Image Captioning with both Object and Scene Information

Xiangyang Li, Xinhang Song, Luis Herranz, Yaohui Zhu, Shuqiang Jiang
Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS)
Institute of Computing Technology, CAS, Beijing, 100190, China
{xiangyang.li, xinhang.song, luis.herranz, yaohui.zhu}@vipl.ict.ac.cn;
sqjiang@ict.ac.cn

ABSTRACT

Recently, automatic generation of image captions has attracted great interest not only because of its extensive applications but also because it connects computer vision and natural language processing. By combining convolutional neural networks (CNNs), which learn visual representations from images, and recurrent neural networks (RNNs), which translate the learned features into text sequences, the content of a image can be transformed into linguistic sequences. Existing approaches typically focus on visual features extracted form an object-oriented CNN (train on ImageNet) and then decode them into natural language. In this paper, we propose a novel model using not only object-related, but also scene-related information extracted from the images. To make full use of both object and scene information, we first combine object information and scene information (extracted from a scene-oriented CNN), and then using as inputs to RNNs. Both types of information provide complementary aspects that help in generating a more complete description of the image. Qualitative and quantitative evaluation results validate the effectiveness of our method.

CCS Concepts

•**Computing methodologies** → **Natural language generation; Computer vision; Image representations;**

Keywords

Image Caption; Long Short-Term Memory; Convolutional Neural Network; Scene and Object

1. INTRODUCTION

Being able to automatically describe the content of an image, a problem known as image captioning, is becoming an important task in the field of artificial intelligence. The task is interesting not only because it has many applications, such as navigation for the blind, human-machine interaction and image retrieval systems to organize and locate images by the form of verbal communication, but also because it is a core problem connecting computer vision and natural language processing. Compared with conventional tasks in the computer vision community such as image classification, object detection, or attribute prediction, generating a meaningful natural language description of an image needs a more sophisticated and holistic understanding of the image. The description not only should describe the objects appearing in the image, but also should depict their spatial configurations, attributes, interactions with other objects, as well as the activities involved in.

Image captioning models take an image as input and output a sequence text that describes its content. The methods of solving this task can be divided into two categories. The first category is based on retrieval [8, 12]. These approaches first find the closest matching images with captions, and then simply transfer these descriptions to the query image. They depend on the training data too much, so it is hard to add or remove some words for a particular image. The

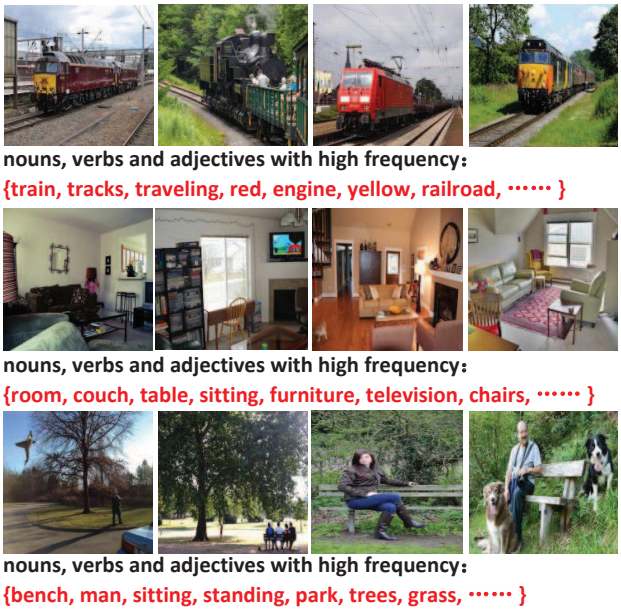

Figure 1: Illustration of the relationship between object information and scene information. The scene categories from top to bottom are railway, bedroom and park respectively.

MM '16, October 15-19, 2016, Amsterdam, Netherlands

DOI: http://dx.doi.org/10.1145/2964284.2984069

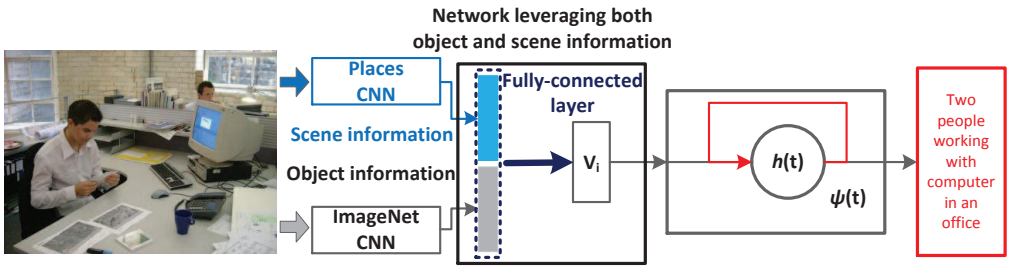

Figure 2: Illustration of the framework. We first use networks pre-trained on ImageNet and Places to extract object information and scene information of the image. And then we train a network to leverage both of the information. The learned feature V_i is used to generate image descriotion.

second category is generating new captions. Some work generates descriptions based on fixed sentence template [2, 9]. They first detect elements such as attributes, objects and actions in the image, and then fill the template with them to generate sentence. The results of these approaches have a fixed format and are less natural. Recently, many researches have adopted CNNs and RNNs to this task [7, 11, 18, 6, 5, 4]. Karpathy *et al.* [7] and Mao *et al.* [18, 11] use a deep CNN to extract visual feature, which is put into a RNN as the initial start word to generate image description. Vinyals *et al.* [18] propose to use LSTM [3]. Jia *et al.* [4] propose gLSTM and adds semantic information to the LSTM model. Xu *et al.* [19] use attention mechanism over the regions of the input image to generate sentence.

However, considering these approaches of generating new captions for images, they just focus on visual features extracted form an object-oriented CNN and then decode them into natural language. Intuitively, the scene information of an image is an invaluable context that can affect the prediction of sequential semantic words. For example, the word of *train* seems unlikely to appear in the *bedroom* scene, while the word of *television* is more likely to appear in the *bedroom* scene than in the *park* scene, as shown in Figure 1.

In this paper, we propose a novel model for image captioning not just using object information, but also using scene information of the images. Similar to Rohrbach's work [14] which learns different oriented classifiers such as verbs, objects and places for movie description, we use CNNs pretrained in ImageNet [15] and Places [20] to extract viusal features. As ImageNet CNN is object-oriented, and Places CNN is scene-oriented, features extracted with different networks contain different information. We include a fully-connected layer to combine object and scene information in visual space, and then feed the RNN units with the enhanced features, as demonstrated in Figure 2. Experimental results demonstrate that the object information and scene information are complementary, and the combination of them can be helpful to improve the performance. Jun *et al.* [6] use scene information from the corpus to factor the LSTM units, while we explicitly use scene information from the images.

The rest of the paper is organized as follow: In section 2, we introduce the framework of our method. In Section 3, we evaluate our method and report the experimental results. Finally, Section 4 gives a summary of this paper.

2. PROPOSED METHOD

In this paper, we proposed a framework to describe the content of an image with both object and scene information. The proposed system is illustrated in in Figure 2. Firstly, networks pre-trained on object-oriented and scene-oriented datasets are used to extract the visual features of the images. And then, a network is trained to leverage both object and scene information. Finally, the enhanced features are put into the RNN units to predict sequential text.

2.1 Object and Scene

Currently, most image captioning approaches [18, 11, 19] only use the CNN pretrained with ImageNet, which focus on the object rather than scene images, to extract visual features. However, the image caption task is not only about the description of particular objects, but is the description of the whole images. While, the scene information can be regarded as the background knowledge and context of the objects contained in the images. Thus, including the scene information into the framework is helpful to generate more complementary and accurate descriptions.

In our framework, we also extract visual features with the CNN pretrained using Places dataset (referred to Places CNN) [20], which is a large scale dataset consisting of 205 scenes. Using Places CNN to extract visual features can be regarded as embedding the scene information into the visual feature, which has never been considered in the previous image captioning approaches.

2.2 Visual Combination

Instead of simple concatenation, we combine the visual features embedded with object and scene information by including another hidden layer in the RNN architecture, which is connected to the both visual features. The connection parameters between input visual features and the hidden layer are learned during the training of the RNN. Thus, the output of the hidden layer can be regarded as the selection among the two input visual features, with the supervision of ground truth sentences during the RNN training.

2.3 Image Captioning

Like most of previous work, our image captioning method follows an encoder-decoder pipeline. The encoder transfers an visual representation into a feature space, and the decoder further converts the feature to a sentence. In our encoding process, CNNs trained on ImageNet and Places are utilized for feature extraction. Then the RNN model is adopted in our decoding process to generate sentence descriptions for images. The core of RNN model is Long Short-Term Memory (LSTM) [3], which has shown state-of-the-art

performance on sequence tasks. The memory cell and gates in a LSTM block are defined as follows:

$i_t = \sigma(W_{ix}x_t + W_{im}m_{t-1})$
$f_t = \sigma(W_{fx}x_t + W_{fm}m_{t-1})$
$o_t = \sigma(W_{ox}x_t + W_{om}m_{t-1})$
$c_t = f_t \odot c_{t-1} + i_t \odot h(W_{cx}x_t + W_{cm}m_{t-1})$
$m_t = o_t \odot c_t$

where $\odot$ represents the elements-wise multiplication, and $\sigma(\cdot)$ represents the sigmoid function and $h(\cdot)$ represents the hyperbolic tangent function. At given time t, LSTM receives different kinds of inputs: current input x_t, the previous hidden state m_{t-1} and the previous memory cell state c_{t-1}. During training, the goal is to maximize the log-likelihood of correct translations given the sentence in the source language. When using this principle to image description generation, the aim is to maximize the log-likehood of the image caption given an image, that is :

$$arg \max_{\theta} \sum_{i} log\ p(s^i_{1:L^i}|x_i, \theta)$$

where x^i denotes an image, $s^i_{1:L^i}$ denotes a sequence of words in a sentence of length L^i and θ denotes the parameters of the model.

3. EXPERIMENTAL RESULTS

3.1 Datasets and Settings

We perform experiments on the the popular large scale *MS COCO* dataset [10]. It is composed of 82,783 images for training and 40,504 images for validation, while each images accompanied with 5 reference sentences annotated by humans. In order to compare the proposed method with previous work, we use the splits used by Karpathy [7], that is, all 82,783 images from the training set for training, and 5,000 images for both validation and testing.

We also perform experiments on the the Caption Prediction Challenge dataset (CPC), which is exclusively drawn from the Yahoo-Flickr Creative Commons 100M (YFCC100M) dataset [17]. It is composed of 1,287,522 images in the training set and 36,884 images in the test set. To train our image caption model, we select 253,890 for training and 30,000 images for validation in the 1,287,522 images.

We use the last fully-connected layer of VGG-16 [16] as the visual features. In order to embed the object and scene information, two CNNs are pre-trained on ImageNet and Places datasets. For *MS COCO*, we remove the words that occurs less than 5 times in the training set, resulting a dictionary of size 8498. As CPC covering a much broader domain, we set the threshold as 2, resulting 14067 words. For the model with both object and scene information, the parameters of the fully-connected layer and the LSTM units are jointly trained from scratch. In our experiments, we set both the size of the fully-connected layer and the size of the hidden units in LSTM models as 512.

3.2 Quantitative Evaluation Results

Although it is sometimes not clear whether a description should be deemed successful or not given an image, previous work has proposed several evaluation metrics The most commonly used metric so far in the image description literature is the BLEU score [13], which evaluates a candidate sentence by measuring the fraction of n-grams that appear in a set of references. So we evaluate our method with the BLEU score. As BLEU is critical to favor short sentences,

Table 1: Performance on MS COCO (%).

Models	B-1	B-2	B-3	B-4	METEOR
Karpathy[7]	62.5	45.0	32.1	23.0	19.5
Vinyals[18]	66.6	46.1	32.9	24.6	-
SF-GREEDY[6]	67.8	49.4	34.8	24.2	21.8
C_O	64.5	45.8	32.1	22.7	20.6
C_S	60.5	41.3	28.1	19.5	18.8
C_OS	**68.1**	**50.6**	**36.8**	**26.8**	**22.8**

Table 2: Results on the validation set of CPC (%).

Models	B-1	B-2	B-3	B-4	METEOR
C_O	10.2	4.3	2.6	1.8	3.1
C_S	9.8	4.0	2.4	1.6	3.0
C_OS	**10.7**	**4.6**	**2.9**	**2.2**	**3.2**

METEOR [1] can avoid this weakness by evaluating a generated sentence by computing a score based on word level matches between the generation and a reference and returning the maximum score over a set of related references. In this paper, beside BLEU, we also use METEOR.

Table 1 shows the performance on MS COCO dataset. C_O indicates the baseline method that generates image description just with object information. For fair comparison, the performance of the method that generates image description just with scene information, referred as C_S, is also evaluated. C_OS is the proposed method which leverages both object and scene information. As image captioning need to describe a image as complete as possible, object information contains more details than scene information, so the performance of C_O surpasses C_S. While, scene information contains the contextual relations between objects, so C_S still has a decent performance. C_OS has the best performance as it contains both object and scene information. Experiments results demonstrate that object information and scene information are complementary, and the combination of them can be helpful to improve the performance. Table 2 shows the performance on the validation set of the Caption Prediction Challenge dataset (CPC). It also demonstrates that the combination of object and scene information can get better performance. The performance of C_OS on the test set of CPC is 9.0%.

3.3 Qualitative Analysis

We hypothesize that object information and scene information are complementary for image captions generation. The results shown in Figure 3 verify this hypothesis. As object information often focuses on particular parts of images and has more details of visual attributes, the generated captions highlight the elements such as people, animal, or related activities. While for scene information, it is more sensitive to the contextual relations between objects and spatial structures, the generated captions highlight more holistic elements. For example, the caption of the image in the first column in Figure 3 (a) generated from object information highlights the girl, while the caption generated from scene information just indicates that it is a kitchen. All the experimental results demonstrate that the combination of object and scene information can improve the performance.

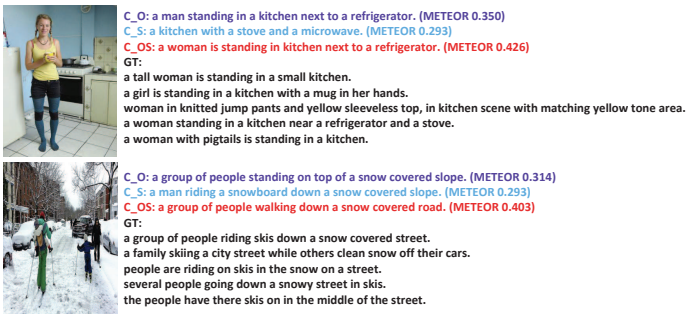

Figure 3: Image descriptions generated with different kinds of information on MS COCO (a) and CPC (b). The METEOR score of each generated sentence is evaluated over the ground truth sentences.

4. CONCLUSION

We have presented a framework to combine both object and scene information to generate the image captions. Both information are embedded in the visual features extracted from the convolutional neural networks pre-trained with ImageNet and Places datasets. And we combine them by include a hidden layer during the recurrent neural network training, which is connected to both input visual features. The experimental results show that including the scene information can get higher performance than only using the object information.

5. ACKNOWLEDGEMENTS

This work is supported in part by the National Basic Research 973 Program of China under Grant No. 2012CB316400, the National Natural Science Foundation of China under Grant Nos. 61532018, 61322212 and 61550110505, the National High Technology Research and Development 863 Program of China under Grant No. 2014AA015202. This work is also funded by Beijing Science And Technology Project under Grant No. D161100001816001 and Lenovo Outstanding Young Scientists Program (LOYS).

6. REFERENCES

[1] M. Denkowski and A. Lavie. Meteor universal: Language specific translation evaluation for any target language. In *EACL 2014 Workshop on Statistical Machine Translation*, 2014.

[2] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In *ECCV*, 2010.

[3] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735 – 1780, 1997.

[4] X. Ji, E. gavves, B. Fernando, and T. Tuytelaars. Guiding Long-Short Term Memoty for image caption genaration. In *ICCV*, 2015.

[5] J. Johnson, A. Karpathy, and L. Fei-Fei. DenseCap: Fully convolutional localization networks for dense captioning. In *CVPR*, 2016.

[6] J. Jun, K. Fu, R. Cui, F. Sha, and C. Zhang. Aligning where to see and what to tell: image aption with region-based attention and scene factorization. In *arXiv:1506.06272*, 2015.

[7] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015.

[8] P. Kuznetsova, V. Ordonez, A. Berg, T. Berg, and Y. Choi. Generalizing image captions for image-text parallel corpus. In *ACL*, 2013.

[9] R. Lebret, P. Pinheiro, and R. Collobert. Pharse-based image captioning. In *ICCV*, 2015.

[10] T. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. ramanan, C. Zitnick, and P. Dollar. Microsoft COCO: Common object in context. In *ECCV*, 2014.

[11] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. Deep captioning with multimodal recurrent neural networks (M-RNN). In *ICLR*, 2015.

[12] R. Mason and E. Charniak. Nonparametric method for data-driven image captioning. In *ACL*, 2014.

[13] K. Papineni, S. Roukos, T. Ward, and W.J. Zhu. Bleu: A method for automatic evaluation of machine translation. In *ACL*, 2002.

[14] A. Rohrbach, M. Rohrbach, and B. Schiele. The long-short story of movie description. In *GCPR*, 2015.

[15] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211 – 252, 2015.

[16] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[17] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L. Li. YFCC100M: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

[18] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and Tell: a neural image caption generator. In *CVPR*, 2015.

[19] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. Show, Attend and Tell: Neural image caption generation with visual attention. In *ICML*, 2015.

[20] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *NIPS*, 2014.