# Generation and Comprehension of Unambiguous Object Descriptions

Zhang, Liqiang

June 28, 2018

## Abstract

*In this paper, the author propose a method that can generate an unambiguous description (known as a referring expression) of a specific object or region in an image, and which can also comprehend or interpret such an expression to infer which object is being described. They show that their method outperforms previous methods that generate descriptions of objects without taking into account other potentially ambiguous objects in the scene. The model they propose is inspired by recent successes of deep learning methods for image captioning, but while image captioning is difficult to evaluate, their task allows for easy objective evaluation.*

## 1. Introduction

There has been a lot of recent interest in generating text descriptions of images. However, fundamentally this problem of image captioning is subjective and ill-posed. With so many valid ways to describe any given image, automatic captioning methods are thus notoriously difficult to evaluate. In this paper, the author focus on a special case of text generation given images, where the goal is to generate an unambiguous text description that applies to exactly one object or region in the image. Such a description is known as a "referring expression" [3]. This approach has a major advantage over generic image captioning, since there is a well-defined performance metric: a referring expression is considered to be good if it uniquely describes the relevant object or region within its context, such that a listener can comprehend the description and then recover the location of the original object.

There are two problems: (1) *description generation*, in which the author must generate a text expression that uniquely pinpoints a highlighted object/region in the image and (2) *description comprehension,* in which this team must automatically select an object given a text expression that refers to this object (see Fig. 1). Most prior work in the literature has focused exclusively on description generation. Golland *et al.* [2] consider generation and comprehension, but they do not process real world images.

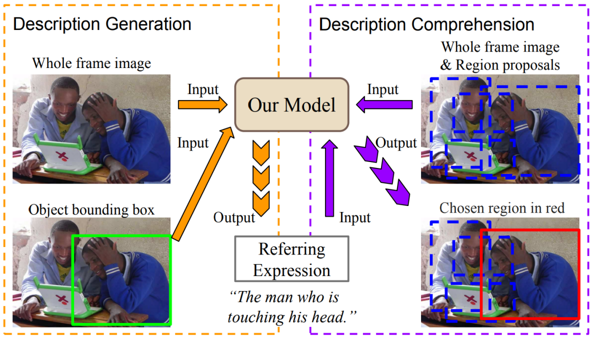

Figure 1. Illustration of their generation and comprehension system. On the left the atuhor see that the system is given an image and a region of interest; it describes it as the man who is touching his head, which is unambiguous (unlike other possible expressions, such as "the man wearing blue", which would be unclear). On the right the ream see that the system is given an image, an expression, and a set of candidate regions (bounding boxes), and it selects the region that corresponds to the expression.

In this paper, the author jointly model both tasks of description generation and comprehension, using state-of-the-art deep learning approaches to handle real images and text. Specifically, their model is based upon recently developed methods that combine convolutional neural networks (CNNs) with recurrent neural networks (RNNs). This team demonstrate that their model outperforms a baseline which generates referring expressions without regard to the listener who must comprehend the expression [1]. They also show that their model can be trained in a semi-supervised fashion, by automatically generating descriptions for image regions.

## 2. Maximum Likelihood Training

Their training data consists of observed triplets, where $I$ is an image, $R$ denotes a region within $I$, and $S$ denotes a referring expression for $R$. To train the baseline model, the author minimize the negative log probability of the referring expressions given their respective region and image as Eq. 1:

$$J(\theta) = -\sum_{n=1}^{N} \log p(S_n|R_n, I_n, \theta), \qquad (1)$$

where $\theta$ are the parameters of the RNN and CNN, and where the atuhor sum over the N examples in the training set. They use ordinary stochastic gradient decent with a batch size of 16 and use an initial learning rate of 0.01 which is halved every 50,000 iterations. Gradient norms are clipped to a maximum value of 10. To combat overfitting, the author regularize using dropout with a ratio of 0.5 for both the word-embedding and output layers of the LSTM.

## 3. Experiments

| Proposals Descriptions | GT | | Multibox | |
|---|---|---|---|---|
| | GEN | GT | GEN | GT |
| ML (baseline) | 0.803 | 0.654 | 0.564 | 0.478 |
| MMI-MM-easy-GT-neg | 0.851 | 0.677 | 0.590 | 0.492 |
| MMI-MM-hard-GT-neg | **0.857** | **0.699** | 0.591 | 0.503 |
| MMI-MM-multibox-neg | 0.848 | 0.695 | **0.604** | **0.511** |
| MMI-SoftMax | 0.848 | 0.689 | 0.591 | 0.502 |

Table 1. The columns show performance on ground truth or multibox proposals, and ground truth (human) or generated descriptions. Thus the columns with GT descriptions evaluate the performance of the comprehension system, and the columns with GEN descriptions evaluate (in an end-to-end way) the performance of the generation system.

In this experiment the team treat UNC-Ref as a validation set to explore various algorithmic options and hyperparameter settings for MMI. Only after having fixed these algorithmic options and hyperparameter settings did they do experiments on their G-Ref dataset. The results are summarized in Table. 1.

## References

[1] X. Chen and C. L. Zitnick. Mind's eye: A recurrent visual representation for image caption generation. In *CVPR*, 2015. 1

[2] B. Hariharan, P. Arbelez, R. Girshick, and J. Malik. Simultaneous detection and segmentation. In *ECCV*, pages 297–312, 2014. 1

[3] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Baby talk: Understanding and generating image descriptions. In *CVPR*, 2011. 1