

Exploring Visual Relationship for Image Captioning

Ting Yao¹, Yingwei Pan¹, Yehao Li², and Tao Mei¹

¹ JD AI Research, Beijing, China

² Sun Yat-sen University, Guangzhou, China

{tingyao.ustc, panyw.ustc, yehaoli.sysu}@gmail.com, tmei@live.com

Abstract. It is always well believed that modeling relationships between objects would be helpful for representing and eventually describing an image. Nevertheless, there has not been evidence in support of the idea on image description generation. In this paper, we introduce a new design to explore the connections between objects for image captioning under the umbrella of attention-based encoder-decoder framework. Specifically, we present Graph Convolutional Networks plus Long Short-Term Memory (dubbed as GCN-LSTM) architecture that novelly integrates both semantic and spatial object relationships into image encoder. Technically, we build graphs over the detected objects in an image based on their spatial and semantic connections. The representations of each region proposed on objects are then refined by leveraging graph structure through GCN. With the learnt region-level features, our GCN-LSTM capitalizes on LSTM-based captioning framework with attention mechanism for sentence generation. Extensive experiments are conducted on COCO image captioning dataset, and superior results are reported when comparing to state-of-the-art approaches. More remarkably, GCN-LSTM increases CIDEr-D performance from 120.1% to 128.7% on COCO testing set.

Keywords: Image Captioning · Graph Convolutional Networks · Visual Relationship · Long Short-Term Memory

1 Introduction

The recent advances in deep neural networks have convincingly demonstrated high capability in learning vision models particularly for recognition. The achievements make a further step towards the ultimate goal of image understanding, which is to automatically describe image content with a complete and natural sentence or referred to as image captioning problem. The typical solutions [7,34,37,39] of image captioning are inspired by machine translation and equivalent to translating an image to a text. As illustrated in Figure 1 (a) and (b), a Convolutional Neural Network (CNN) or Region-based CNN (R-CNN) is usually exploited to encode an image and a decoder of Recurrent Neural Network (RNN) w/ or w/o attention mechanism is utilized to generate the sentence, one word at each time step. Regardless of these different versions of CNN plus

