Image2Text: A Multimodal Caption Generator

Chang Liu[1]* , Changhu Wang[2], Fuchun Sun[1], Yong Rui[2]
[1]Department of Computer Science, Tsinghua University, Beijing, P.R. China
[2]Microsoft Research, Beijing, P.R. China
cliu13@mails.tsinghua.edu.cn, chw@microsoft.com
fcsun@mail.tsinghua.edu.cn, yongrui@microsoft.com

ABSTRACT

In this work, we showcase the Image2Text system, which is a real-time captioning system that can generate human-level natural language description for any input image. We formulate the problem of image captioning as a multimodal translation task. Analogous to machine translation, we present a sequence-to-sequence recurrent neural networks (RNN) model for image caption generation. Different from most existing work where the whole image is represented by a convolutional neural networks (CNN) feature, we propose to represent the input image as a sequence of detected objects to serve as the source sequence of the RNN model. Based on the captioning framework, we develop a user-friendly system to automatically generated human-level captions for users. The system also enables users to detect salient objects in an image, and retrieve similar images and corresponding descriptions from a database.

CCS Concepts

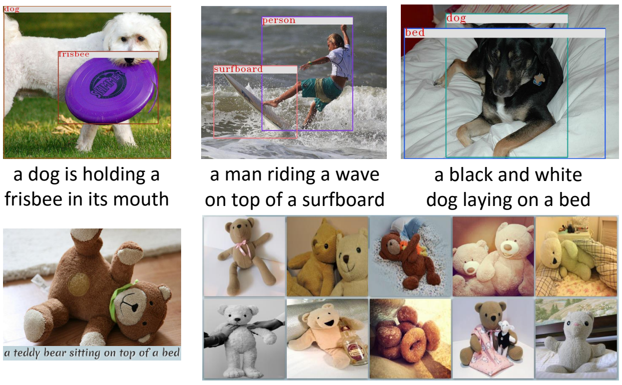
•Information systems → Language models; •Computing methodologies → Neural networks; 5

Keywords

Image Captioning; Object Detection; Deep Neural Networks

1. INTRODUCTION

Image captioning is a challenging research topic due to the requirements of the knowledge of both vision modality and natural language modality. The ultimate goal of image captioning is to generate natural language description for any given image in a real-time manner, just like what we humans do. Moreover, the generated language should be capable of describing the objects and their relations in the image in a grammar error-free and fluent way.



a dog is holding a frisbee in its mouth

a man riding a wave on top of a surfboard

a black and white dog laying on a bed

a teddy bear sitting on top of a bed

Figure 1: Example results of the system. Given any image, the system is able to generate human-level caption sentences, show detected objects with bounding boxes and labels, and return similar images with corresponding sentences in real time.

Despite the difficulty of this task, there have been emerging efforts recently due to the success of introducing the deep neural networks to this field. Most existing work leverages the deep convolutional neural networks (CNN) and the recurrent neural networks (RNN) in an encoding-decoding scheme [6, 4, 1, 5, 3]. In these work, the input image is usually encoded by a fixed length of CNN feature vector, functioning as the first time-step input to the RNN; the description is generated by conditioning the output word at each time step on the input visual vector as well as the previous generated words.

However, since the power of RNN lies in its capability in modeling the contextual information between each time step [2], the encoded visual vector extracted from the whole image may weaken the RNN's memory of the visual information as it contains no temporal concept. To encode more visual information to balance the source and the target sequence of the recurrent neural networks, a natural idea is to leverage the object information in the image. We propose in this work to leverage the high-level features of the detected objects in the image, in order to enrich the visual part of the source sequence. The objects are arranged in a sequence manner according to their saliency, and thus expand the encoding phase of the recurrent neural networks for the image from one time step to multiple time steps. The object features are then mapped into a common hidden

*This work was performed at Microsoft Research Asia.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

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

© 2016 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-3603-1/16/10.
DOI: http://dx.doi.org/10.1145/2964284.2973831

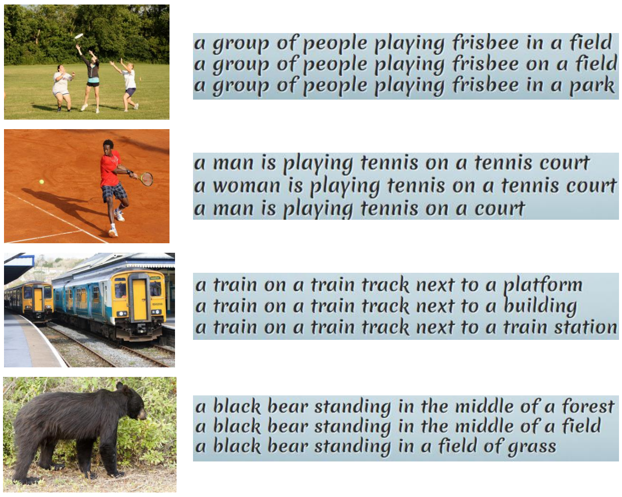

a group of people playing frisbee in a field
a group of people playing frisbee on a field
a group of people playing frisbee in a park
a man is playing tennis on a tennis court
a woman is playing tennis on a tennis court
a man is playing tennis on a court
a train on a train track next to a platform
a train on a train track next to a building
a train on a train track next to a train station
a black bear standing in the middle of a forest
a black bear standing in the middle of a field
a black bear standing in a field of grass

Figure 2: Examples of image caption generation. Given any image, the model can generate the natural language description of the image in real time.

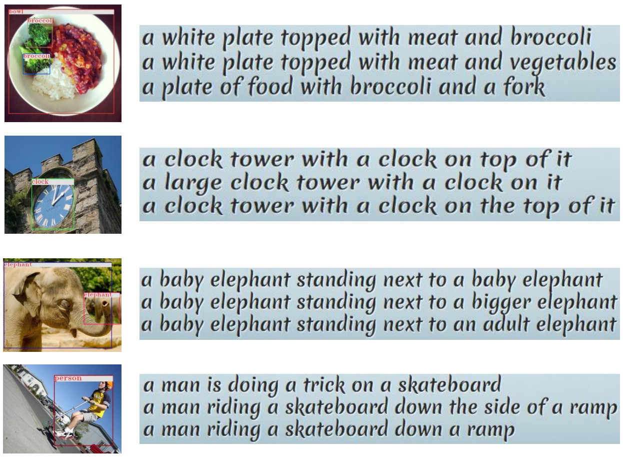

a white plate topped with meat and broccoli
a white plate topped with meat and vegetables
a plate of food with broccoli and a fork
a clock tower with a clock on top of it
a large clock tower with a clock on it
a clock tower with a clock on the top of it
a baby elephant standing next to a baby elephant
a baby elephant standing next to a bigger elephant
a baby elephant standing next to an adult elephant
a man is doing a trick on a skateboard
a man riding a skateboard down the side of a ramp
a man riding a skateboard down a ramp

Figure 3: Alongside the image caption generation, the system is also capable of detecting objects in the image, whose bounding boxes and labels are shown at the same time.

space to serve as the 'visual words' in the source sequence. Thus the problem is formulated as a multimodal translation problem, in which we use the detected objects in the image to serve as the source input 'language', and the corresponding words as the target language.

We develop an image captioning system, which allows the users to 1) generate human-level natural language description of any input image, 2) detect objects in the given image alongside caption generation, and 3) retrieve similar images and descriptions from a database, which holds over 120,000 image-description pairs, all in real time. The core of the system is a combination of pre-trained deep convolutional neural networks for object detection, and recurrent neural networks for caption generation.

2. SYSTEM OVERVIEW

Func.1 Examples The first function of the system is a simple example section, which demonstrates the basic image captioning task by presenting several example images for users to choose. Once the user selects some image among example images, a caption for the image will be generated in the background process of the system, and the response time is generally less than 0.2s. The main purpose of this function is to provide users with ease to use the system at a glance.

Func.2 Capioning The system provides an interface for the users to upload an image or copy the image URL for futher captioning. Once uploaded, a human-level description result of the image will be shown, also in a real-time manner. The results contain three best candidates given by the model, ordered by the log probability of the generation result, as shown in Fig. 2.

Func.3 Detection and Captioning The third part of the system provides a further functionality for the input image. Beyond generating the caption of the image, the model will also detect the objects in the image, and presents the detection result by drawing bounding boxes around the detected objects. The object names are also shown on top of the bounding boxes. Similar to the captioning function, three candidate captions will be shown on the bottom of the image, with the same ordering strategy, as shown in Fig. 3.

Func.4 Retrieval The system enables users to retrieve similar images and corresponding descriptions from a database given any input image. The user can have a glimpse of the retrieved image by clicking one of them, and a pop-up box will present a sharper version of the image, with the caption below it, as shown in Fig. 4

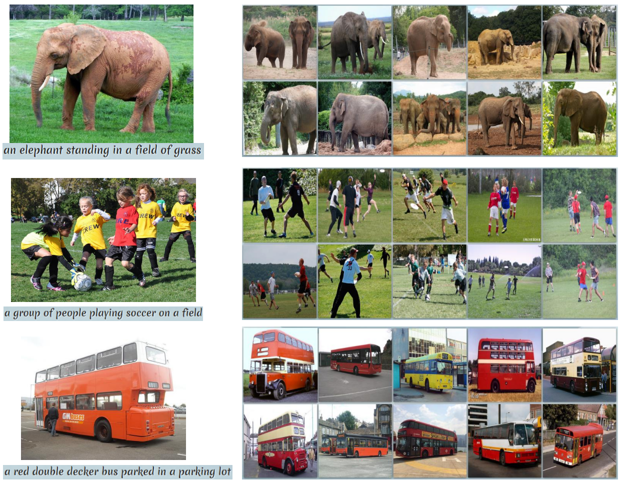

an elephant standing in a field of grass
a group of people playing soccer on a field
a red double decker bus parked in a parking lot

Figure 4: Examples of image retrieval. Similar images to the query image and corresponding sentences are returned.

Acknowledgement

Acknowledgement: This paper is jointly supported by National Natural Science Foundation of China under with Grants No.61621136008, 61327809,61210013, 91420302 & 91520201

3. REFERENCES

[1] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

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

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

[4] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

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

[6] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.