Journal of Computers Vol. 30 No. 4, 2019, pp. 206-216
doi:10.3966/199115992019083004020

FOF: Fusing Object Features into Deep Learning Model to Generate Image Caption

Hang Zhou[1], Xue-Qiang Lv[1], Xin-Dong You[1*], Zhi-An Dong[1], Kai Zhang[2]

[1] Beijing Key Laboratory of Internet Culture and Digital Dissemination Research,
Beijing Information Science & Technology University, Beijing, China
zhouhang5555@126.com, lxq@bistu.edu.cn, youxindong@bistu.edu.cn, dong.zhian@163.com

[2] China Language Intelligence Research Center, Capital Normal University, Beijing, China
irs_zhangkai@163.com

Received 13 January 2019; Revised 11 March 2019; Accepted 12 March 2019

Abstract. To solve the problem of category errors and number errors of objects in the sentences generated by existing image captioning model, we propose an image captioning model fused with object features. In particular, we integrate object statistical feature and object regional feature extracted from the image into the Convolutional Neural Networks (CNNs) plus Recurrent Neural Networks (RNNs) image captioning framework. Using object detection network to extract object statistical feature and object regional feature, the object statistical feature and the image convolutional feature are used as the input of Long Short-Term Memory (LSTM), and Attention Mechanism (AM) is used to concatenating the object regional feature with the output of LSTM to generate sentences, so that the model obtains additional information about objects categories, objects numbers and objects regions, which helps to improve the quality of the generated description. Experiments are conducted on MSCOCO dataset. Especially compared with the Hard-attention model, BLEU3/4 increase 4.5%, 4.9%, respectively and compared with the g-LSTM model, BLEU3/4 increase 4.4%, 3.5%, respectively. The proposed model is of great significance to solve the problem of object category errors and object number errors in image description.

Keywords: convolutional neural network, image caption, object detection, recurrent neural network

1 Introduction

With the rapid development of information science and technology, image data presents explosive growth. As there is a "semantic gap" [1] between that low-level visual feature and the semantic information of the image, how to understand and manage a large number of untagged image data has become an urgent problem for academia and industry.

Image caption is that computer automatically generates descriptive sentences according to the image content, which is considered as one of the important methods to narrow the "semantic gap" [2]. It combines the two research directions of computer vision and natural language processing. The generated descriptive sentences not only conform to the natural language expression habits, but also capture important information in images, such as people and objects. Therefore, it is difficult to generate high-quality descriptive sentences, which attracts many researchers to study. Image caption can be applied in many fields, such as image understanding of children, visually impaired people, visual intelligent chat robot and image retrieval, which has great application prospects and commercial value.

With the rise of deep learning, computer vision and natural language processing have made a series of breakthroughs in the field of research, image caption have evolved from template-based methods to deep learning-based methods. The current mainstream deep learning-based methods are inspired by machine

* Corresponding Author

translation, machine translation is to translate that source language sentence S into the target language sentence T by maximize the probability $p(T \mid S)$. The RNN is use as an encoder and a decoder, the encoder RNN encodes the source language sentence S into eigenvector of fixed length, the decoder RNN treats the eigenvector as an initial hidden state, and converts the eigenvector into the target language sentence T. The image captioning architectures use CNN as the encoder to encode the image into feature vectors, and then use RNN as the decoder to decode the feature vectors into target description sentences.

Those CNN plus RNN image captioning methods directly convert from image representations to language without explicitly considering more object category information and object number information from the image. Our contributions are as follows.

(1) *we propose the concept of object statistical feature and we devise variants of architectures by feeding it into RNN in different moments and placements.*

(2) *our model combines state-of-art object detection network Faster R-CNN [18], we use Faster R-CNN to extract object regional feature that is a set of salient image regions, with each region represented by a pooled convolutional feature vector, and then integrate object regional feature into deep model.*

(3) *our model outperforms other state-of-the-art methods on MSCOCO. Especially compared with the Hard-attention model, BLEU3/4 increase 4.5%, 4.9%, respectively and compared with the g-LSTM model, BLEU3/4 increase 4.4%, 3.5%, respectively.*

We review the related work about image caption in the second part of the article, then introduce the proposed model FOF in the third part, and show the experimental results in the fourth part, finally give the conclusion in the last part.

2 Related Work

The current image captioning methods can be summarized into three categories: template-based methods, similarity retrieval-based methods and deep learning-based methods.

Template-based image captioning method detects the objects and its attribute in the image through object recognition, and then embeds this information into the pre-designed template in an appropriate way. In 2010, Farhadi et al. [3] used detectors to detect objects in images to infer the < object, action, scene > triple, and converted it into descriptive text using templates. In 2011, Li et al. [4] formed relational phrases using language models after acquiring objects through detectors, and then combined relational phrases and description templates to generate descriptions. In 2013, Kulkarni et al. [5] proposed the Baby Talk model, which uses Conditional Random Field (CRF) to label the detected objects, attributes and relationships, and finally uses templates to generate descriptive statements. However, the descriptions obtained by the above approaches are limited by the templates, and they are stiff and inflexible.

The similarity retrieval-based method utilizes the similarity of image visual features to retrieve, and takes the description text of the image with high similarity as the candidate answer, or maps the image features and text features to the same feature space, and retrieves the high similarity image text as the candidate result. In 2011, Ordonez et al. [6] proposed to use the global features of images to retrieve in millions of images, and take the description of the most similar image as the description of the image to be described. In 2013, Hodosh et al. [7] proposed to align images with description using Kernel Canonical Correlation Analysis (KCCA). In 2014, Gong et al. [8] used Canonical Correlation Analysis (CCA) to map images and text to the same feature space and retrieve the text most similar to images from databases. However, such methods cannot generate sentences entirely from the image content, nor can they generate descriptions that do not exist in the database.

With the development of deep learning, researchers have proposed an image captioning method based on deep learning. In 2015, Mao et al. [9] proposed multimodal Recurrent Neural Network (m-RNN), using convolutional neural network to extract image convolutional features, and put the features into the multimodal recurrent neural network at each time to generate description words. In the same year, Vinyals et al [10] proposed the image description generation model (NIC) based on convolutional neural network and Long Short-Term Memory (LSTM). Unlike Mao, NIC uses LSTM to build language model to generate description sentences, and achieves good results.

Subsequently, the researchers improved the NIC model and the quality of the generated description. In 2015, Xu et al. [11] introduced Attention Mechanism into the model for the first time, which enables the

model to capture the local information of the image. In 2016, Jia et al. [12] used semantic information to guide LSTM in generating descriptions. In 2017, Tang et al. [13] put the scene priori information in the image into the model, and cooperated to generate the description sentences of the image. In 2018, Liu et al. [14] added stack hidden layer and common hidden layer to the model to improve the learning ability of the language model. Liu [15] and Lan [16] studied the Chinese image caption, and optimized the NIC, which improved the effect of the model.

Most image captioning models use CNN to extract image convolutional features as the initial input of RNN, and then use RNN to generate description sentences step by step. Although the model schemas are different, the generated sentences still have errors in describing the category of objects and the number of objects. To solve this problem, we propose an image captioning model fused with object features in this paper. In order to obtain more information about object categories, object numbers and object regions. We use Faster R-CNN to extract the object statistical feature and object regional feature. Our work not only explores the use of image representation and object features for image captioning, but also explores the relationship between the two to design the architecture better.

3 FOF Model

We first describe the CNN-LSTM-based image captioning model in Sec.3.1, then introduce our proposed image caption generator in Sec.3.2 & Sec3.3, finally the training process of the model is explained in Sec.3.4.

3.1 CNN-LSTM-Based Image Captioning Model

The model is composed of two parts: CNN and LSTM. CNN is used as the encoder to extract the image convolutional features, and LSTM is used as the decoder to decode the image convolutional features as the initial input to generate the target description. The LSTM inputs a vector at each time and outputs a vector. More specifically, LSTM first accepts the image convolutional feature and ignores the output at this time. After inputting a start symbol < Start >, LSTM outputs a vector composed of the predicted probability of each word in the vocabulary, and selects the word with the highest probability as the current output according to the output vector. Then, the output vector is used as an input to the next time, and LSTM continues the prediction until the end symbol < End > is predicted. The overall structure of the CNN-LSTM-Based model is shown in Fig. 1.

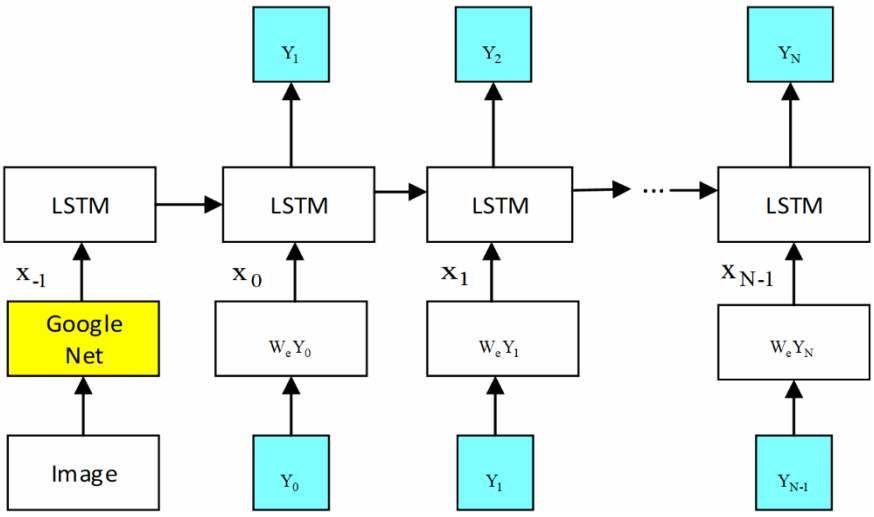


Fig. 1. The overall structure of the CNN-LSTM-Based model

The encoder CNN is a kind of neural network used to process grid data. Typical convolutional neural network structure consists of input layer, convolutional layer, pooling layer, fully connected layer and output layer. The input image is passed through convolutional layer, pooling layer and full connected layer to obtain image convolutional feature that has powerful descriptive ability, and can be used in

image classification, image segmentation and other tasks.

The decoder LSTM is a kind of neural network used to process sequential data. It can be used to build language models, that is, to predict what the next word most likely for a given sentence. At the same time, LSTM can effectively solve long-term dependency problems. The core of LSTM is cell state C. Cell state can selectively retain the input information at each moment and run through the whole training process.

LSTM adds or removes information to cell state C through different "gates" structures, including "forget gate", "input gate" and "output gate", which control whether the current cell state is forgotten, whether the current input is read, and whether the new cell state is output. The formula is as follows:

$$f_t = \sigma_g(W_f x_t + U_f h_{t-1} + b_f). \tag{1}$$

$$i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i). \tag{2}$$

$$o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o). \tag{3}$$

$$c_t = f_t * c_{t-1} + i_t * \sigma_h(W_c x_t + U_c h_{t-1} + b_c). \tag{4}$$

$$h_t = o_t * c_t. \tag{5}$$

Where * denotes dot product, t denotes moment t, x_t denotes input vector of LSTM, f_t denotes activation vector of LSTM "forggt gate", i_t denotes activation vector of LSTM "input gate", o_t denotes activation vector of LSTM "output gate", h_t denotes hidden state of LSTM, c_t denotes cell state of LSTM, W, U, b denotes parameters of LSTM to be trained, σ_g denotes Sigmoid function, σ_h denotes Tanh function.

The model needs to maximize the probability that a given image generates a target description statement during the "encode-decode" process, represented by Equation (6):

$$\theta^* = \arg\max_\theta \Sigma_{(I,Y)} \log p(Y \mid I; \theta) \tag{6}$$

Where l denotes the input image, Y denotes any indefinite length of the target description sentence, and consists of the words $Y_0, Y_1, \ldots, Y_N$ and θ denotes the model parameters.

3.2 Object Feature

Object feature include object statistical feature I_r and object regional feature V_o. We propose the concept of object statistical feature. Through object detection network, we can obtain information on categories and numbers of all objects in the image, and statistically analyze the information to form the object statistical feature, so it contains the category information and number information of objects. The dimensions of object statistical feature are the number of all categories of objects that the object detection network can detect, and the value of each dimension is the number of objects in a class. As shown in Fig. 2, two dogs and a Frisbee can be detected by the object detection network, so the value of the dimension representing the dog is 2, the value representing the dimension of the Frisbee is 1, and the remaining dimension is 0. Object regional feature are a set of pooled convolutional feature vectors for classification and border regression extracted by object detection network.

Specifically, we use Faster R-CNN, which is the most accurate object detection network. Its network architecture includes Convolutional Layers, Region Proposal Network (RPN), Region of Interest Pooling (RoI Pooling) and Classifier. The process of detecting objects by Faster R-CNN is as follows. Firstly, the convolutional features of the detected image are extracted by Convolutional Layers. Secondly, the convolutional features are input into the RPN to predict the object region proposals. Then, the convolutional features of the object region proposals (RoI feature) are obtained by putting the convolutional features and object region proposals into RoI pooling. Finally, the RoI features are input into Classifier for classification and the border regression.

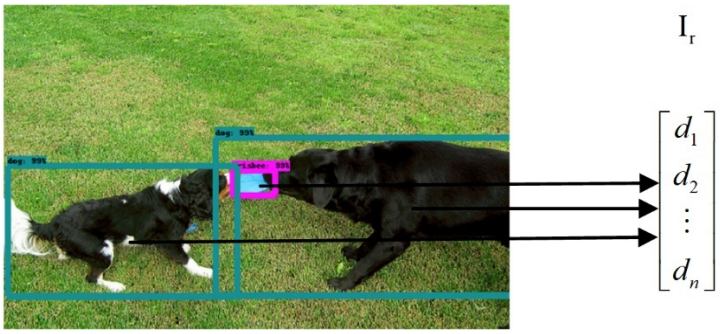


Fig. 2. Process of constructing object statistical feature

All categories and numbers of objects in the Faster R-CNN detection results are statistically analyzed to form the object statistical feature I_r. The RoI features which used for classification and border regression in Faster R-CNN are used as the object regional feature.

3.3 FOF: Fusing Object Features into Deep Learning Model

Different from the existing CNN-LSTM-based image captioning model that uses image convolutional features to generate sentences, we propose an image caption generator to integrate the detected object features into LSTM. We designed four variants for two purposes. The first goal is about where to feed object statistical feature to the LSTM and three architectures, FOF-RC, FOF-CR, and FOF-Fusion. The second is how to integrate the object regional feature into LSTM and we design FOF-Ensemble. These four model architectures are shown in Fig. 3.

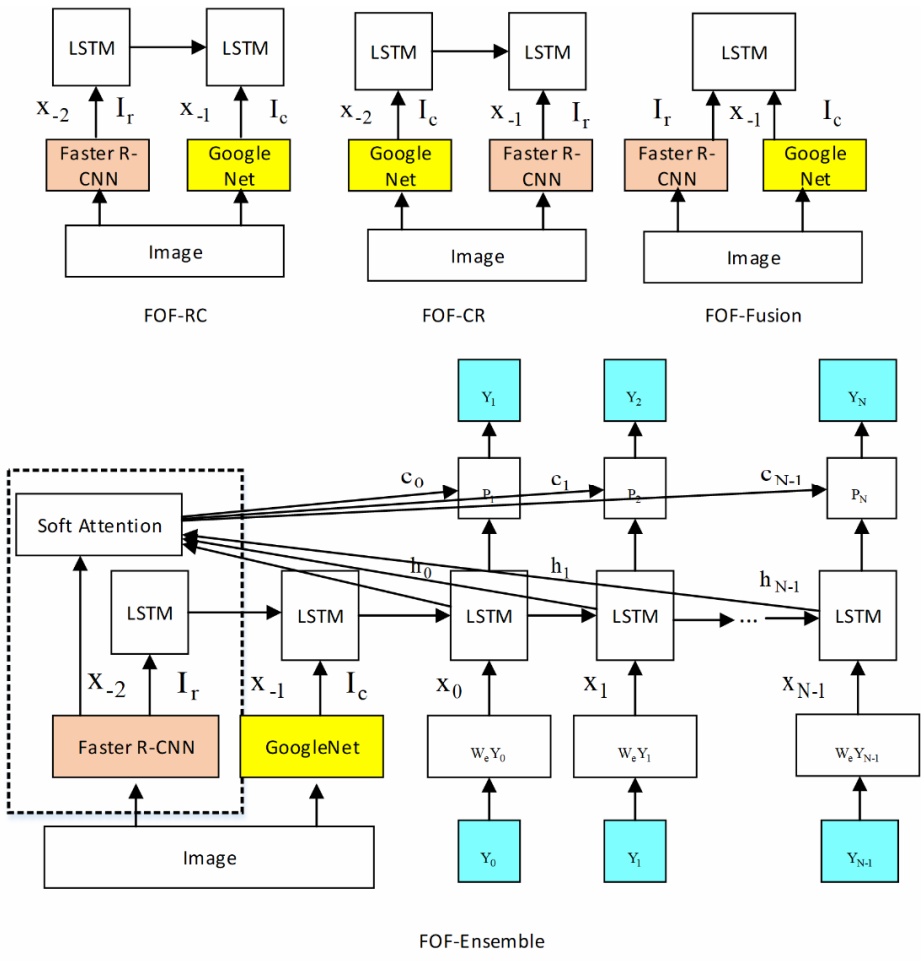


Fig. 3. Overview of four model architectures

FOF-RC. As with the image convolutional feature, it is natural to directly feed object statistical feature into LSTM as input. The first method FOF-RC is to treat the object statistical feature I_r as the input of the first moment of the LSTM, and the image convolutional feature I_c as the input of the second moment of the LSTM. The formulas are as follows:

$$x_{-2} = W_r I_r. \tag{7}$$

$$x_{-1} = W_c I_c. \tag{8}$$

$$x_t = W_e Y_t, t \in \{0, \ldots, N-1\}. \tag{9}$$

$$h_t = LSTM(x_t, h_{t-1}). \tag{10}$$

$$p_{t+1} = softmax([h_t]). \tag{11}$$

In the formulas, I_r is the object statistical feature, I_c is the image convolutional feature, W_r, W_r are fully connected layer parameters, W_e is word embedding vector, and we represent each word as a one-hot vector Y_t of dimension equal to the size of the dictionary.

FOF-CR. The second method FOF-CR is similar to FOF-RC. FOF-RC takes the image convolutional feature I_c as the input of the first moment of LSTM, and the object statistical feature I_r as the input of the second moment of LSTM. The formulas are as follows:

$$x_{-2} = W_r I_r. \tag{12}$$

$$x_{-1} = W_c I_c. \tag{13}$$

FOF-Fusion. Different from the former two methods, the third method FOF-Fusion is to concatenate the object statistical feature I_r and image convolutional features I_c, and send them to LSTM at the same time. The formulas are as follows:

$$x_{-1} = [W_r I_r; W_c I_c]. \tag{14}$$

FOF-Ensemble. The last method is based on FOF-RC. FOF-Ensemble takes the object regional feature matrix V_o as the input of attention mechanism, focuses on the image region through attention mechanism, and then concatenates the focusing result with the output of LSTM to generate sentences, so that the model can obtain more information of object region. The formulas are as follows:

$$x_{-2} = W_r I_r. \tag{15}$$

$$x_{-1} = W_c I_c. \tag{16}$$

$$x_t = W_e Y_t, t \in \{0, \ldots, N-1\}. \tag{17}$$

$$h_t = LSTM(x_t, h_{t-1}). \tag{18}$$

$$z_t = w_h^t \sigma_h (W_v v_o + (W_g h_t) 1^T). \tag{19}$$

$$\alpha_t = softmax(z_t). \tag{20}$$

$$c_t = \sum_{i=1}^{k} \alpha_{ti} v_{toi}. \tag{21}$$

$$p_{t+1} = softmax([c_t : h_t]). \tag{22}$$

3.4 Training Process

Taking FOF-Ensemble model training process as an example: Faster R-CNN is used to extract the object statistical feature I_r and the object regional feature matrix V_o from the image. Image convolutional feature I_c extracted by convolutional neural network GoogleNet; The object statistical feature I_c and the image convolutional feature I_c are used as the input of the first two moments of LSTM; The word feature vector obtained by word embedding is used as the input of LSTM at other time; The hidden state h_t of the LSTM and the object regional feature matrix V_o are input to the attention mechanism at each moment (except the first two moments) for focusing image region; Finally, the hidden state h_t of LSTM and the output c_t of attention mechanism are used to predict the word, and then the loss of the model is calculated.

When LSTM predicts that the next word is the terminator, it indicates that a complete sentence has been generated. The loss function is the negative logarithmic sum of the probability of outputting the correct word at each moment and minimizes the loss function by a stochastic gradient descent, as follows:

$$L(I,Y) = -\sum_{t=1}^{N} \log p_t(Y_t). \tag{23}$$

4 Experiment

4.1 Datasets and Evaluation Metrics

The image caption dataset used in this paper is MSCOCO dataset. The MSCOCO dataset contains 123,287 annotated images, of which 82,783 are training set and 40,504 are validation set. Each image in the datasets has five descriptive sentences corresponding to it, as shown in Fig. 4.

(1) A woman posing for the camera standing on skis.
(2) a woman standing on skiis while posing for the camera.
(3) A woman in a red jacket skiing down a slope.
(4) A young woman is skiing down the mountain slope.
(5) a person on skis makes her way through the snow.

Fig. 4. Example of MSCOCO dataset

The MSCOCO dataset is segmented according to the segmentation method used in Kapathy et al. [19]. For MSCOCO dataset, 5000 images in the original validation set are taken as the validation set, 5000 images in the original validation set are taken as the test set, and the rest images in the original validation set and the original training set are taken as the training set.

The evaluation metrics uses in this paper are BLEU [20], METEOR [21] and CIDEr [22]. Among them, BLEU is a machine translation method to measure the similarity between generated sentences and reference sentences by calculating the co-occurrence degree of n-gram words in generated sentences and reference sentences. METEOR is used to measure the quality of machine translation generated sentences.

The method matches the 1-gram words in the generated sentence and the reference sentence in different ways to find the maximum value of the match. When the same matching value exists, the matching method with the least number of word crosses in the corresponding sentences is selected to obtain the matching set m. The accuracy and recall rate of the corresponding sentence are calculated by the matching set m. The METEOR value is the harmonic average of the accuracy and recall of the corresponding generated and reference sentences. CIDEr (Consultative-based Image Description Evaluation) calculates the TF-IDF weight of each n-gram phrase to measure the similarity between the generated sentence and the reference sentence.

4.2 Experiment Details

The lab environment is configured as follows: Intel Xeon E5-2603 v4 processor, 64G RAM, Nvidia Tesla k80 graphics card, operating system Ubuntu 16.03.1, development language python 2.7, and deep learning framework Tensorflow 1.4.

Network configuration parameters, as shown in Table 1.

Table 1. Network configuration parameters

Parameter name	Parameter value
number of LSTM units	512
dimension of I_c	2048
dimension of word embedding	512
dimension of I_r	80
dimensions of V_o	20×2048
MSCOCO vocabulary size (word frequency more than 4 times)	11348

During training, the batch size of the model is set to 32, and the initial learning rate is set to 2. The initial state and all its parameters in LSTM, the parameters of fully connected layer after GoogleNet, the parameters of fully connected layer after Faster R-CNN, the parameters of attention mechanism w_h^l, W_v, W_g and the parameters of word embedding matrix $W-e$ are all randomly initialized by standard normal distribution, and optimized by random gradient descent method. In order to optimize the model parameters better, the learning rate decreases to 1/2 of the original one after every 8 epochs training. The maximum iteration times of MSCOCO dataset is set to 500k. Finally, the experimental results are obtained.

4.3 Experiment Results

Quantitative analysis. As shown in Table 2, the benchmark model NIC, FOF-RC, FOF-CR, FOF-Fusion and FOF-Ensemble proposed in Chapter 3 are compared experimentally on MSCOCO dataset. Experiments show that the three models fusing object statistical feature are better than the baseline model NIC, and FOF-RC is the best. Compared with NIC model, FOF-RC improves BLEU3-4 by 6.4% and 4.6%, respectively. It shows that the method of integrating object statistical feature by FOF-RC can improve the accuracy between the generated sentences and the reference sentences more effectively, and the quality of the generated sentences is higher. However, FOF-Ensemble is better than FOF-RC. BLEU3-4, METEOR and CIDEr are improved by 0.9%, 0.7%, 0.5%, 2.5% respectively. It shows that the model can further improve the accuracy of sentences by fusing object regional feature, and verify the validity of the model proposed in this paper.

Table 2. Compare on MSCOCO Dataset

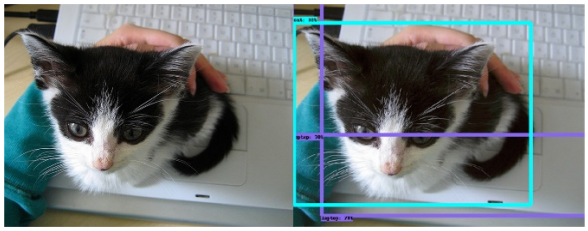
Method	B-3	B-4	M	C
NIC	32.9	24.6		
FOF-RC	39.3	29.2	24.6	93.3
FOF-CR	36.4	26.8	22.3	87.9
FOF-Fusion	38.3	28.4	24.3	90.3
FOF-Ensemble	**40.2**	**29.9**	**25.1**	**95.8**

As shown in Table 3, the proposed model is compared with m-RNN, Soft-Attention, Hard-Attention, g-LSTM, F-SOCPK, Liu Chang [14] on the MSCOCO dataset. Compared with Hard-attenuation model and g-LSTM model, B3-4 and METEOR are improved by 4.5%, 4.9%, 2.1% and 4.4%, 3.5%, 2.4% respectively. Compared with F-SOCPK model, each evaluation metric is improved by 1.2%, 1.8%, 1.2%, 7.6% respectively. First of all, the BLEU shows that the generated sentences and reference sentences have more co-occurrence words, and the description is more accurate. Secondly, the improvement of METEOR indicates that the generated sentences have higher accuracy and recall rate. Finally, a significant increase in the CIDEr indicates that the generated sentences is more similar to the reference sentences.

Table 3. Compare with other methods on MSCOCO dataset

Method	B-3	B-4	M	C
m-RNN [9]	35.0	25.0		
NIC [10]	32.9	24.6		
Soft-attention [11]	34.4	24.3	23.9	
Hard-attention [11]	35.7	25.0	23.0	
g-LSTM [12]	35.8	26.4	22.7	81.3
F-SOCPK [13]	39.0	28.1	23.9	88.2
Liu Chang [14]	39.3	28.6	24.1	95.2
FOF-Ensemble	**40.2**	**29.9**	**25.1**	**95.8**

Qualitative Analysis. Taking MSCOCO dataset as an example, the results of object detection network and experiment are presented, and compared with the results of benchmark model NIC and ground truth. As shown in Fig. 6, that results are labeled (1), (2), (3) and (4) from top to bottom and from left to right, respectively.



NIC: a black and white cat sitting on a desk.
FOF-Ensemble: a black and white cat laying on top of a laptop computer.
Ground Truth: black and white kitten sitting on the keyboard of a laptop.

NIC: a group of people sitting around a wooden table.
FOF-Ensemble: a herd of cattle laying on top of a dirt field.
Ground Truth: A group of colorful cows laying inside of a barn.

NIC: a black dog with a frisbee in its mouth.

FOF-Ensemble: two dogs playing with a frisbee in the grass.
Ground Truth: Two dogs playing tug of war with a blue frisbee.

NIC: a horse grazing in a field with a fence in the background.
FOF-Ensemble: a group of horses grazing in a field.

Ground Truth: There are some horses grazing on grass in a pen.

Fig. 6. Comparison of MSCOCO results

(1), (2) The description results of two images using NIC model have wrong description of object category. “Laptop” is described as “desk” in (1), “colorful cows” is described as “people”, “wooden

table" in (2). The object category information is fused in the model proposed in this paper. From the right figures of (1) and (2), the object detection network detects the object in the image, including "cat", "laptop", "cow", which makes the description result contain the information of "laptop computer", "cattle", and so on, and it can describe the image correctly, which reflects the validity of the model proposed.

(3), (4) There are errors in describing the number of objects in the results of two images using NIC model. "two dogs" is described as "a dog" in (3), "some horses" is described as "a horse" in (4), and the object number information is fused in the model proposed in this paper. From the right figures of (3) and (4), it can be seen that the number of objects is detected correctly by the object detection network, which makes the description result accord with the image content more, and further demonstrates the validity of the model proposed.

5 Conclusions and Future Work

In this paper, an image description generation model fused with object features is proposed to solve the problem in the current mainstream deep learning model when generating sentences. The object detection network and attention mechanism are introduced to obtain the information about object category, object number and object region in the image, which effectively reduces the errors of object category and number in generated description and improves the quality of generated sentences. However, the proposed model still has room for improvement. In the further work, we will improve the proposed model to increase the diversity of description sentences and enrich the vocabulary in description sentence. Moreover, we can further fuse the information of human action and the relationship between objects in the image, in order to improve the quality of description sentences.

Acknowledgements

This work is supported by National Natural Science Foundation of China under Grants No. 61671070, National Language Committee of China under Grants ZDI135-53, and Project of Cycle Economy and Knowledge Management Based on Big Data in Promoting the Developing University Intension – Disciplinary Cluster No. 5111823517.

References

[1] Y.-X. Xie, X.-D. Luan, L.-D. Wu, Multimedia data semantic gap analysis, Journal of Wuhan University of Technology 33(6)(2011) 859-863.

[2] X.-R. Li, T. Uricchio, L. Ballan, Socializing the semantic gap: a comparative survey on image tag assignment, refinement, and retrieval, ACM Computing Surveys 49(1)(2016) 14: 1-14: 39.

[3] A. Farhadi, M. Hejrati, M.A. Sadeghi, Every picture tells a story: generating sentences from images, in: Proc. 2010 11th European Conference on Computer Vision, 2010.

[4] S.-M. Li, G. Kulkarni, T.L. Berg, Composing simple image descriptions using web-scale n-grams, in: Proc. 2011 15th Conference on Computational Natural Language Learning, 2011.

[5] G. Kulkarni, V. Premraj, V. Ordonez, Babytalk: understanding and generating simple image descriptions, IEEE Transactions on Pattern Analysis & Machine Intelligence 35(12)(2013) 2891-2903.

[6] V. Ordonez, G. Kulkarni, T.L. Berg, Im2Text: describing images using 1 million captioned photographs, in: Proc. 2011 International Conference on Neural Information Processing Systems, 2011.

[7] M. Hodosh, P. Young, J. Hockenmaier, Framing image description as a ranking task: data, models and evaluation metrics, Journal of Artificial Intelligence Research 47(1)(2013) 853-899.

[8] Y.-C. Gong, L.-W. Wang, M. Hodosh, Improving Image-Sentence Embeddings Using Large Weakly Annotated Photo Collection, in: Proc. 2014 European Conference on Computer Vision, 2014.

[9] J.-H. Mao, W. Xu, Y. Yang, Deep captioning with multimodal recurrent neural networks (m-RNN), in: Proc. 2015 International Conference on Learning Representations, 2015.

[10] O. Vinyals, A. Toshev, S. Bengio, Show and tell: a neural image caption generator, in: Proc. 2015 IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[11] K. Xu, J.L. Ba, R. Kiros, Show, Attend and tell: neural image caption generation with visual attention, in: Proc. 2015 32nd International Conference on Machine Learning, 2015.

[12] X. Jia, E. Gavves, B. Fernando, Guiding the long-short term memory model for image caption generation, in: Proc. 2016 IEEE International Conference on Computer Vision, 2016.

[13] P.-J. Tang, Y.-L. Tan, J.-Z. Li, Image description based on the fusion of scene and object category prior knowledge, Journal of Image and Graphics 22(9)(2017) 1251-1260.

[14] C. Liu, X.-D. Zhou, B.-L. Shi, Image caption based on image semantic similarity network, Computer Applications and Software 35(1)(2018) 211-216.

[15] Z.-Y. Liu, L.-L. Ma, J. Wu, L. Sun, Chinese image captioning method based on multimodal neural network, Journal of Chinese Information Processing 31(6)(2017) 162-171.

[16] W.-Y. Lan, X.-X. Wang, G. Yang, X.-R. Li, Improving chinese image captioning by tag prediction, Chinese Journal of Computers 42(1)(2019) 136-148.

[17] C. Szegedy, W. Liu, Y.-Q. Jia, Going deeper with convolutions, in: Proc. 2015 IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[18] S.-Q. Ren, K.-M. He, R. Girshick, Faster R-CNN: towards real-time object detection with region proposal networks, in: Proc. 2015 International Conference on Neural Information Processing Systems, 2015.

[19] A. Karpathy, F.-F. Li, Deep visual-semantic alignments for generating image descriptions, in: Proc. 2015 IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[20] K. Papineni, A method for automatic evaluation of machine translation, in: Proc. 2002 40th Annual Meeting on Association for Computational Linguistics, 2002.

[21] M. Denkowski, A. Lavie, Meteor universal: language specific translation evaluation for any target language, in: Proc. 2014 9th Workshop on Statistical Machine Translation, 2014.

[22] R. Vedantam, C.L. Zitnick, D. Parikh, CIDEr: consensus-based image description evaluation, in: Proc. 2015 IEEE Conference on Computer Vision and Pattern Recognition, 2015.