Image Caption with Global-Local Attention*

Linghui Li,[1,2] **Sheng Tang,**[1,†] **Lixi Deng,**[1,2] **Yongdong Zhang,**[1] **Qi Tian**[3]
[1]Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing 100190, China
[2]University of Chinese Academy of Sciences, Beijing 100039, China
[3]Department of Computer Science, University of Texas at San Antonio, TX 78249-1604
{lilinghui,ts,denglixi,zhyd}@ict.ac.cn,qitian@cs.utsa.edu

Abstract

Image caption is becoming important in the field of artificial intelligence. Most existing methods based on CNN-RNN framework suffer from the problems of object missing and misprediction due to the mere use of global representation at image-level. To address these problems, in this paper, we propose a global-local attention (GLA) method by integrating local representation at object-level with global representation at image-level through attention mechanism. Thus, our proposed method can pay more attention to how to predict the salient objects more precisely with high recall while keeping context information at image-level cocurrently. Therefore, our proposed GLA method can generate more relevant sentences, and achieve the state-of-the-art performance on the well-known Microsoft COCO caption dataset with several popular metrics.

Introduction

Recently, image description has received much attention in the field of computer vision. It is a high-level and complicated task which involves computer vision and natural language processing technologies. Generating a meaningful description requires that the algorithm not only recognizes objects contained in an image, but also obtains the relationships among these objects, their attributes and activities, and then describes these semantic information with natural language.

So far, many pioneering approaches have been proposed for image caption. They can be broadly divided into three categories according to the way of sentence generation (Jia et al. 2015): template-based method, transfer-based method and neural network-based method. Template-based method (Farhadi et al. 2010; Yang et al. 2011; Kulkarni et al. 2013) firstly recognizes the objects contained in an image, their attributes and their relationships by using several classifiers respectively. Then it uses a rigid sentence template to form a complete sentence. This kind of method is simple and intuitive, but is not enough flexible to generate meaningful sentence due to the limitation of sentence template. Transfer-based method (Kuznetsova et al. 2012; Mason and Charniak 2014; Ordonez et al. 2015) utilizes image retrieval approaches to obtain similar image and then directly transfers the description of the retrieved image to the query image. This kind of method can generate more grammatically correct and natural sentences. However, the generated sentences may not correctly express the visual content of query image. Inspired by the recent advances of neural network in image recognition (Simonyan and Zisserman 2014; Ren et al. 2015) and machine translation (Bahdanau, Cho, and Bengio 2015; Cho et al. 2014; Sutskever, Vinyals, and Le 2014), neural network-based method (Xu et al. 2015; Vinyals et al. 2015; Jia et al. 2015; Mao et al. 2015; Karpathy and Fei-Fei 2015; Donahue et al. 2015) has been rapidly applied to image caption and has been made great progress. This kind of method is primarily based on the Convolutional Neural Network (CNN)-Recurrent Neural Network (RNN) framework which first extracts the image-level feature using CNN, and then utilizes a language model, RNN or its variants, such as Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU) et al., to generate meaningful sentences.

Baseline: A group of people standing on top of a snow covered slope.
Ours: A group of people on **skis** standing on a snow covered slope.

Baseline: A group of young men playing a game of soccer.
Ours: A boy and a **child** fly **kites** in a field.

Figure 1: Illustration of problems of the object missing (left: missing "skis") and misprediction (left: mispredicting "kite" as "soccer"). The baseline caption is generated by LRCN (Donahue et al. 2015). Our caption is generated by utilizing the proposed GLA method.

*This work was supported in part by National Nature Science Foundation of China (61525206, 61572472, 61429201), Beijing Natural Science Foundation (4152050), Beijing Advanced Innovation Center for Imaging Technology (BAICIT-2016009), and in part to Dr. Qi Tian by ARO grant W911NF-15-1-0290 and Faculty Research Gift Awards by NEC Laboratories of America and Blippar.

†Corresponding author

Compared with the previous two methods, neural network-based method can generate more coherent and relevant sentences thanks to the ability of capturing dynamic temporal information of RNN and the good representation ability of CNN.

However, most of the existing neural network-based methods only utilize global features at image-level with which some objects may not been detected. Therefore, it can cause the problem of object missing when generating image description. As shown in the left picture of Fig. 1, the "skis" is missed. Besides, global features are extracted at a coarse level which may result in incorrect recognition and can cause the problem of object misprediction during the process of description generation. As shown in the right picture of Fig. 1, the "kite" is mispredicted as "soccer". In order to make description more accurate, we take advantage of local features at object-level to address the problem of object missing. Moreover, we integrate local features with global features to reserve context information to address the problem of misprediction.

The main contribution of this paper is that we propose a global-local attention (GLA) method for image caption. Our proposed GLA method can selectively focus on semantically more important regions at different time while keeping global context information through integrating local features at object-level with global features at image-level via attention mechanism. The proposed GLA method achieves the state-of-the-art performance on Microsoft COCO caption datasets with different evaluation metrics in our experiments.

Related Work

Our proposed GLA method is based on neural network and attention mechanism, so we mainly introduce the related work about image caption with them.

Neural network-based image caption. Inspired by the successful application of neural network in image recognition and machine translation, several methods (Xu et al. 2015; Vinyals et al. 2015; Jia et al. 2015; Mao et al. 2015; Karpathy and Fei-Fei 2015; Donahue et al. 2015) have been proposed for generating image description based on neural network.

These approaches directly translate an image to a sentence by utilizing the encoder-decoder framework (Cho et al. 2014) introduced in machine translation. This paradigm first uses a deep CNN to encode an image to a static representation, and then uses a RNN to decode the representation to a meaningful sentence which can well describe the content of the image.

Mao et al. (Mao et al. 2015) propose a multimodal RNN (m-RNN) for image description. NIC (Vinyals et al. 2015) has been proposed to automatically generate image description with an end-to-end model by combining deep CNN with LSTM. Karpathy et al. (Karpathy and Fei-Fei 2015) propose a bidirectional RNN model to align segments of sentences with the regions of the image that they describe, and a mulitmodal RNN model to generate description of an image. Jia et al. (Jia et al. 2015) propose gLSTM, an alternative extension of LSTM, to guide LSTM to generate descriptions by using semantic information of an image. Donahue et al. (Donahue et al. 2015) propose Long-term Recurrent Convolutional Networks (LRCNs) which combines convolutional layers and long-range temporal recursion for visual recognition and description.

However, all the above mentioned approaches encode the whole image to a global feature vector. Therefore, these methods may suffer from the problems of object missing and misprediction as shown in Fig. 1. To address these problems, we propose a GLA method which integrates image-level features with object-level features for image description instead of only using the global features.

Attention mechanism in image caption and machine translation. Attention mechanism has been proved to be effective and important in the field of computer vision (Xu et al. 2015; You et al. 2016; Yao et al. 2015; Jin et al. 2015) and natural language processing (Bahdanau, Cho, and Bengio 2015). Bahdanau et al. (Bahdanau, Cho, and Bengio 2015) exploit BRNN to align a source sentence with the corresponding target sentence. The proposed method can automatically (soft-)search the parts of a source sentence that are most relevant to a target word. Xu et al. (Xu et al. 2015) explore two attention-based image caption methods, soft-attention and hard-attention, and analyze how attention mechanism works for descriptions generation. Yao et al. (Yao et al. 2015) exploit a temporal attention mechanism to capture global temporal structure among video frames based on soft-alignment method introduced in (Bahdanau, Cho, and Bengio 2015). The temporal attention mechanism makes the decoder selectively focus on some key frames which are most relevant to the predicted word in some degree. ATT (You et al. 2016) first utilizes different approaches (k-NN, multi-label ranking and Fully Convolutional Network) to obtain semantic concept proposals, and then integrates them into one vector through attention mechanism to guide language model for description generation.

To address the aforementioned problem of object missing and misprediction, our proposed GLA method integrates local representation at object-level with global representation through attention mechanism, which is sufficiently different from soft/hard attention (Xu et al. 2015), Yao et al. (Yao et al. 2015). These methods use only global frame-level features which cannot avoid the problem of object missing and misprediction. Instead of considering semantic concepts or attributes used in ATT (You et al. 2016), we directly apply image visual feature with attention mechanism. RA (Jin et al. 2015) proposes a complicated pipeline to obtain important regions from selective search region proposals (Uijlings et al. 2013) and combines them with scene-specific contexts to generate image caption. Compared with ATT and RA methods, our GLA method is simpler and the performance is much better than RA method.

Global-local Attention Model

In this section, we describe our global-local attention (GLA) method for image caption in details. An overview of our image caption pipeline is shown as Fig. 2. Our proposed GLA approach consists of the following three processes:

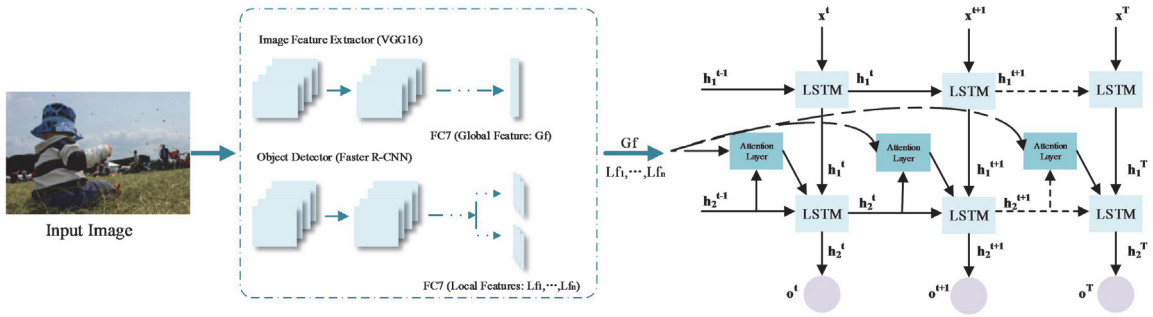

Figure 2: Illustration of our proposed image caption framework based on global-local attention mechanism.

(a) Global and local features extraction: Obtaining global feature and local features of an input image using deep CNNs.

(b) Global-local attention: Integrating local features with global feature through attention mechanism.

(c) Image description generation: Generating a sentence to describe the content of an input image by using LSTM.

Global and Local Features Extraction

Global features and local features are important for representing an image. Global features usually contain the context information around objects, and local features always contain the fine-grained information of objects. Therefore, in this paper, we explore both features for describing the content of an image. Benefitting from the powerful representation of CNNs, image classification and object detection (Ren et al. 2015) have made great progress. In this paper, we extract global feature with VGG16 (Simonyan and Zisserman 2014) and local features with Faster R-CNN (Ren et al. 2015).

We extract global feature from fc7 layer of VGG16 net, a 4096-dimension vector denoted as Gf. The VGG16 net is trained on ImageNet classification dataset. For local features denoted as $\{Lf_1, ..., Lf_n\}$, we select top-n detected objects to represent important local objects according to their class confidence scores obtained from Faster R-CNN. Then, we represent each object as a 4096-dimension CNN feature vector extracted from fc7 layer for each object bounding box. The Faster R-CNN model is pre-trained on ImageNet classification dataset and then fine-tuned on the MS COCO detection dataset. Therefore, each image can be finally represented as a set of 4096-dimension vectors I=$\{Gf$,Lf_1,...,$Lf_n\}$. In our experiments, we set n to 10 since the number of object contained in an image is usually below 10.

Global-local Attention Mechanism

How to integrate the local features with global features is important for describing images. In our proposed method, we adopt attention mechanism to fuse these two kinds of features according to the following Eq. 1:

$$\Psi^{(t)}(I) = \alpha_0^{(t)} Gf + \sum_{i=1}^{n} \alpha_i^{(t)} Lf_i, \tag{1}$$

where $\alpha_i^{(t)}$ denotes the attention weight of each feature at time t and $\sum_{i=0}^{n} \alpha^{(t)}$=1.

This mechanism dynamically weights each feature by assigning it with one positive weight $\alpha_i^{(t)}$ along with the sentence generation procedure. Through this manner, our method can selectively focus on some salient objects at different time and consider their context information at the same time.

The attention weight $\alpha_i^{(t)}$ measures the importance degree of each feature at time t and the relevance of each feature to the previous information. Thus, it can be computed based on the previous information and each feature $f_i \in \{Gf, Lf_0, ..., Lf_n\}$ with the following equations:

$$\beta_i^{(t)} = w^T \varphi(W_h h^{(t-1)} + W_o f_i + b), \tag{2}$$

$$\alpha_i^{(t)} = \frac{\beta_i^{(t)}}{\sum_{j=0}^{n} \beta_j^{(t)}}, \tag{3}$$

where $\beta_i^{(t)}$ denotes the relevance score of feature f_i with the previous generated words.

The attention weight $\alpha_i^{(t)}$ is obtained by normalizing $\beta_i^{(t)}$ with softmax regression. $h^{(t-1)}$ is the previous hidden state output which will be introduced in the next section. W, W_h, W_o and b are the parameters to be learned by our model and shared by all the features at all the time steps. φ is activation function. Here, we use the element-wise Hyperbolic Tangent function. Compared with existing image/video caption methods with using frame-level features, the way of using global-local features can capture both salient objects and context information. This makes our model better describe the context of a given image.

Image Description Generation

Inspired by the good performance of RNN for capturing dynamic temporal information in neural machine translation (Bahdanau, Cho, and Bengio 2015), we use a stacked two-layer LSTM with global-local attention for image caption.

LSTM is an advanced RNN with distinctive unit to manipulate the long-term information.

The right figure of Fig. 3 shows that the basic RNNs essentially maintain a state h_t in time t and overwrite the state h_t with input x_t and state h_{t-1} which reserves the history information. When training the basic RNNs with Back-Propagation Through Time (BPTT) algorithm, the gradients of RNNs tend to vanish because of the chain rule of derivative. Therefore, the basic RNNs suffer from the long-range dependency problem caused by vanishing and exploding gradients in the learning process. The LSTMs are designed to combat these problems through a gating mechanism. As the left figure of Fig. 3 shows, a LSTM unit consists of a memory cell, and three gates (input gate, output gate and forget gate). The forget gate decides what information we should discard or persist. It puts the input x_t and the previous hidden state h_{t-1} into an activation function and determines how to handle the previous information of cell c_{t-1}. Input gate also takes x_t and h_{t-1} as input, but it has a different function to update the memory cell. With these two gates, the state of memory cell has been changed. Finally, LSTM unit employs output gate by taking the same input with other two gates to get result based on cell state.

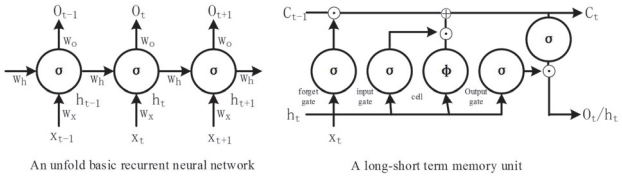

Figure 3: Illustration of a un-fold basic RNN and a LSTM unit. σ represents logic sigmoid function. ϕ represents hyperbolic tangent function. $\bigodot$ represents the multiplication operation and $\oplus$ represents sum operation.

The LSTM is trained to predict word s_t of the description on the condition that it has known the image visual information I as well as all predicted words $\{s_0, s_1, ..., s_{t-1}\}$ which is defined by $p(s_t|I, s_0, s_1, ..., s_{t_1})$. Specifically, the integrated image feature is only used as the input of the second LSTM layer. The detailed operations of our second LSTM layer is as follows:

$$\begin{aligned} x^t &= w_x s_t, I^t = \Psi^t(I) \\ i^t &= \sigma(w_{is}x^t + w_{iI}I^t + w_{ih}h^{(t-1)} + b_i) \\ f^t &= \sigma(w_{fs}x^t + w_{fI}I^t + w_{fh}h^{(t-1)} + b_f) \\ o^t &= \sigma(w_{os}x^t + w_{oI}I^t + w_{oh}h^{(t-1)} + b_o) \\ c^t &= f^t \otimes c^{t-1} \oplus i^t \otimes \phi(w_{cs}x^t + w_{ch}h^{t-1}) \\ h^t &= o^t \otimes c^t \\ P(s_t|I, s_0, s_1, ..., s_{t-1}) &= Softmax(w_p h^t) \end{aligned} \tag{4}$$

Where $\Psi^t(I)$ is the image representation defined in Eq. 1 at time *t*. w_* and b_* are the parameter learned by our model and shared by all time steps. Each word is represented as an one-hot vector s_t whose dimension is equal to the vocabulary size.

Our goal is to predict the probability of observing the sentence defined as Eq. 5. In summary, the probability of a sentence is the product of the probability of each word given the image and the words before current time.

$$p(s_0, s_1, ..., s_m) = \prod_{i=0}^{m} p(s_i|I, s_0, .., s_{i-1}) \tag{5}$$

Thus, we define the loss by using the sum of the log likelihood of the correct word at each time step. The optimization can be formed as Eq. 6.

$$L(I, S) = \sum_{i=0}^{m} log(p(s_t|I, s_0, ..., s_i)) \tag{6}$$

The above objective function is optimized over the whole training caption set by using stochastic gradient descent with a momentum of 0.9. The learning rate is initially set to 0.01 and then is decreased by step. For sentence generation, there are two strategies for sampling sentence of a given image. The first approach is essentially a greedy method in which we sample the next word with the maximum probability from the probability distribution at each time step until the end sign word is sampled or the maximum sentence length is reached. The other approach is beam search method which selects the top-k best sentences at each step and then generates new best top-k sentences based on the previous top-k sentences. In this paper, we evaluate our method by using these two sentence generation approaches respectively. Particularly, we can obtain the best run when the value of k is set to 3.

Experiments

In this section, we introduce our experiments and analyze the results in details. We implement our global-local attention model based on LRCN framwork (Donahue et al. 2015), an open-source implementation of RNN. We prove the effectiveness of our model through extensive comparison experiments with several popular metrics.

Evaluation Metrics

Various methods for the evaluation of generated sentences have been employed. However, how to evaluate the quality of the descriptions is also challenging. Thus, in order to verify the performance of our model, we use multiple metrics to evaluate the proposed GLA approach, *i.e*. BLEU-1,2,3,4 (Papineni et al. 2002), METEOR (Banerjee and Lavie 2005), CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015) and ROUGE-L (Lin 2004).

BLEU is the most popular metric for the evaluation of machine translation which is only based on the n-gram precision. Here, we choose 1,2,3,4-gram to validate the performance. METEOR is based on the harmonic mean of unigram precision and recall with which the recall is weighted higher than precision. It is designed to fix some of the problems of BLEU metric. Different with the BLEU metric, the METEOR seeks correlation at the corpus level. CIDEr is designed for evaluating image descriptions using human consensus. ROUGE-L is used to measure the common subse-

Table 1: Comparison experiments on exploiting global feature, local features and fusion of the two features.

Method	Bleu1	Bleu2	Bleu3	Bleu4	METEOR	CIDEr	ROUGE-L
GlobFeat	67.3	49.1	34.4	24.0	22.2	76.9	49.2
LocAtt	66.3	47.5	33.1	22.9	21.6	73.6	47.8
GloLocAtt	69.7	51.7	37.1	26.3	23.8	85.7	51.4

Table 2: Comparison experiments to exploring the effect of dropout mechanism.

Method	Bleu1	Bleu2	Bleu3	Bleu4	METEOR	CIDEr	ROUGE-L
GloLocAttEmb	70.1	52.4	37.7	26.6	23.7	87.3	51.4
GloLocAttEmb+OneDrop	71.0	53.4	38.6	27.6	23.8	89.2	51.5
GloLocAttEmb+TwoDrop	71.8	54.3	39.5	28.6	24.2	91.2	52.3

quence with maximum length between target sentence and source sentence.

Datasets

We conduct experiments on the well-known MS COCO caption dataset, a popular large scale dataset. This dataset contains 82,783 images for training and 40,504 images for validation. Each image is associated with 5-sentence annotated in English by AMT workers. Compared with other existing image caption dataset, such as, flickr8k and flickr30k, COCO dataset has much more images and annotations for both training and testing. Therefore, we only choose COCO dataset in our experiments. In order to fairly compare with existing methods, we keep the same splits as the previous work (Karpathy and Fei-Fei 2015) - 5,000 images for validation and another 5,000 images from validation for testing.

Experiments and Results

Evaluation on Image-level and Object-level Information for Image Description. Global features and local features are important for image recognition task. In this section, we conduct three experiments to test the effect of global feature, local feature and fusion feature for image description. The configurations of the model are listed as follows:

- **GloFeat:** Only using image-level feature Gf extracted from VGG16 to generate description.
- **LocAtt:** Only using object-level features $\{Lf_1, ..., Lf_n\}$ extracted from Faster R-CNN with attention mechanism to generate description.
- **GloLocAtt:** Integrating GF and $\{Lf_1, ..., Lf_n\}$ with attention mechanism to generate description.

The results of the three experiments are shown in Tab. 1. Through comparing the result of GloFeat with LocATT, we find that it is better to use global feature than only use local features. Our conjecture is that we choose only the most important objects. Perhaps there are some less important objects which can not be ignored. In our experiments, we choose top-n (n=10) detected objects by Faster R-CNN. By comparing GloLocAtt with the other two runs, we find that integration of local features with global feature can achieve best performances. The reason is that the GloLocAtt model captures both local object information and global context information.

Evaluation on Dropout Mechanism for Our Proposed Method. When we train our two-layer LSTM language model with global-local attention mechanism, we note that there would be overfitting which does not appear in experiments with only using global features. Dropout is an import mechanism for regularizing deep network to reduce overfitting. As introduced in (Zaremba, Sutskever, and Vinyals 2014), we employ dropout in our model and explore the performance with different forms. Besides, we also add one linear transform layer to reduce the integrated 4096-dimension feature to 1000-dimension to keep consistent with the dimension of LSTM hidden layer which is denoted with "Emb". Therefore, we conduct three experiments in this section as follows:

- **GloLocAttEmb:** Adding a linear transform layer to reduce the feature dimension.
- **GloLocAttEmb+OneDrop:** Adding one dropout layer after the second LSTM layer.
- **GloLocAttEmb+TwoDrop:** Adding one dropout layer after the first LSTM layer and one dropout layer after the second LSTM layer, respectively.

Tab. 2 shows the comparison results. Compared with GloLocAtt, GloLocAttEmb improves the performance slightly in that the linear transform layer makes the feature more distinctive. Through comparing the dropout experiments with GlobLocAttEmb experiment, we note that dropout can reduce the overfitting in some degree and it is better by adding two dropout layers.

Comparison with the State-of-the-art Methods. We finally compare the proposed GLA method with several state-of-the-art methods: NIC (Vinyals et al. 2015), LRCN (Donahue et al. 2015), m-RNN (Mao et al. 2015), soft/hard attention (Xu et al. 2015), g-LSTM (Jia et al. 2015), DeepVS (Karpathy and Fei-Fei 2015) and ATT (You et al. 2016). We show our results in Tab. 3. "GLA" represents that the model configuration is same with the previous "GloLocAttEmb+TwoDrop" model. GLA samples sentence with greedy method. Since beam search is a heuristic search strategy which can approximately maximize the probability of generated sentence, we try beam search in our experiments, and finally get the best result when the k is set to 3. The best run is denoted as "GLA+BEAM3".

Baseline: A group of birds are standing in the water.
Ours: A group of ducks swimming in a lake.

Baseline: A group of people playing a game of frisbee.
Ours: A group of people playing soccer on a field.

Baseline: A man is riding a elephant with a trunk.
Ours: A man is standing next to a large elephant.

Baseline: A bear is walking through a tree in the woods.
Ours: A bear is sitting on a tree branch.

Figure 4: The sample images and their descriptions. The original caption is generated without object attention by LRCN. The new caption is generated with our GLA.

Table 3: Comparison with several state-of-the-art models in terms of BLEU-1,2,3,4, METEOR, CIDEr, ROUGE-L and METEOR over MS COCO dataset. - indicates unknown scores. † indicates that the model has the same decoder with ours, that is, the same CNN model for image representation. $*$ indicate that the model has the same encoder - the language model for generating sentence description with ours.

Method	Bleu1	Bleu2	Bleu3	Bleu4	METEOR	CIDEr	ROUGH-L
NIC (Vinyals et al. 2015)	66.6	46.1	32.9	24.6	-	-	-
LRCN $*$ (Donahue et al. 2015)	62.79	44.19	30.41	21	-	-	-
DeepVS † (Karpathy and Fei-Fei 2015)	62.5	45	32.1	23	19.5	66	-
m-RNN † (Mao et al. 2015)	67	49	35	25	-	-	-
soft attention † (Xu et al. 2015)	70.7	49.2	34.4	24.3	23.9	-	-
g-LSTM Gaussian (Jia et al. 2015)	67	49.1	35.8	26.4	22.74	81.25	-
(RA+SF)-GREEDY $*$† (Jin et al. 2015)	69.1	50.4	35.7	24.6	22.1	78.3	50.1
(RA+SF)-BEAM10 $*$† (Jin et al. 2015)	69.7	51.9	38.1	28.2	23.5	83.8	50.9
ATT (You et al. 2016)	70.9	53.7	40.2	30.4	24.3	-	-
GLA (ours) $*$†	71.8	54.3	39.5	28.5	24.2	91.2	52.3
GLA-BEAM3 (ours) $*$†	**72.5**	**55.6**	**41.7**	**31.2**	**24.9**	**96.4**	**53.3**

For these comparison methods, there are some differences with each other. The first difference is the encoder for images representation. NIC, g-LSTM and ATT use GoogLeNet, to obtain image-level features. LRCN exploits AlexNet to extract image-leval features. DeepVS, m-RNN and soft/hard attention utilize the VGG16 as the same with our model to get image-level representation. To make fair comparison, we first compare our method with these methods which use VGG16 encoder, and find that our method has significant improvement on different metrics due to the stacked two-layer LSTM and the use of global-local features.

The second difference is the decoder structure for generating sentence descriptions. NIC, g-LSTM, and soft/hard attention use the LSTM network as language model to generate image sentence description. ATT and m-RNN exploit the basic RNN as decoder. As same with our model, LRCN uses a stacked two-layer LSTM to translate image to sentence description. DeepVS employs a BRNN to obtain the image caption. Here, as for the same decoder, since we use integrated global-local feature to generate image caption, our model provides more distinctive features so as to have better performances.

By comparing our GLA model with the existing methods, we note that our approach achieves best performance. From Fig. 4 which illustrates the sample images and their descriptions generated by LRCN (Donahue et al. 2015) model and GLA model, we can see that GLA model can generate more relevant descriptions. The results show our method can solve the problems of objects missing and misprediction in some degree.

Conclusion

This paper proposes a novel method via combining image-level and object-level information through attention mechanism with the encoder-decoder framework for image description, which achieves better performance on the MS COCO benchmark compared with the previous approaches. Compared with existing methods, our method not only captures the global information, but also obtains local object information. Consequently, our method generates more relevant and coherent natural language sentences which can describe the context of images.

However, our current GLA is not end-to-end. Thus, we will try how to integrate the object detector with image caption so as to train and test our model end-to-end.

References

Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural machine translation by jointly learning to align and translate.

ICLR.

Banerjee, S., and Lavie, A. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, volume 29, 65–72.

Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *EMNLP*.

Donahue, J.; Anne Hendricks, L.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Saenko, K.; and Darrell, T. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2625–2634.

Farhadi, A.; Hejrati, M.; Sadeghi, M. A.; Young, P.; Rashtchian, C.; Hockenmaier, J.; and Forsyth, D. 2010. Every picture tells a story: Generating sentences from images. In *European Conference on Computer Vision*, 15–29. Springer.

Jia, X.; Gavves, E.; Fernando, B.; and Tuytelaars, T. 2015. Guiding long-short term memory for image caption generation. *arXiv preprint arXiv:1509.04942*.

Jin, J.; Fu, K.; Cui, R.; Sha, F.; and Zhang, C. 2015. Aligning where to see and what to tell: image caption with region-based attention and scene factorization. *arXiv preprint arXiv:1506.06272*.

Karpathy, A., and Fei-Fei, L. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3128–3137.

Kulkarni, G.; Premraj, V.; Ordonez, V.; Dhar, S.; Li, S.; Choi, Y.; Berg, A. C.; and Berg, T. L. 2013. Babytalk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35(12):2891–2903.

Kuznetsova, P.; Ordonez, V.; Berg, A. C.; Berg, T. L.; and Choi, Y. 2012. Collective generation of natural image descriptions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, 359–368. Association for Computational Linguistics.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, volume 8. Barcelona, Spain.

Mao, J.; Xu, W.; Yang, Y.; Wang, J.; Huang, Z.; and Yuille, A. 2015. Deep captioning with multimodal recurrent neural networks (m-rnn). *ICLR*.

Mason, R., and Charniak, E. 2014. Nonparametric method for data-driven image captioning. In *ACL (2)*, 592–598.

Ordonez, V.; Han, X.; Kuznetsova, P.; Kulkarni, G.; Mitchell, M.; Yamaguchi, K.; Stratos, K.; Goyal, A.; Dodge, J.; Mensch, A.; et al. 2015. Large scale retrieval and generation of image descriptions. *International Journal of Computer Vision* 1–14.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, 311–318. Association for Computational Linguistics.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Neural Information Processing Systems*.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *CoRR* abs/1409.1556.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, 3104–3112.

Uijlings, J. R.; van de Sande, K. E.; Gevers, T.; and Smeulders, A. W. 2013. Selective search for object recognition. *International journal of computer vision* 104(2):154–171.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4566–4575.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3156–3164.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhutdinov, R.; Zemel, R. S.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. *CoRR* 2(3):5.

Yang, Y.; Teo, C. L.; Daumé III, H.; and Aloimonos, Y. 2011. Corpus-guided sentence generation of natural images. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 444–454. Association for Computational Linguistics.

Yao, L.; Torabi, A.; Cho, K.; Ballas, N.; Pal, C.; Larochelle, H.; and Courville, A. 2015. Describing videos by exploiting temporal structure. In *Proceedings of the IEEE International Conference on Computer Vision*, 4507–4515.

You, Q.; Jin, H.; Wang, Z.; Fang, C.; and Luo, J. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Zaremba, W.; Sutskever, I.; and Vinyals, O. 2014. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*.