Natural Language Object Retrieval

Zhang, Liqiang

June 14, 2018

Abstract

In this paper, the author address the task of natural language object retrieval, to localize a target object within a given image based on a natural language query of the object. Natural language object retrieval differs from text-based image retrieval task as it involves spatial information about objects within the scene and global scene context. To address this issue, the auhtor propose a novel Spatial Context Recurrent ConvNet (SCRC) model as scoring function on candidate boxes for object retrieval, integrating spatial configurations and global scene-level contextual information into the network. The model processes query text, local image descriptors, spatial configurations and global context features through a recurrent network, outputs the probability of the query text conditioned on each candidate box as a score for the box, and can transfer visual-linguistic knowledge from image captioning domain to our task.

1. Introduction

Significant progress has been made in object detection in recent years; with the help of Convolutional Neural Networks (CNNs), it is possible to detect a predefined set of object categories with high accuracy, and the number of categories in object detection has grown over 10K to 100K with the help of domain adaptation and hashing [2]. However, in practical application scenarios, instead of using a predefined fixed set of object categories, one would often prefer to refer to an object with natural language rather than use a predefined category label. Such natural language query can include different types of phrases such as categories, attributes, spatial configurations and interactions with other objects, such as the young lady in a white dress sitting on the left or white car on the right in Fig. 1. In this paper, the author address the problem of natural language object retrieval: given an image and a natural language description of an object as query, they want to retrieve the object by localizing the object in the image [3]. Natural language object retrieval can be seen as a generalization of generic object detection and has a wide range of applications, such as handling natural language commands in robotics where the user may ask to a robot to pick up "the TV remote control on the shelf".

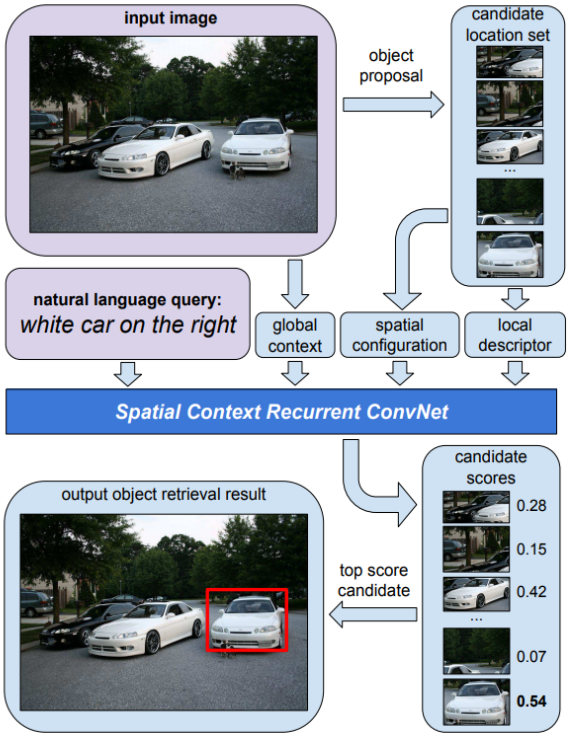

Figure 1. Overview of our method. Given an input image, a text query and a set of candidate locations (e.g. from object proposal methods), a recurrent neural network model is used to score candidate locations based on local descriptors, spatial configurations and global context. The highest scoring candidate is retrieved.

2. The model

Inspired by the architecture of LRCN , the Spatial Context Recurrent ConvNet (SCRC) model for natural language object retrieval consists of several components as illustrated in Fig. 2.

At test time, given an input image I, a query text S and

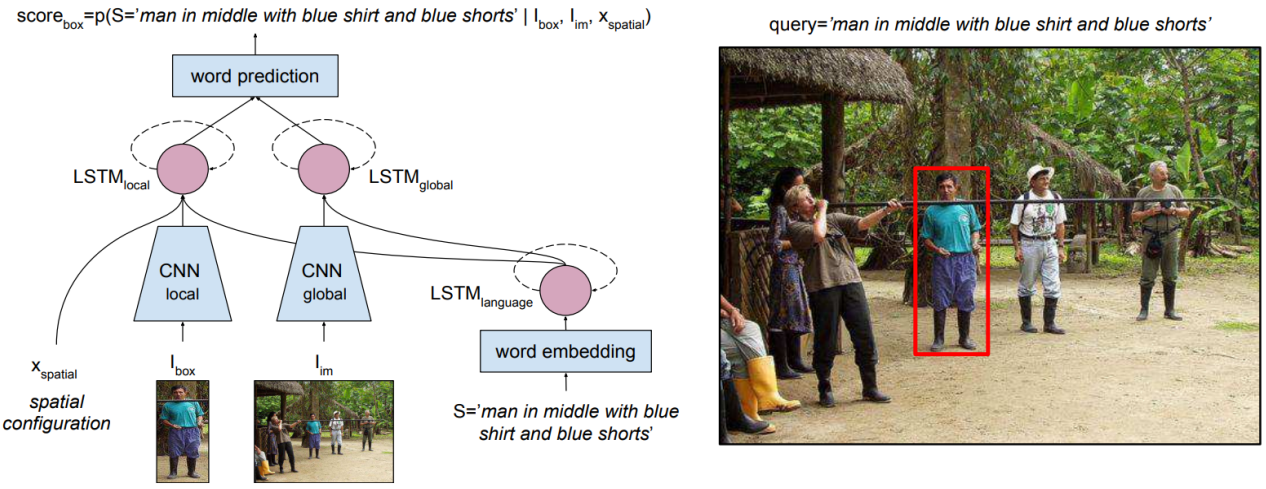

Figure 2. The Spatial Context Recurrent ConvNet (SCRC) for natural language object retrieval. The recurrent network in the author's model contains three LSTM units [1]. Two CNNs are used to extract local image descriptors and global scene-level contextual feature respectively. Parameters in word embedding, word prediction and three LSTM units are initialized by pretraining on image captioning dataset.

a set of candidate bounding boxes $\{b_i\}$, the query text S is scored on i-th candidate box using the likelihood of S conditioned on the local image region, the whole image and the spatial configuration of the box, computed as

$$\begin{aligned} s &= p(S|I_{box}, I_{im}, x_{spatial} \\ &= \prod_{w_t \in S} p(w_t|w_{t-1}, \cdots, w_1, I_{box}, I_{im}, x_{spatial}). \end{aligned} \quad (1)$$

and the highest scoring candidate boxes are retrieved.

$$\begin{aligned} & p(w_{t+1}|w_t, \cdots, w_1, I_{box}, I_{im}, x_{spatial}) \\ &= \text{Softmax}(W_{local} h^{(t)}_{local} + W_{global} h^{(t)}_{global} + r) \end{aligned} \quad (2)$$

where W_{local} and W_{global} are weight matrices for word prediction and r is a bias vector. Softmax$(\cdot)$ is a softmax function over a vector to output a probability distribution.

3. Experiments

Method	P@1-NR	P@1
CAFFE-7K	32.53%	27.73%
LRCN [4]	-	38.38%
SCRC(w/o context, spatial, transfer)	-	61.03%
SCRC(w/o context, spatial)	-	64.09%
SCRC(w/o context)	-	70.15%
SCRC	-	**72.74%**

Table 1. Top-1 precision of our method compared with baselines on annotated bounding boxes in ReferIt dataset.

In Table 1, it can be seen that pretraining on image captioning, adding spatial configuration, and adding scene-level context all improve the performance, with adding spatial configuration $x_{spatial}$ leading to the most significant performance boost.

References

[1] R. Girshick. Fast R-CNN. In *ICCV*, pages 1440–1448, 2015. 2

[2] A. Karpathy, A. Joulin, and L. Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. In *NIPS*, pages 1889–1897, 2014. 1

[3] C. Kong, D. Lin, M. Bansal, R. Urtasun, and S. Fidler. What are you talking about? Text-to-image coreference. In *CVPR*, pages 3558–3565, 2014. 1