Rich Image Description Based on Regions

Xiaodan Zhang[†§], Xinhang Song[‡], Xiong Lv[‡], Shuqiang Jiang[‡], Qixiang Ye[†], Jianbin Jiao[†]
[†]University of Chinese Academy of Sciences, Beijing 101408, P. R. China
[‡]Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, P. R. China
[§]City University of Hong Kong, Kowloon, Hong Kong
zhangxiaodan10@mails.ucas.ac.cn; {xinhang.song, xiong.lv}@vipl.ict.ac.cn;
sqjiang@ict.ac.cn; {qxye, jiaojb}@ucas.ac.cn

ABSTRACT

Automatically describing the content of an image is a fundamental problem in artificial intelligence that connects computer vision and natural language processing. In contrast to the previous image description methods that focus on describing the whole image, this paper presents a method of generating rich image descriptions from image regions. First, we detect regions with R-CNN (regions with convolutional neural network features) framework. We then utilize the RNN (recurrent neural networks) to generate sentences for image regions. Finally, we propose an optimization method to select one suitable region. The proposed model generates several sentence description of regions in an image, which has sufficient representative power of the whole image and contains more detailed information. Comparing to general image level description, generating more specific and accurate sentences on the different regions can satisfy more personal requirements for different people. Experimental evaluations validate the effectiveness of the proposed method.

Categories and Subject Descriptors

I.4.8 [**Image Processing And Computer Vision**]: Scene Analysis — Object recognition; I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding — perceptual reasoning

Keywords

Image Description; Object Detection; Region Optimization; Convolutional Neural Networks; Recurrent Neural Networks

1. INTRODUCTION

An image contains a lot of information from various aspects. Automatically describing the content of an image using natural language is a challenging task which should on the one hand accurately provide the object and activity information to avoid delivering the wrong message; on the other hand, the description should be as more coverage as possible to avoid miss important information. Compared with conventional image classification and object recognition tasks, which have been studied for a long time in the multimedia and computer vision community, an image description should capture not only the objects contained in an image, but it also must express how these objects relate to each other, as well as how their attributes and the activities are involved in. Moreover, the above semantic knowledge (e.g. objects, attributes, activities) has to be expressed in a natural language like English, which means that a language model is required in addition to visual understanding. Image description has many important application requirements. For example, it can help visually impaired people better understand the content of the image content. The vivid and informative image description is also helpful to satisfy our daily needs such as image searching, human-to-machine interaction, and mobile visual assistance.

Recently, some works [2, 1, 14, 7] have been proposed to generate image descriptions with natural sentences. M.

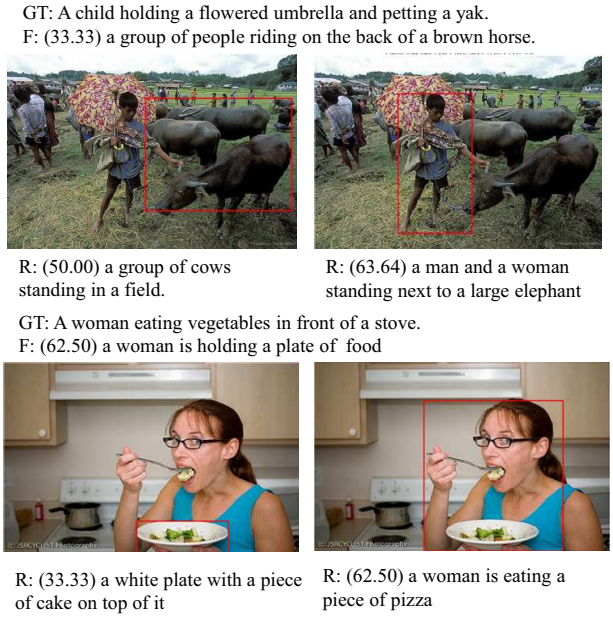

Figure 1: Image description based on regions, compared to full image description. GT: ground truth. F: full image description. R: region based description

MM'15, October 26–30, 2015, Brisbane, Australia.

DOI: http://dx.doi.org/10.1145/2733373.2806338.

Mitchell et al. [2] introduce a novel system by leveraging syntactically informed word co-occurrence statistics and generating syntactic trees from visual methods, which composes human-like descriptions of images. Socher et al. [1] introduce a max-margin structure based predicting architecture, which is recovered by recursive neural networks both in complex scene images and sentences. Oriol Vinyals et al. [14] present a novel model training recurrent neural networks (RNN) model base on the convolutional neural networks (CNN) features to generate natural sentences for describing images. Girish Kulkarni et al. [7] present a system to automatically generate natural language descriptions for images, which exploits statistics gleaned from parsing large quantities of text data and image recognition algorithms.

However, most of the existing works only focus on image description of the whole image, which is limited to represent the the rich information in the image content and are deficiency on personalized applications. For example, when someone wants to get the detailed information of specific position in front of a camera or Google glass, the full image description will be insufficient. Kapathy and Li [6] present a model that generates free-form natural language description of image regions. However, their region descriptions are only annotated with lists of keywords, which is still not suitable enough for practical applications. Detailed image description on image regions to acquire rich visual explanations with sentences have not been investigated yet.

In this paper, we focus on describing images with regions obtained from object detection, aiming to fill the gap and conquer the problem mentioned above. In the proposed approach, an object detection procedure is firstly used to generate candidate regions. An RNN is then trained to learn the describing models between image regions and sentences. Finally, all the regions are used for generating the image descriptions. Region based image description can give more informative representation of an image, which includes some descriptions even the ground truth sentences of benchmark datasets have not recorded. As is shown in Figure 1, where GT means ground truth image description, values like 33.33 are evaluation scores of the corresponding descriptions. The region description can generate sentences like, 'a group of cows standing in a field', 'a white plate with a piece of cake on top of it'. Note that the description of the two regions in the images are not appeared in ground truth sentences. From the experimental analysis, it can be observed that description of regions can be better than or comparable to full image description.

The rest of the paper is structured as follows. In Section 2, the proposed framework is described in detail. Evaluation and experimental results are provided in Section 3. Finally, Section 4 concludes our paper.

2. PROPOSED FRAMEWORK

2.1 Overview

In this paper, we proposed a visual to semantic framework to conduct detailed image description, which aims to generate rich image descriptions through object detection and sentence generation. The architecture of proposed system is shown in Figure 2. Firstly, R-CNN [3], the state-of-the-art object detection method, is adopted to detect image regions of objects. Then a 16 layers convolutional neural network model (VGG) [12] is used to extract visual features for each

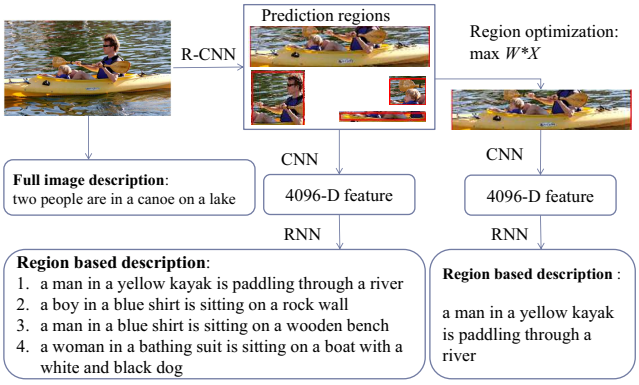

Figure 2: The framework of our region based image description model

region. Finally, the language model of RNN [9] is used to generate the sentences for image regions. In addition, we propose an optimization method to select the optimal region to generate sentence.

2.2 Object Detection

Object detection has been extensively studied in the last decade. Currently, R-CNN [3] based method is the state-of-the-art framework, which uses selective search [13] to find candidate object window proposals, then extracts CNN features for detection.

In this paper, the R-CNN model [3] trained on the 200-category ILSVRC2014 detection challenge dataset is utilized to detect objects and generate image regions. As the corresponding regions of detected objects represent the prominent and significant area of an image, the image description based on these regions can be meaningful and informative.

2.3 Region Optimization

Although R-CNN is the state-of-the-art object detection methods, it can not guarantee the perfect performance in our task. On one hand, the detection results may contain some fault objects. On the other hand, even the right detection objects may contribute little to the final image description.

To reduce the number of the useless detection regions, we define $X = \{x_1, x_2, x_3, x_4\}$ to be the four cues of region selection, x_1: "size: area ratio", x_2: "location: coordinate center position ratio", x_3: "the universality of label", x_4: "prediction score" , and $W = \{w_1, w_2, w_3, w_4\}$ be the optimization parameter. The objective function is: $Y = max\, W * X$, and Y is the final description score of regions.

2.4 Image Description

For image description, the RNN model [6] trained on Flickr8K and MSCOCO datasets are adopted for our image description task. First, the 16-layers CNN network (VGG) [12] trained on the ImageNet dataset [11], is utilized for feature extraction. Then the RNN model is adopted to generate sentence descriptions for image regions. The core of RNN model is Long-Short Term Memory (LSTM) [4], which has shown state-of-the-art performance on sequence tasks such as translation.

The RNN takes the region pixels R and a sequence of input vectors $(x_1, ..., x_T)$. It then computes a sequence of hidden

states $(h_1, ..., h_t)$ and a sequence of outputs $(y_1, ..., y_t)$ by iterating the following recurrence relation for $t = 1$ to N:

$x_{-1} = VGG(R)$,

$h_t = LSTM(h_{t-1}, x_{t-1}), t \in \{1, ..., N\}$,

$y_t = softmax(h_t, x_t), t \in \{1, ..., N\}$.

With the raw generated regions for an image by object detection, there will be several sentences, which contain more detailed and informative contents of the image. And through our region optimization method, there will be one final region and a sentence description for an image, which is comparable with the conventional full image description. Both of the results will be described in our experiments to validate the effectiveness of our method.

2.5 Evaluation Metrics

The most commonly used metric so far in the image description literature is the BLEU score [10], 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.

3. EXPERIMENTS

3.1 Image Datasets

We evaluate the proposed methods on two datasets: Flickr8K [5] and MSCOCO [8]. Flickr8K consists of 8091 images and 5 ground truth sentence descriptions for each image, and we follow the common train/test split. For the MSCOCO dataset, we use 80k images and corresponding sentence descriptions for training and 1k images for testing. For each dataset, we run two experiments for comparison: one is the conventional full image description, the other is the proposed region based rich image description, and both methods use the same CNN and RNN model.

3.2 Image Description of Regions

Through RCNN, each image proposes an average of 4 regions (range from 1 to 9). Then a 16-layers VGG net is adopted to extract features of the regions. At last the 4090-D features are imported to the LSTM model for sentence generation and there will be several sentences for each image. These sentences can be totally different from ground truth sentences. As are shown in Figure 1 and Figure 3 .We can generate more specific sentences from the different regions in the image, while the ground truth sentence only focus on the whole image.

We also use the region optimization method described in section 2.3 to select one best region for evaluation. And the reason why we choose the four cues as the region evaluation criteria is that, in experiment it is observed that regions with bigger size and closer to the image center perform well in image description, which is also unsurprising with the truth that salient region with large area and close to the center of image can represent the whole image. In addition, regions with universal label such as person, animal, fruit, furniture are also more representative. Finally, sentence score of region is also an important index to evaluate description performance.

To make comparison of our method and the conventional full image description, we also exert the full image description use the same CNN and RNN model, except the detection procedure.

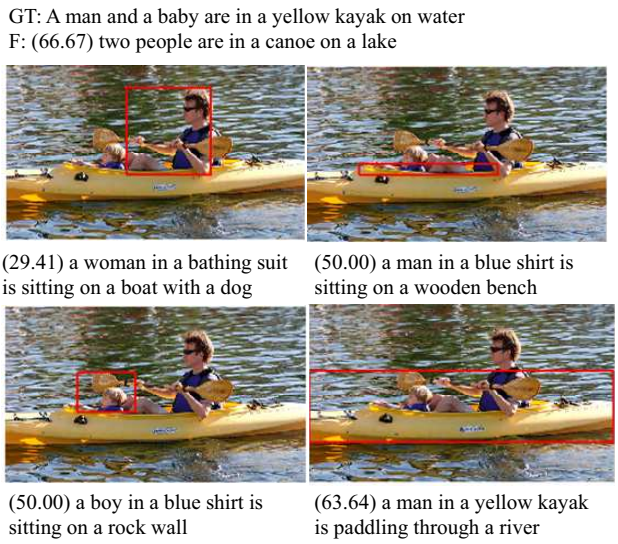

Figure 3: Visual display of the result. GT: ground truth sentence. F: full image description. The numbers indicate the BLEU1 score of sentences.

Table 1: Result of Flickr8K dataset

Types	BLEU1	BLEU2	BLEU3	BLEU4
ID	51.9873	30.7134	13.6014	5.9712
IDR(mean)	46.6835	25.327	10.2291	4.0732
IDR(max)	52.5900	32.1969	15.0766	6.2731
ID+IDR	62.2139	42.7559	23.9343	11.3113

3.3 Evaluation Results

As the most commonly used metric so far in the image description literature is the BLEU score, we choose four indicators, BLEUn (n={1, 2, 3, 4}), as the evaluation metrics. And the bigger the values of BLEU1, BLEU2, BLEU3, BLEU4 are, the better.

Table 1 and Table 2 illustrate the comparison results of Flickr8K and MSCOCO datasets, respectively.

ID indicates the baseline image description method which generate sentence from the whole image. IDR is our approach of image description based on regions, and IDR (mean) is the mean value of BLEU score of all regions, IDR (max) is the max value of the regions score can be. ID+IDR means the combination of full image description and region based image description, and the value is the max BLEU score it can be. Experiments on the two datasets demonstrate that image description based on regions can contain more content than full image description.

To display the effectiveness of our region optimization method, we compare the results of different methods as shown in Figure 4. IDR-O is our approach with region opti-

Table 2: Result of MSCOCO dataset

Types	BLEU1	BLEU2	BLEU3	BLEU4
ID	57.2717	36.1942	17.5965	7.8787
IDR(mean)	50.6205	27.4891	10.7494	4.6179
IDR(max)	59.2985	38.5430	18.8187	8.9793
ID+IDR	64.0047	44.5745	24.8785	12.5477

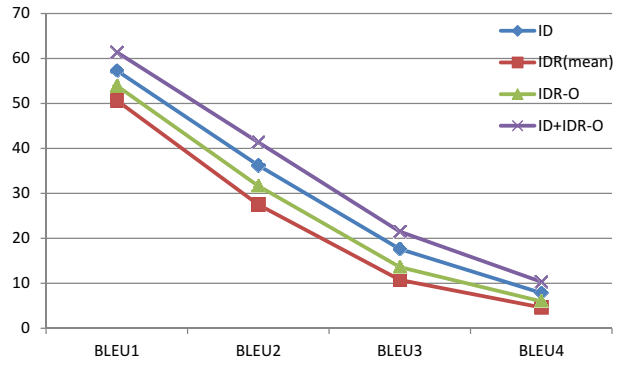

Figure 4: Comparison of IDR-O and ID for MSCOCO dataset.

mization. ID+IDR-O means the combination of full image description and our method with region optimization. It can be seen that our region optimization method can generate a comparable sentence as full image description, which demonstrate that image can be described by representative regions. Image description with region optimization (IDR-O) is better than the mean of region descriptions (IDR-mean), which is due to the reduction of noisy regions of images. And the the combination of two methods(ID+IDR-O) achieves the best performance. It is worth mentioning that for some pure scene image without significant objects, the region based description may degrade severely. So the combination of full image description and region based description gets the best performance.

4. CONCLUSIONS

In this paper, we propose a method to generate rich descriptions for image regions. In the proposed model, three deep neural networks are adopted to generate image regions (R-CNN), extract features (VGG) and generate descriptive sentences (RNN). The experimental results on Flickr8K and MSCOCO datasets verify the effectiveness of the proposed method. We showed that image description based on regions can represent the whole image sufficiently to a great extent, even contains more detailed information out of the ground truth sentences. The proposed region optimization method can select a suitable region for an image and achieve comparable descriptive effect to the full image description. And the full image description and region based description can be complementary.

5. ACKNOWLEDGMENTS

This work was supported in part by the National Basic Research Program of China (973 Program) under Grant 2011CB706900 and Grant 2012CB316400, in part by the National Science Foundation of China under Grant 61271433, Grant 61202323 and Grant 61322212, and in part by National Hi-Tech Development Program (863 Program) of China under Grant 2014AA015202. This work is also funded by Lenovo Outstanding Young Scientists Program (LOYS).

6. REFERENCES

[1] Richard Socher andCliff Chiung-Yu Lin, Andrew Y. Ng, and Christopher D. Manning. Parsing natural scenes and natural language with recursive neural networks. In *ICML*, 2011.

[2] Walter Daelemans, Mirella Lapata, and Lluís Màrquez, editors. *EACL 2012, 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France, April 23-27, 2012*. The Association for Computer Linguistics, 2012.

[3] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 580–587, 2014.

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

[5] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *J. Artif. Intell. Res. (JAIR)*, 47:853–899, 2013.

[6] Andrej Karpathy and Fei-Fei Li. Deep visual-semantic alignments for generating image descriptions. *CoRR*, abs/1412.2306, 2014.

[7] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, Siming Li, Yejin Choi, A.C. Berg, and T.L. Berg. Babytalk: Understanding and generating simple image descriptions. *TPAMI*, 2013.

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.

[9] Tomas Mikolov, Martin Karafiát, Lukás Burget, Jan Cernocký, and Sanjeev Khudanpur. Recurrent neural network based language model. In *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, pages 1045–1048, 2010.

[10] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[11] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. Imagenet large scale visual recognition challenge. *CoRR*, abs/1409.0575, 2014.

[12] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[13] Jasper R. R. Uijlings, Koen E. A. van de Sande, Theo Gevers, and Arnold W. M. Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 104(2):154–171, 2013.

[14] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. *CoRR*, abs/1411.4555, 2014.