The Accurate Guidance for Image Caption Generation

Xinyuan Qi, Zhiguo Cao(✉), Yang Xiao, Jian Wang, and Chao Zhang

National Key Lab of Science and Technology of Multispetral
Information Processing, School of Automation,
Huazhong University of Science and Technology, Wuhan 430074, Hubei, China
{silliam_qi, zgcao, Yang_Xiao, M201572352,
zhangC_22}@hust.edu.cn

Abstract. Image caption task has been focusing on generating a descriptive sentence for a certain image. In this work, we propose the accurate guidance for image caption generation, which guides the caption model to focus more on the principle semantic object while making human reading sentence, and generate high quality sentence in grammar. In particular, we replace the classification network with object detection network as the multi-level feature extracter to emphasize what human care about and avoid unnecessary model additions. Attention mechanism is utilized to align the feature of principle objects with words in the semantic sentence. Under these circumstances, we combine the object detection network and the text generation model together and it becomes an end-to-end model with less parameters. The experimental results on MS-COCO dataset show that our methods are on part with or even outperforms the current state-of-the-art.

Keywords: Image caption · Object detection · Attention mechanism
Deep learning

1 Introduction

Image caption task aims at automatically generating a descriptive sentence to describe the content of an image with an English sentence [1]. With the explosive increase in digital images and the rapid development in deep learning, teaching machines to understand images as humans is drawing great interests. At the outset, computer vision task aims at classifying the category of a single image (image classification). Hereafter, researchers try to locate the position of objects in more complicated scenes (object detection). After that, researchers further want to distinguish the category of per-pixel (semantic segmentation). Along with this fruitful development route, researchers owe it to comprehending the semantic information of the picture better and better. Meanwhile, another understanding of the images' semantic information is to describe an image's content with a human-like sentence (image caption). This idea is closer to human's habit when there is a scene in front of their eyes. While caption task seems obvious for human beings, it is much more difficult for machine since it requires the 'translation' model to capture several semantic information from a certain image. Such as scenes,

J.-H. Lai et al. (Eds.): PRCV 2018, LNCS 11258, pp. 15–26, 2018.
https://doi.org/10.1007/978-3-030-03338-5_2

objects, attributes, relative position and so on. Another challenge of caption task is to generate descriptive sentence meeting the grammar rules.

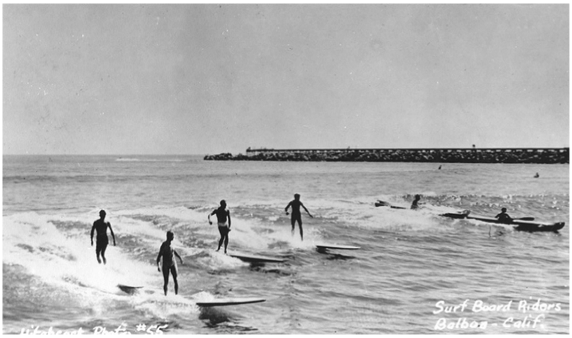


Fig. 1. This is an example picture in MS-COCO dataset. The caption ground truth is “Several surfers riding a small wave into the beach”. The proportion of principal object (humans and surfboards) is well low. There are too much redundant information, such as sky, which will make it harder for attention mechanism to align the principal object with the noun composition in the descriptive sentence.

Recently, Neural network methods [2, 3] dominate the literature in image captioning. The encoder-decoder architecture in Neural Machine Translation [4] inspire these methods very much. In contrast to original Neural Machine Translation model, image caption model replace the recurrent neural network (RNNs) with convolutional neural network (CNNs) as encoder. CNNs encode the input image into a feature vector, which represents the semantic information of the image. Then a sequence modeling approach (e.g., Long Short-Term Memory (LSTM) [5]) decodes the semantic feature vector into a sequence of words. Such architecture applies to the vast majority of image caption model.

The method to combine CNNs and RNNs together directly will result that the information of the input image decreases by iterations. In this situation, researchers start to utilize image guidance [3], attributes [6] or region attention [7] as the extra input into LSTM decoder for better performance. The original intention comes from visual attention, which has been known in Psychology and Neuroscience for a long time. Attention mechanism highly relies on the quality of the input image. If there are too much redundant information in the image, it will be hard for attention mechanism to capture the principal information. As shown in Fig. 1, the proportion of principal objects (humans and surfboards) is very low. CNNs-encoder usually reduce the dimension of feature vector a lot, which will make it harder for attention mechanism to capture the information for subject, object and other noun composition. In this condition, if we insist on applying attention mechanism to the whole image like [7], caption model may not know what to describe.

In Natural Language Processing, scientists take the noun composition in a sentence as the focus, which people care more about. In image caption task, the noun composition corresponds to the principal object in an image. To help image caption model to capture the principal object more accurate, we propose to get help from object detection task. Object detection task has been studied for a long time. CNNs framework is widely used and rapidly developed in object detection task, such as R-CNN [11], Fast-RCNN [12], Faster-RCNN [13]. These models are able to capture principal objects in the image very well. So we propose to make use of the feature of object detection methods to encode the image and generate guidance for the language generate model. We call it as accurate guidance. This advance also means to combine the higher level of semantic information in computer vision task with the semantic meaning in human-reading sentence.

We implement our model based on a single state-of-the-art object detection network Faster-RCNN [13], for accuracy and speed. Simultaneously, our model can be trained end-to-end, which will make the object detection module to adjust itself to suit for the image caption task. We take the Google NIC [7] as the baseline and compare our methods with popular attention models on the commonly used MS-COCO dataset [9] with publicly available splits of training, validation and testing sets. We evaluate methods on standard metrics. Our proposed methods outperform all of them and achieve state-of-the-art across different evaluation metrics.

The main contributions of our paper are as follows. First, we propose accurate guidance mechanism to help the caption model capture the principal object more precisely and infer their relationships from global information simultaneously. Second, the proposed method utilize a single object detection network as the multi-level feature extracter and demonstrates a less complicated way to achieve end-to-end training of attention-based captioning model, whereas state-of-the-art methods [3, 6, 19] involve LSTM hidden states or image attributes for attention, which compromises the possibility of end-to-end optimization.

2 Related Work

Recent successes of deep neural networks in machine translation catalyze the adoption of neural networks [8] in solving image caption problems. Early works of neural networks-based image caption include the multimodal RNN [10] and LSTM [5]. In these methods, neural networks are used to both image-text embedding and sentence generating.

Attention mechanism has recently attracted considerable interest in LSTM-based image captioning [3, 6]. Xu et al. [7] proposed to integrate visual attention through the hidden state of LSTM model. You et al. [6] propose to fusion visual attributes extracted from images with the input or output of LSTM. These methods achieve state-of-the-art performance but they highly rely on the quality of the pre-specified visual attributes. Our method also use attention mechanism. Different from the predecessors, we consider the object detection-dependent attention to generate high quality guidance rather than search at the whole noisy image. It is an adaptive method to obtain high quality features.

Reinforcement Learning has recently been introduced into image caption task [20] and achieved state-of-the-art performance due to optimize the evaluation metrics directly. These methods are generally applicable training approach not the improvement for the caption model. Thus, we don't compare with them but believe that our model will gain much higher performance with Reinforcement Learning.

[19] first proposes to utilize object detection method in image caption task. However, it utilize Fast-RCNN to detect and VGG net [15] to locate. The caption model is very redundant. While generating guidance, it keep the region of its bounding box unchanged and set remaining regions to mean value of the training set for each object in image. This process will bring much interference to the caption model. Our method solves these puzzles by taking the single object detection network as the multi-level feature extracter. In this way, our method is a clean architecture for the ease of end-to-end learning.

3 Methods

Our accurate guidance model includes a multi-level feature extraction module (MFEM) and a principal object guiding LSTM (po-gLSTM). Figure 2 shows the structure of our model. We first describe how to use object detection network as MFEM to simultaneously extract the features of the whole image (fea_w) and principle objects (fea_o) in Sect. 3.1. Then, we introduce our po-gLSTM which will take advantage of the multi-level feature to guide the LSTM to describe the image more precise in Sect. 3.2.

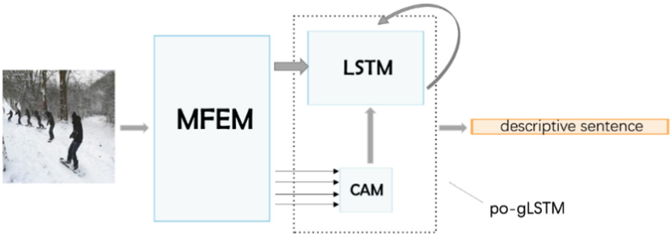


Fig. 2. The structure of our accurate guidance model

3.1 Multi-level Feature Extraction Module

Figure 3 shows the framework of multi-level feature extraction module. The MFEM consists of two parts: (1) fea_o extraction network (above the red dotted line); (2) fea_w extraction network (below the red dotted line). It is a variant of Faster-RCNN [13]. In order to capture the principle objects better, for an input image I, we suppose to utilize object detection network to find the potential objects and extract fea_o, which denoted as $fea_o = \{obj_1, \ldots, obj_N\}$ and formulated as formula (1). N is the number of potential objects. RPN (Region Proposal Network) splits the principle object parts from the whole image. $CNN_{\theta 2}$ is to further extract the features after RPN.

$$fea_0 = CNN_{\theta 2}\{RPN[CNN_{\theta 1}(I)]\} \quad (1)$$

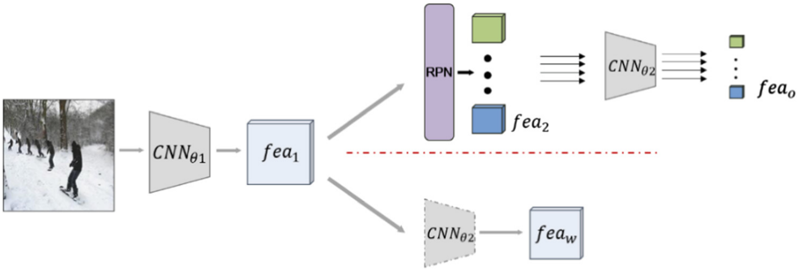


Fig. 3. The structure of the MFEM (Color figure online)

Simultaneously, we also need fea_w so that the po-gLSTM can get the information of scenes and infer the relationship between objects. In this situation, the original output of Faster-RCNN framework does not meet the conditions. Thus, we try to fix it's framework so that it can extract fea_w at the same time. As shown in the part below the red dotted line in Fig. 3, we get a copy of the feature after $CNN_{\theta 1}$ and take it into $CNN_{\theta 2}$ directly. Then we get an imitation classification network followed with fea_w, formulated as formula (2).

$$fea_w = CNN_{\theta 2}[CNN_{\theta 1}(I)] \tag{2}$$

Notice that the $CNN_{\theta 2}$ with dotted border (below the red dotted line) is the same with the $CNN_{\theta 2}$ with solid border (above the red dotted line). We do not increase the model parameters but obtain fea_w successfully. Faster-RCNN argues the size of input image should be larger than 600 pixel × 600 pixel. For reducing the model parameters, we replace its' fully connected layer with the Global Average Pooling layer to embedding fea_w and resize it to fit the size of the principle object guiding LSTM's input, formulated as follows:

$$x_0 = Pool_{ave}(fea_w) \tag{3}$$

x_0 is utilized to initialize decoder in Sect. 3.2. Here, we have already gotten the multi-level feature of the input image. The multi-level feature carries the multi-level semantic information. As later experiments will demonstrate, multi-level feature extraction module will help the model to focus more on the principle objects and achieve better performance.

3.2 Principal Object Guiding LSTM (po-gLSTM)

As shown in Fig. 4, the function of po-gLSTM is to decode the multi-level semantic information of the image and generate corresponding descriptive sentence. In this section, we will first introduce the condition attention module to obtain the principle object information for the current word. Then we will introduce how to make use of the principle object information to guide the LSTM to generate sentence. Both of above, we treat them as a whole and call it as po-gLSTM.

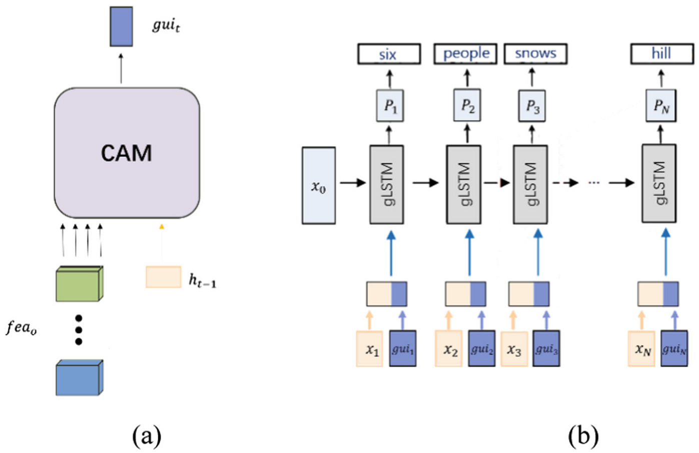


Fig. 4. (a) CAM is the condition Attention Module, which is to generate guidance information (gui_t) by principle object features (fea_0) and the information of hidden layer at previous step (h_{t-1}). (b) This sketch map shows how to utilize x_0 and gui_t to generate descriptive sentence. Both of (a) and (b) make up the po-gLSTM.

Condition Attention Module

With the multi-level feature extraction module, fea_o and fea_w of an input image will be extracted easily. Each word in caption is represented by a one-hot vector and the captioning sentence is a sequence of input vectors $(x_1, \ldots, x_T)$. Same as previous methods, we utilize fea_w to initialize the decoder (LSTM), the decoder then computes a sequence of hidden states $(h_1, \ldots, h_t)$ and a sequence of outputs $(y_1, \ldots, y_t)$. The primer decoder only accesses fea_w (encoded as x_0) once at the beginning of the learning process, which will loss most of the information of image I by iterations, and output incorrect words or stop too early. To avoid this, we proposed to utilize condition attention module (CAM) [6] to stress the role of principle objects and supply necessary information lost by iterations. CAM is formulated as followed:

$$a_t^i = Wtanh(W_{ao}obj_i + W_{ah}h_{t-1})i = 1, \ldots, N \tag{4}$$

$$\alpha_t = softmax(a_t) \tag{5}$$

$$gui_t = \sum\nolimits_{i=1}^{N} \alpha_t^i obj_i \tag{6}$$

W, W_{ao}, W_{ah} are learnable parameters. N is the number of principle object in an image. a_t^i is the relevance of obj_i and h_{t-1}. The elements of α_t is utilized to combine the guiding information (principle objects). gui_t is the guidance at iteration t.

With attention mechanism, model will know "where to see" while generating every word. We also make a visualization of attention mechanism to prove it in later experiment.

Guiding LSTM

The generated sentence by the LSTM model may lose track of the original image content since it only accesses the image content once at the beginning of the learning process, and forgets the image even after a short time. To make use of gui_t mentioned above and supplement the forgotten information if necessary, we propose to utilize an extension of the LSTM model, named the guiding LSTM (gLSTM) [3], which extracts semantic information from the input image and feeds it into the LSTM model every time step as extra information. Its' gate and memory cell can be formulated as follows:

$$i'_t = \sigma\left(W'_i\left[h'_{t-1}, x'_t, gui_t\right]\right) \tag{7}$$

$$f'_t = \sigma\left(W'_f\left[h'_{t-1}, x'_t, gui_t\right]\right) \tag{8}$$

$$o'_t = \sigma\left(W'_o\left[h'_{t-1}, x'_t, gui_t\right]\right) \tag{9}$$

$$\widetilde{C'_t} = tanh\left(W'_c\left[h'_{t-1}, x'_t, gui_t\right]\right) \tag{10}$$

$$C'_t = f'_t C'_{t-1} + i'_t \widetilde{C'_t} \tag{11}$$

$$h'_t = o'_t * tanh\left(C'_t\right) \tag{12}$$

$$x'_{t+1} = W'_{emb}\left(\log softmax\left(W'_h h'_t\right)\right) \tag{13}$$

Where W'_s denote learnable weighs, $*$ represent element-wise multiplication, $\sigma(\cdot)$ is the sigmoid function, $tanh(\cdot)$ is the hyperbolic tangent function, x'_t stands for input at t-th iteration, i'_t for the input gate, f'_t for the forget gate, o'_t for the output gate, C'_t for state of the memory cell, h'_t for the hidden state.

o'_t decides what to forget in C'_t. Its' decision is up to h'_{t-1} and x'_t. In original LSTM, when o'_t decides that forgetting some information is helpful for x'_{t+1}, it will be impossible for $x'_{t'}(t' > t+1)$ to utilize the forgotten information. The longer the descriptive sentence, the worse the condition like this is.

gLSTM is able to supplement the forgotten information if necessary. Condition attention module will also help to pick the most helpful principle object for x'_{t+1}. And we call our gLSTM with principle object condition attention module as op-gLSTM. Somebody may doubt weather emphasizing the principle object so much is helpful. Our experiment will verify that the model can infer the relationship better with stronger principle object information and it will cause no trouble for extracting the scene from fea_w.

One benefit of op-gLSTM is that it allows the language model to learn semantic attention automatically through the back-propagation of the training loss. While [19] only utilize objects and locations, other semantic information, such as scenes and motion relationship, is discarded.

4 Experiments

4.1 Dataset and Experiment Setup

Dataset

We use MS-COCO dataset [9] in our experiments. The dataset contains 123287 images respectively and each is annotated with 5 sentences using Amazon Mechanical Turk. There are 80 classes included in the dataset. We use 113287 images for training, 5000 images for validation and 5000 images for testing.

Experiment Setup

The inputting image is resized to 600 pixel × 600 pixel. The training process contains three stages: (1) pre-train the object detection network (Faster RCNN) on MS-COCO dataset. (2) combine the multi-level feature extract module (a variant of the pre-trained Faster RCNN) with our po-gLSTM and train the po-gLSTM to equip it with the ability to decode. (3) train the integral model end-to-end to help our multi-level feature extract module and po-gLSTM fusion better. Four standard evaluation metrics, e.g. BLUE, METEOR, ROUGE_L, and CIDER, are used evaluate the property of the generated sentence.

4.2 Comparison Between Different CNNs Encoders

Encoder is used to extract the semantic feature of the input image. The property of the extracted feature is decisive to our caption model. To explore which encoder is more proper, we use three different CNNs in our experiments, including 50-layer and 101-layer ResNets [14] and 16-layer VGGNet [15]. Table 1 shows the experimental result.

Table 1. Results of different CNNs encoders. All values are reported as percentage (%).

CNNs encoders	MS-COCO dataset						
	B1	B2	B3	B4	M	R	C
Ours-VGG16	70.9	53.1	38.4	27.4	23.5	51.3	88.0
Ours-RESNET50	72	54.4	39.8	28.9	24.1	52.3	90.8
Ours-RESNET101	**72.9**	**55.6**	**41.0**	**29.9**	**24.7**	**53.1**	**96**

The experimental results show that deeper CNNs achieves higher scores on all metrics. This indicates that deeper CNNs can capture better semantic features, which contain more and better information for descriptive sentence generation. The guidance of deeper CNNs is much more accurate.

4.3 Comparison to the State-of-the-Art

Several related models have been proposed in Arxiv preprints since the original submission of this work. We also include these in Table 2 for comparison.

Table 2 shows the comparison results. Our models, both VGG16-based and RESNET101-based, outperform other models at the same scale in most metrics by a large margin, ranging from 1% to 5%. Models with attention mechanism, such as ATT [6], Det+Loc [19] achieve better score than models without attention mechanism, such as NIC [7] and LRCN [16]. Det+Loc [19] also utilize the object detection network whose scores are better than the models with classification network. Notice that, our VGG16-based model gets comparable performance with FC-2 K [20] (Resnet-101 based). Meanwhile, our RESNET101-based model outperforms FC-2 K in all metrics. it's up to 5.1% in CIDER. Det+Loc is an object detection-based model, which utilize beam search (beam size 4) while testing. Without beam search, our VGG16-based model outperforms it in Blue_1 and CIDER and slightly inferior to it in other metrics. Det+Loc. introduce too much redundant information, which results in that its' poorer performance.

Table 2. Results of different caption models. All values are reported as percentage (%).

Caption models	MSCOCO dataset						
	B1	B2	B3	B4	M	R	C
NICs	66.6	46.1	32.9	24.6	–	–	–
LRCN	62.8	44.2	30.4	21.0	–	–	–
m-RNN	67.0	49.0	35.0	25.0	–	–	–
Soft-Attention	70.7	49.2	34.4	24.3	23.9	–	–
Hard-Attention	71.8	50.4	35.7	25.0	23.0	–	–
g-LSTM	67.0	49.1	35.8	26.4	22.7	–	81.3
ATT	70.9	53.7	40.2	**30.4**	24.3	–	–
RA-SF	69.1	50.4	35.7	24.6	22.1	50.1	78.3
(RA-SF)-BEAM10	69.7	51.9	38.1	28.2	23.5	50.9	83.8
(Det.+Loc.)-BEAM4	70.4	53.1	39.2	29.0	23.8	52.1	85.0
FC-2K	–	–	–	28.6	24.1	52.3	90.9
Ours-VGG16	70.9	53.1	38.4	27.4	23.5	51.3	88.0
Ours-RESNET101	**72.9**	**55.6**	**41.0**	29.9	**24.7**	**53.1**	**96**

The results of comparison are strong evidence that (1) the object detection task does have the ability to help with image caption model and our multi-level feature extract module is better suitable for caption task. (2) Our end-to-end model can help the two modules merge to get better performance in caption task.

4.4 Comparison Between Different Beam Search Size

In this section, we introduce Beam Search (BS) to replace Maximum Probability Sampling Mechanism. BS is a heuristic algorithm, which will consider more situations to generate better sentence while testing. The larger the beam size is, the more situation will be considered. We take gLSTM as comparison and Table 3 shows the experimental results.

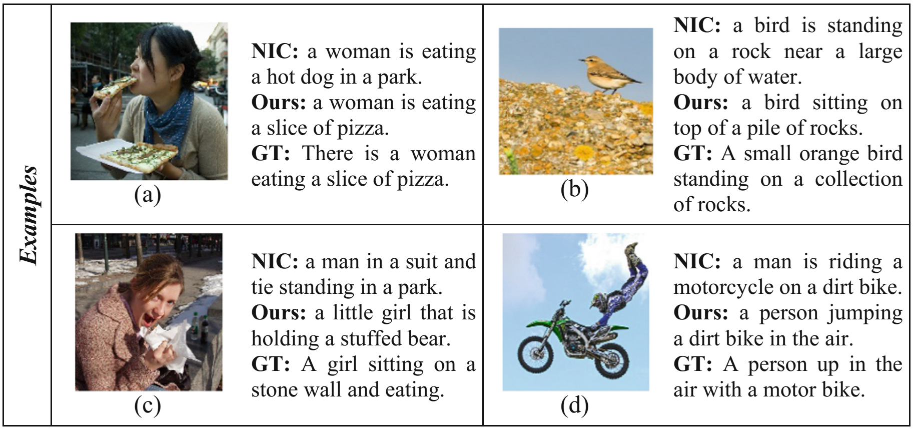
Table 3. Results of different Beam Size. All values are reported as percentage (%).

Beam size	Model	MS-COCO dataset						
		B1	B2	B3	B4	M	R	C
2	gLSTM	70.2	52.7	38.8	28.7	24.1	51.6	88.5
	Ours-VGG16	**71.7**	**54.3**	**40.3**	**29.8**	**24.2**	**52.2**	**92.5**
3	gLSTM	70.2	52.8	39.1	29.0	24.1	51.6	88.9
	Ours-VGG16	**71.1**	**53.9**	**40.2**	**30.0**	**24.2**	**52.3**	**92.6**
4	gLSTM	69.9	52.6	39.0	29.0	24.0	51.4	88.4
	Ours-VGG16	**70.7**	**53.5**	**39.9**	**30.0**	**24.2**	**52.2**	**92.1**

From Table 3, we can see that the performance of a model varies in different beam size. Simultaneously, our model always outperforms gLSTM and it surpass Det+Loc. at beam size = 4. This is another evidence that our accurate guided model is better than other methods.

4.5 Qualitative Results

Figure 5 shows qualitative captioning results. To emphasize the effectiveness of our accurate guidance model and for fair comparison, we compare our VGG16-based model with the baseline model (NIC).



(a) **NIC:** a woman is eating a hot dog in a park. **Ours:** a woman is eating a slice of pizza. **GT:** There is a woman eating a slice of pizza.

(b) **NIC:** a bird is standing on a rock near a large body of water. **Ours:** a bird sitting on top of a pile of rocks. **GT:** A small orange bird standing on a collection of rocks.

(c) **NIC:** a man in a suit and tie standing in a park. **Ours:** a little girl that is holding a stuffed bear. **GT:** A girl sitting on a stone wall and eating.

(d) **NIC:** a man is riding a motorcycle on a dirt bike. **Ours:** a person jumping a dirt bike in the air. **GT:** A person up in the air with a motor bike.

Fig. 5. Qualitative results: **NIC** is the baseline model; **Ours** means our VGG16 based model; **GT** is the ground truth.

The example images include similar colors and rare actions. Our proposed model can better capture objects in the target image, such as “a slice of pizza” in image (a) and “a little girl” in image (b). Our po-gLSTM can better capture the scenes and relationships between objects, such as “on a pile of rocks” in image (b), “in the air” in image (d) and "holding" in image (c), "jumping" in image (d). Assuredly, our model may fail in some cases, such as "bear" in image (c). It is mainly due to there is no class named as "hamburger" while training the object detection network and the hamburger is covered with a white wrapping paper, which is hard for object detection task. If the performance of object detection task gets better, our proposed model can achieve better performance simultaneously. The qualitative result shows that object detection network does do much help to capture the principle objects. Our model does not loss the information of scenes and relationships between objects but it can even do better.

4.6 Visualization of Condition Attention Mechanism

In this section, we visualize the focus of CAM. The brighter part refers to higher attention. Taking the first row as example, our proposed model focus exactly on the bus in the image while generating the word-"bus". When generating "parked", the CAM focus more on where the car and ground contact. This indicates that our po-gLSTM does have the ability to focus on the effective objects all the time (Fig. 6).

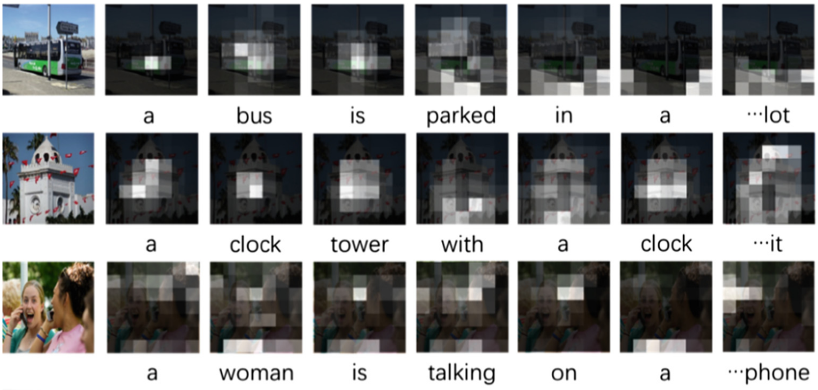


Fig. 6. The visualization of condition attention mechanism on feature maps.

5 Conclusion

In this work, we propose the framework of accurate guidance for image caption. It combines a variety of object detection network (MFEM) and gLSTM with the help of attention mechanism (po-LSTM). We show in our experiments that the proposed methods significantly improve the baseline method and outperform the current state-of-the-art on MS-COCO dataset, which supports our argument of explicit consideration of getting help from object detection task.

References

1. Kulkarni, G., et al.: BabyTalk: understanding and generating simple image descriptions. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1601–1608. IEEE Computer Society (2011)
2. Donahue, J., et al.: Long-term recurrent convolutional networks for visual recognition and description. IEEE Trans. Pattern Anal. Mach. Intell. **39**(4), 677–691 (2015)
3. Jia, X., et al.: Guiding the long-short term memory model for image caption generation. In: IEEE International Conference on Computer Vision, pp. 2407–2415. IEEE Computer Society (2015)
4. Bahdanau, D., et al.: Neural machine translation by jointly learning to align and translate. Comput. Sci. (2014)
5. Hochreiter, S., et al.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
6. You, Q., et al.: Image Captioning with Semantic Attention. In: IEEE Computer Vision and Pattern Recognition, pp. 4651–4659. IEEE Computer Society (2016)
7. Xu, K., et al.: Show, attend and tell: neural image caption generation with visual attention. Comput. Sci. 2048–2057 (2015)
8. Cho, K., et al.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. Comput, Sci (2014)
9. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
10. Karpathy, A., et.al.: Deep visual-semantic alignments for generating image descriptions. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3128–3137. IEEE Computer Society (2015)
11. Ross, G., et al.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: IEEE Computer Vision and Pattern Recognition, pp. 580–587. IEEE Computer Society (2014)
12. Girshick, R.: Fast R-CNN. In: IEEE International Conference on Computer Vision, pp. 1440–1448. IEEE Computer Society (2015)
13. Ren, S., et al.: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**(6), 1137–1149 (2015)
14. He, K., et al.: Deep Residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778. IEEE Computer Society (2016)
15. Simonyan, K., et al.: Very deep convolutional networks for large-scale image recognition. Comput. Sci. (2014)
16. Donahue, J., et al.: Long-term recurrent convolutional networks for visual recognition and description. IEEE Trans. Pattern Anal. Mach. Intell. **39**(4), 677–691 (2017)
17. Mao, J., et al.: Deep captioning with multimodal recurrent neural networks (m-RNN). arXiv preprint arXiv:1412.6632 (2014)
18. Fu, K., et al.: Aligning where to see and what to tell: image captioning with region-based attention and scene-specific contexts. IEEE Trans. Pattern Anal. Mach. Intell. **39**(12), 2321–2334 (2015)
19. Yang, Z., Zhang, Y.-J., Rehman, S., Huang, Y.: Image captioning with object detection and localization. In: Zhao, Y., Kong, X., Taubman, D. (eds.) ICIG 2017. LNCS, vol. 10667, pp. 109–118. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-71589-6_10
20. Rennie, S.J., et al.: Self-critical sequence training for image captioning. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1179–1195. IEEE Computer Society (2017)