Rich Image Captioning in the Wild

Shuang Sha

June 24,2018

Abstract

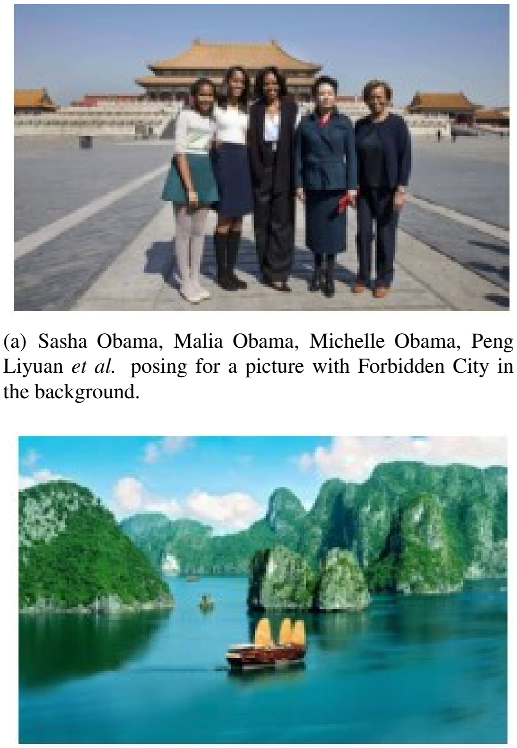
In this paper, authors proposed an image caption system that automatically describe images in the wild. And this research solved new challenges. The challenges include emerging high quality caption about human judgments, out-of-domain data handing, and low latency required in many applications. Built on top of an advanced framework, Authors developed three models. The three models include a deep vision model that detects a broad range of visual concepts, an entity recognition model that identifies celebrities and landmarks, and a confidence model for the caption output. Experimental results show that this caption system can obtain better performs.

1. Introduction

Captioning is a fundamental task in Artificial Intelligence (AI) which describes objects, attributes and relationship in an image, or in a natural language form. It has many applications such as semantic image search, visual intelligent chat, or helping some people who eyes have problems to see the world. Recently, many researches have researched image captioning, and it has a hot topic, as shown in [2, 1, 3, 4].

The leading approaches mainly use two forms. One forms takes an end-to end, encoder-decoder framework adopted from machine translation, [4] use this form. The other form applies a compositional framework [2]. However, the two forms have some abuses, it is unclear how these systems perform in open-domain images. Furthermore, most of the image captioning systems only describe generic visual content without identifying key entities. The entities, such as celebrities and landmarks, are important pieces in common sense and knowledge. As shown in Figure 1, in many situations, the entities are the key information in an image.

In this paper, authors proposed a captioning system for open domain images. They take a compositional approach by starting from one of the advanced image captioning framework [2]. In order to solve the challenges when describing images in the wild, they developed a visual model by detecting a boarder range of visual concepts, and an entity recognition model that generates caption by recognizing celebrities and landmarks, as shown in Figure 1. Further, in order to provide graceful handling for images that are difficult to describe, they built a confidence model to estimate a confidence score for the caption output based on the vision and text features, and provide a back-off caption for these difficult cases.



(a) Sasha Obama, Malia Obama, Michelle Obama, Peng Liyuan *et al.* posing for a picture with Forbidden City in the background.

(b) A small boat in Ha-Long Bay.

Figure 1. Rich captions enabled by entity recognition.

In order to measure the quality of the caption from the humans perspective, they carried out a series of human evaluations through crowd souring, and report results based on humans judgments. The results of the experiment shown that the system proposed by this paper can gain better results than other system.

References

[1] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, T. Darrell, and K. Saenko. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, pages 2625–2634, 2015. 1

[2] H. Fang, J. C. Platt, C. L. Zitnick, G. Zweig, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollar, and J. Gao. From captions to visual concepts and back. In *CVPR*, pages 1473–1482, 2015. 1

[3] A. Karpathy and F. F. Li. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pages 3128–3137, 2015. 1

[4] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *CVPR*, pages 3156–3164, 2015. 1