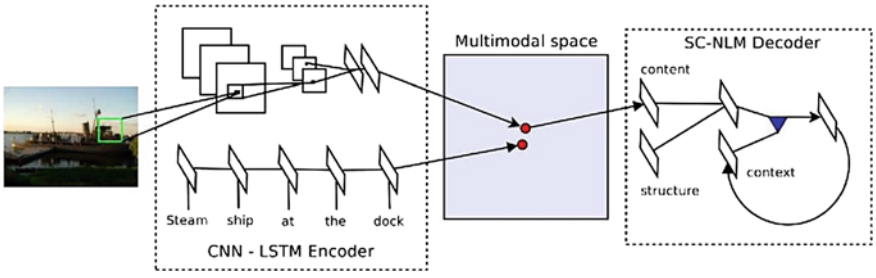

Fig. 1 The encoder-decoder method proposed by [4]

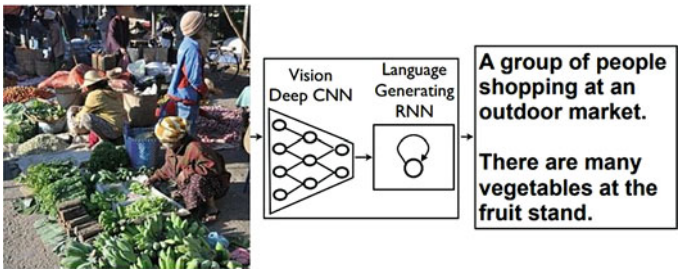

Fig. 2 Neural image caption generator [5]

Working

It contains two stages: ***encoder*** and ***decoder.*** Firstly, the encoder phase makes a combined multimodal space which is used to order the images along with its descriptions. This encoder encodes the sentences by using the idea of machine translation using LSTM model [15]. Features of an image are embedded using a CNN. The encoder tries to minimize the pairwise ranking loss that will help to learn the ranking of images and along with its descriptions. In the second stage, the method uses the multimodal representation so that it can generate novel descriptions. The decoder part uses a new type of a neural network-based language method named as Structure-Content Neural Language Method by [4] and can generate novel descriptions.

(b) ***Neural Image Caption (NIC) Generator***

This method is proposed by [5] which uses a CNN as an encoder for image representations and RNN as a decoder for generating captions of an image shown in Fig. 2. The encoder in this method follows a novel approach where the last hidden layer in the model is fed as an input to the decoder [16].

Working

The encoder (RNN) translates the input of variable length into a fixed dimensional vector [5] and decodes this representation into required output which is the description. The probability of the right caption is calculated using Eq. 1 [17], where I is an image, and S sentence of its length is unbounded.

$$\theta^* = \underset{\theta}{\arg\max} \sum_{(I,S)} \log p(S|I;\theta) \tag{1}$$

Sampling is one of the approaches used in [17] where the first word was sampled according to p_1, equivalent embed was supplied as an input for sample p_2, continuing in the same order like this until all the samples reach a special end-of-sentence token or it has reached with some maximum length. Second approach uses a search

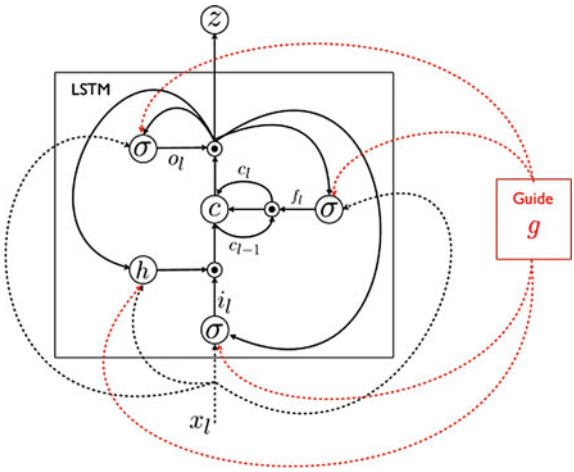

Fig. 3 gLSTM network proposed by [6]

technique called a Beam Search, by iteratively taking all the *k*-best description till some time *t*.

(c) ***gLSTM***

This method is an extension of LSTM proposed by Jia et al. [6]. The author [6] used a concept known as guided LSTM [6] which could create long description in form of sentence by adding global semantic info. This information was then added to each LSTM's gates and cells shown in Fig. 3. It also takes into consideration various normalization strategies to manage the caption length.

Working

Firstly, in this method, for describing an image and extracting the semantic information from an image a Cross-Modal Retrieval (CRM) is used. Multimodal embedding space can also be used to extract the information of the image. Secondly, semantic information is added to the computation of each gates and cell state. Thus, the information is obtained together from the images and its descriptions, aiding as a guide in the procedure of generating a word sequence. In the LSTM method, the generation of a word generally depends on the embedding word at the present time step and the previous hidden state.

$$i'_l = \sigma\left(W_{ix}x_l + W_{im}m_{l-1} + W_{iq}g\right) \tag{2}$$

$$f'_l = \sigma\left(W_{fx}x_l + W_{fm}m_{l-1} + W_{fq}g\right) \tag{3}$$

$$o'_l = \sigma\left(W_{ox}x_l + W_{om}m_{l-1} + W_{oq}g\right) \tag{4}$$

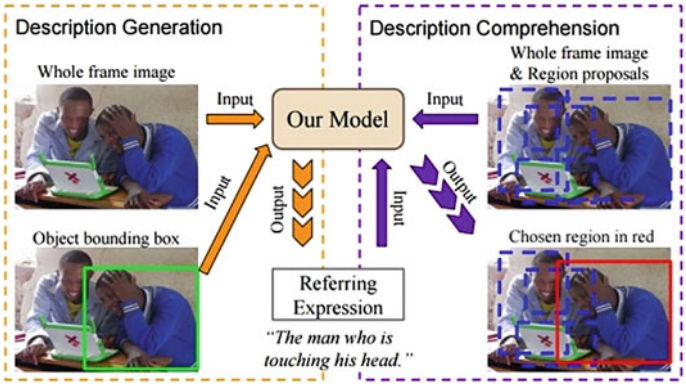

Fig. 4 Illustration of referring expression proposed by [7]

$$c'_l = f'_l \odot c'_l + i'_l \odot h\left(W_{cx}x_l + W_{cm}m_{l-1} + W_{cq}g\right) \tag{5}$$

$$m_l = o'_l \odot c'_l \tag{6}$$

In the above equations by [6], vector representation of semantic information is denoted by g. While $\odot$ denotes the element-wise multiplication, σ represents the sigmoid function, and h is the hyperbolic tangent function. The variable i_l represents the input gate, f_l, is forget gate o_l is output gate, c_l, and m_l memory cell unit and hidden layer.

(d) ***Referring expression***

Mao et al. [7] proposed a new method called as a referring expression. The expression determines a unique description for a particular object or may be an area specified in a given image shown in Fig. 4. Further, the expression can be interpreted to infer which object is being described [7]. This method used well-defined performance metric which gives more detailed than just image captions and as a result provides more helpful.

Working

This method generally considers two characteristics: description generation and description comprehension. In first characteristics, a text expression is generated that exclusively identifies object which is highlighted or some specific region emphasized in the image. In the second characteristics, the method inevitably chooses an object from a given expression that refers to this chosen object. Alike other image captioning methods this method uses a CNN model to represent the image latter followed by an LSTM. The method computes feature for the whole image, to serve as context [7]. It is considered a novel dataset which they termed as ReferIt dataset [18].

(e) ***Variational Auto Encoder (VAE)***

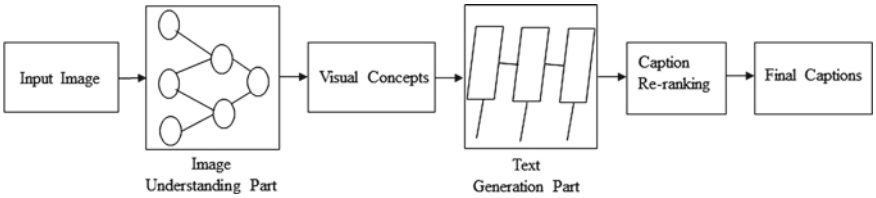

Fig. 5 Compositional-based framework [1]

This method is proposed by [8] using a semi-supervised learning technique. The encoder considered here is a deep CNN and Deep Generative Deconvolutional Neural Network (DGDN) as a decoder. The framework may also even allow unsupervised CNN learning, based on an image [8].

Working

CNN is used as an image encoder for captioning, whereas a recognition method was established for the DGDN as a decoder which decodes the latent image features [8]. The encoder supplies an approximation of distribution for the latent DGDN features which is then related to generative methods for labels or captions. They used Bayesian Support Vector Machine for generating the labels of an image and RNN for giving captions. In the process of generating a label or a caption for any new images, the task of calculating the average across the distribution of latent codes is performed.

1.1.2 Compositional-Based Framework

The second type of architecture is mainly composed of several individual functional components [1]. This approach used CNN that extracts the meaning from an image using a language method illustrated in Fig. 5.

This framework performs the following steps:

(i) Extraction of unique visual features from the image.
(ii) Derived visual attributes from the extracted features.
(iii) Use the visual features and the visual attributes in a language method to generate probable captions.
(iv) Provides ranking for the probable caption using a deep multimodal similarity method, to determine the best suitable captions.

(a) ***Generation-based image captioning from sample***

This method is proposed by [15], which is composed of several components: (i) visual detectors, (ii) a language method, and (iii) multimodal similarity method so as to train the method on dataset of an image captioning.

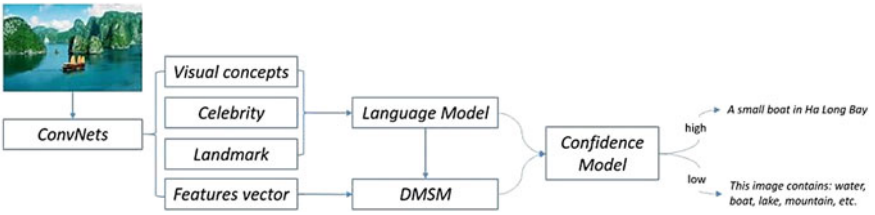

Fig. 6 Illustration of image caption pipeline [10]

Working

A Multiple Instance Learning (MIL) [19] for training the visual detectors for word that are commonly occurs in a caption. This includes several parts of speech like a noun, verb, and adjectives. An image sub-region was considered by this method rather than the complete image. Outputs of a word detector act as conditional inputs to a maximum-entropy language method. The features extracted from the sub-regions are represented with the words probably present in the image captions. Maron and Lozano-Pérez [19] performed re-ranking of the captions using sentence-level features and a deep multimodal similarity method to acquire the semantic information of an image.

(b) ***Generation of description from wild***

This method is introduced by [10] for different image captioning that automatically describes images in the wild. It used the compositional framework like [9], where in [10] the image caption systems are established using different components which are trained independently and latter combined in the main structure shown in Fig 6.

Working

In this method, for identifying a comprehensive visual concept the authors have considered a deep residual network-based vision method. On the other hand, to identify images of celebrities and landmarks; an entity recognition method is used. A classifier for estimating the confidence score for each output caption [10] for generating candidates a language method and for ranking the caption deep multimodal semantic method are considered.

(c) ***Generation of descriptions with structural words***

A compositional network-based image captioning method is proposed by [11]. This method follows some structural words format as: <*object, attribute, activity, scene*> [11]. These structural words are used to generate semantically meaningful descriptions using multi-task method which is comparable to MIL [19] method. Then LSTM [20] machine translation method is used to translate the structural words into image captions.

Working

Figure 7 describes the framework with two stages capable of identifying structural words and generating descriptions from image. Identification of the structural words sequence <objects, attributes, activities, scene> is carried out in the first stage. The image captions that contain comparatively larger information by deep RNN are translated from the word sequence (recognized in first stage) in the second stage.

(d) ***Parallel-fusion RNN-LSTM architecture***

Wang et al. [12] proposed a method based on deep convolutional networks and recurrent neural networks shown in Fig. 8. The main idea is to combine the benefits of RNN and LSTM which leads to decrease in complexity and increase in performance. RNN hidden units are composed of several equal dimension components that work parallel. The outputs are then merged with corresponding ratios to generate final output.

Working

This method follows the following strategies:

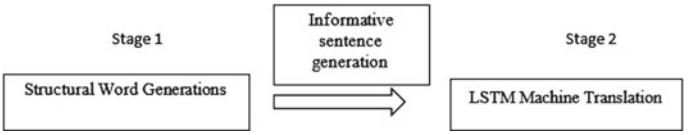

Fig. 7 Stages of generating sentence [11]

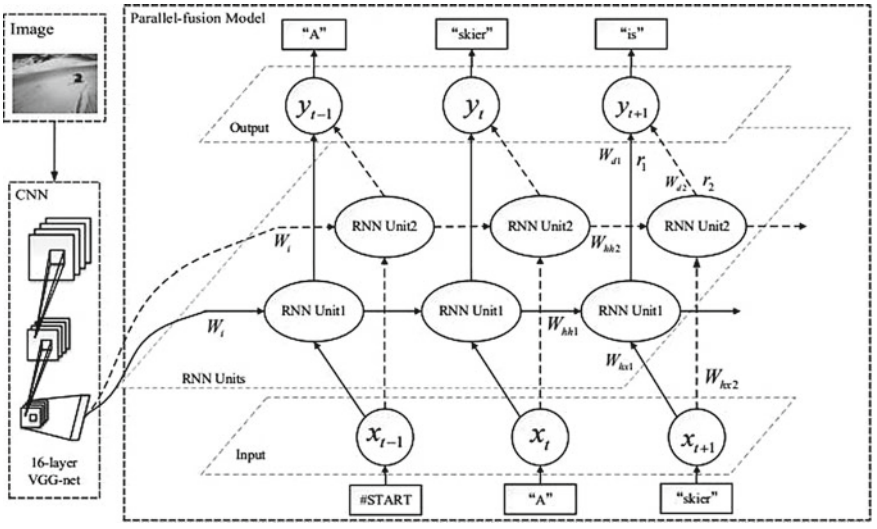

Fig. 8 Method proposed by [12]

1. Splits the hidden layer into 2 parts with both the parts remaining uncorrelated until the output unit.
2. From source data, identical features vectors are fed to the hidden layers along with feedback outputs of the respective hidden layers from the past.
3. Send the generated output of the RNN unit to y_t component of the overall output module.

$$h_{1_t} = \max\left(W_{hx1}x_t + W_{hh1}h_{1_{t-1}} + b_{h1}, 0\right) \tag{7}$$

$$h_{2_t} = \max\left(W_{hx2}x_t + W_{hh2}h_{2_{t-1}} + b_{h2}, 0\right) \tag{8}$$

$$y_t = \text{softmax}\left(r_1 W_{d1}h_{1_t} + r_2 W_{d2}h_{2_t} + b_d\right) \tag{9}$$

$$dy_1 = r_1 \times dy \tag{10}$$

$$dy_2 = r_2 \times dy \tag{11}$$

As per [12], the parameters considered are $\{h1, h2\}$ as the hidden units, $\{W_{hh1}, W_{hh2}, b_{h1}, b_{h2}, W_{d1}, W_{d2}\}$ denotes the weighted parameters while dy is the Matrix for a softmax derivative. The ratios considered are r_1, r_2: ratios.

(e) ***Fusion-based Recurrent Multimodal (FRMM) method***

This method proposed by [13] introduced an end-to-end trainable Fusion-based Recurrent MultiModal (FRMM) method that can address multimodal applications which allow each input modality to be independent w.r.t architecture, parameters and length of input sequences shown in Fig. 9.

Working

The method has separate stages whose outputs are mapped to a common description so that it can be associated with one another during the fusion stage. The outputs are predicted by the fusion stage based on the association. Supervised learning occurs in each all stage. Figure 9 describes how the FRMM method works by taking a video description method as an example. The FRMM method learns the behavior in separate stages.

1.2 Summary

Tables 1 and 2 show the different image captioning methods using deep neural algorithms. The table is categorized into two parts based on the framework used to generate a caption for an image experimented on datasets such as *Flickr8K*,

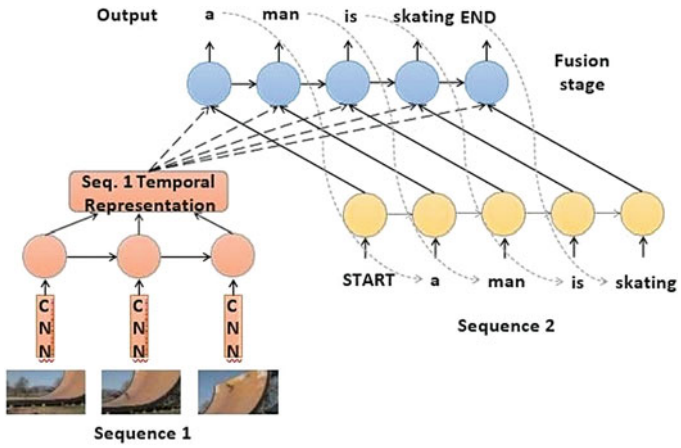

Fig. 9 Method proposed by [13]

Table 1 Summary of generating image caption based on encoder-decoder framework on the dataset using evaluation metric

Method	Dataset	BLEU-1	R@10
Encoder-decoder pipeline [4]	Flickr8K Flickr30K	– 	0.55 0.629
NIC [5]	PASCAL Flickr8K Flickr30K	0.59 0.63 0.66	0.61 0.56
gLSTM [6]	Flickr8K Flickr30K MS COCO	0.647 0.646 0.67	–
Referring expression [7]	MSCOCO	–	–
Variational autoencoder [8]	Flickr8K Flickr30K MSCOCO	0.70 0.69 0.71	–

Table 2 Summary of generating image caption based on compositional-based framework on the dataset

Method	Dataset	BLEU-1	BLEU-4
Generation-based image captioning from sample [9]	MSCOCO	–	0.291
Generation of description from wild [10]	MSCOCO Adobe-MIT FiveK Instagram Image	–	–
Generation of descriptions with structural words [11]	UIUC Pascal Flickr8k	0.40 0.621	–
Parallel-fusion RNN-LSTM [12] architecture	Flickr8k	0.647	–
FRMM [13]	Flickr30k MSCOCO	0.589 0.702	0.177 0.276

Flickr30K, PASCAL, UIUC PASCAL, and *MSCOCO* considering only **BLEU** and **Recall@k** evaluation metrics. The first category, encoder-decoder framework uses to generate a caption which got inspired from the concepts of translating sentences from one language into another language. Under this framework, various authors [4–8] have been successful in generating captions of images. Contrary to the encoder-decoder framework, the second category that has been considered is a Compositional Architecture proposed by [9–13] image captioning.

In Table 1, among the method that follows encoder-decoder framework, variational autoencoder [8] has shown higher BLEU-1 value compared with the other methods experimented in MS COCO datasets. In Table 2, among the compositional-based architecture: FRMM [13] has higher BLEU-1 evaluation value experimented on MSCOCO dataset.

1.3 Conclusion

On carrying out a comprehensive survey on image captioning methods based on some deep learning methods. The following ideas were derived from: encoder-decoder framework and compositional architecture. The encoder-decoder framework is used to generate various captions from images. It first encodes an image to an intermediate representation and then generates a sentence word by word from the representation using the decoder. The compositional image captioning uses a method to detect concepts that visually appear in the input image. The detected concepts are then forwarded to the language method to generate various candidate captions from where one probable caption is chosen as the final caption or description for a given input image.

References

1. Zakir Hossain, M., Sohel, F., Shiratuddin, M.F., Laga, H.: A comprehensive survey of deep learning for image captioning. ACM Comput. Survey **51**(6), 1–36 (2019)
2. Bal, S., An, S.: A survey on automatic image caption generation. Neurocomputing **311**, 291–304 (2018)
3. Kulkarni, G., Premraj, V., Dhar, S., Li, S., Choi, Y., Berg, A.C., Berg, T.L.: Baby talk: understanding and generating simple image descriptions. CVPR (2011)
4. Kiros, R., Salakhutdinov, R., Zemel, R.: Unifying visual-semantic embeddings with multimodal neural language models. arXiv:1411.2539 (2018)
5. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: a neural image caption generator. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3156—3164 (2015)
6. Jia, X., Gavves, E., Fernando, B., Tuytelaars, T.: Guiding long-short term memory for image caption generation. In: Computer Vision (ICCV), IEEE International Conference on Computer Vision (ICCV) (2015)

7. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A.L., Murphy, K.: Generation and comprehension of unambiguous object descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 11–20 (2016)
8. Pu, Y., Gan, Z., Henao, R., Yuan, X., Li, C., Stevens, A., Carin, R.: Variational autoencoder for deep learning of images, labels and captions. In: Proceedings of the Advances in Neural Information Processing Systems, pp. 2352–2360 (2016)
9. Fang, H., Gupta, S., Iandola, F., Srivastava, R.K., Deng, L., Dollār, P., Gao, J., He, X., Mitchell, M., Platt, J.C.: From captions to visual concepts and back. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1473–1482 (2015)
10. Tran, K., He, X., Zhang, L., Sun, J., Carapcea, C., Thrasher, C., Buehler, C., Sienkiewicz, C.: Rich image captioning in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 49–56 (2016)
11. Ma, S., Han, Y.: Describing images by feeding LSTM with structural words. In: 2016 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6. IEEE (2016)
12. Wang, M., Song, L., Yang, X., Luo, C.: A parallel-fusion RNN-LSTM architecture for image caption generation. In: IEEE International Conference on Image Processing (ICIP), pp. 4448–4452. IEEE (2016)
13. Oruganti, R.M., Sah, S., Pillai, S., Ptucha, R.: Image description through fusion based recurrent multi-modal learning. In: IEEE International Conference on Image Processing (ICIP), pp. 3613–3617 (2016)
14. 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. PMLR, pp. 2048–2057 (2015)
15. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**, 1735–1780 (1997)
16. Donahue, J., Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S.: Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2625–2634 (2015)
17. Cho, K., Merrinboer, B.V., Gulcehre, C.: Learning phrase representations using RNN encoder—decoder for statistical machine translation. arXiv:1406.1078v3 (2014)
18. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.L.: ReferItGame: referring to objects in photographs of natural scenes. EMNLP, pp. 787–798 (2014)
19. Maron, O., Lozano-Pérez, T.: A framework for multiple-instance learning. In: Advances in Neural Information Processing Systems, pp. 570–576 (1998)
20. Wang, C., Yang, H., Bartz, C., Meinel, C.: Image captioning with deep bidirectional LSTMs. In: Proceedings of the 2016 ACM on Multimedia Conference, pp. 988–997. ACM (2016)