Retrieval of Natural Language Problems

Lina Zhu

June 18 ,2018

Abstract

In this article, we address the task retrieval of natural language tasks to locate the object-based natural language query images within a given target object. Natural language object retrieval is different from text-based images because it involves spatial information about the objects in the context of the scene and the global scene. Our model handles query text, local image descriptors, the spatial configuration and global context features query the text that is conditional on each candidate box through the probability of a regular network output as a score for this box, and can transmit the visual language knowledge image caption field to our task. Experimental results demonstrate that our method effectively uses local and global information, significantly differs from previous benchmark methods in different data sets and scenarios, and can utilize large-scale visual and linguistic data sets for knowledge transfer.

1. Introduction

Significant progress has been made in target detection in recent years; with the help of convolutional neural networks (CNN), a set of predefined high-precision object classes can be detected[2]and the number of categories in object detection has grown to more than 10K to 100K with domain adaptation and hash help. However, in a practical application scenario, rather than using a predefined, fixed set of object categories, one would generally prefer to refer to objects in natural language rather than using pre-defined category labels. This natural language query can include different types of phrases, such as categories, attributes, spatial configurations, and interactions with other objects, such as this young lady in a white dress sitting on the left or right side of the white car in Figure 1. In this article, we solve the problem of natural language retrieval of objects: given images and natural language descriptions as the objects of the query, we want to retrieve objects by localizing the objects in the image.

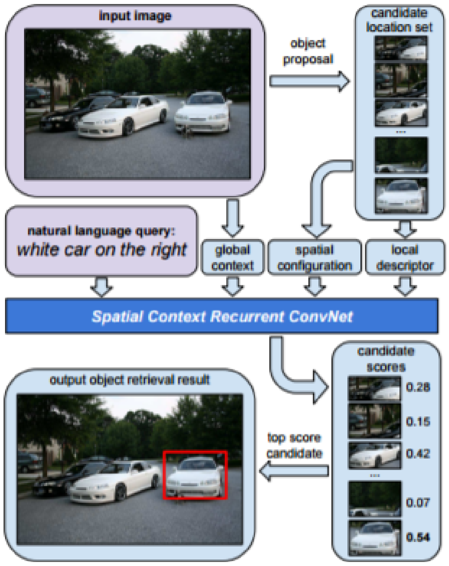

Figure 1. Overview of our method. Given an input image, a text query and a set of candidate locations , a recurrent neural network model is used to score candidate locations based on local descriptors, spatial configurations and global context. The highest scoring candidate is retrieved.

Natural language object retrieval can be viewed as generic generic object detection and has a wide range of applications. For example, a local user who handles natural language commands in robots can request the robot to pick up the TV remote control on the shelf. We construct natural language object retrieval as a set of candidate locations for a retrieval task in a given image, as shown in Figure 1. Where candidate locations can come from object proposal methods, we observe that only text-based image retrieval is applied from candidate locations. The system on the cropped image area causes low performance because of this task, as the natural language object retrieval involves spatial configuration objects and global scenes as contexts. Although both text-based image retrieval and natural language object retrieval involve co-modeling images and texts, they are different visions and linguistic domains that transfer domain names from the entire image to bounding boxes.

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

Method	P@1-NR	P@1
CAFFE-7K	32.53%	27.73%
LRCN	-	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%

2. Related Work

Natural language object retrieval. Based on a bag of word sentence models and an embedded classifier derived from ImageNET solves problems similar to ours and localizes queries within the image based on text. Given a set of candidate target areas, generates a textual representation from the candidate as a bagged text using the category name predicted from the large-scale pre-training the classifier compares the word bag with the query text. Other methods generate visual features from the query text and match them to the image area, for example, by text embedding phrases and visual features for the primary image search engine[1] or a combination of learning texts. In our work at the same time, [7] also proposed a circular network model to locate the object from the given description. The object from the image description is grounded. Specific to an image and its description statement, [4]aligns the sentence by embedding the detection result into a segment of the image from a pre-trained object detector and dependency tree from the loss of the resolver. [5]builds on[4] and replaces the dependency tree with a bidirectional RNN. Canonical correlation analysis is used to learn the joint embedding of image regions and text segments to locate each object mentioned in the title. Use a structural prediction model to keep the text consistent with the image and the reason about the common reference of objects in the text parsed by the 3D scene.

At the same time as this paper, uses the attention model to provide reference phrases in the participating image descriptions to the areas where the phrases can be best reconstructed. The image title method enters the image and generates a text description of it. Recently, a method based on a recurrent neural network [8]has proved effective in this task. Image retrieval is a text-based image retrieval system that selects the most suitable image query text from a set of images. In image retrieval, the learning ranking function is through a recurrent neural network [3], metric learning, correlation analysis[6] and other methods.

3. Experiments

experiment We evaluated our method on a small dataset with a large scale. More experimental results can be found in supplementary materials. A 7K large-scale fine-grained classifier object class is trained on ImageNET. In each box, the candidate set is classified into one of the 7K classes, and a packet word is extracted from the predicted object class based on ImageNET and a synonym containing its category name. The word bag is then projected into the vector space and a score is obtained using the cosine matching to the expected query text distance. Sentence projection in is predefined and is the only training involved in training 7K object classifiers. Please note that also proposes an instance matching model that depends on the online API at the time of the test. As in this work, we assume an independent system without resorting to other APIs.

4. Result

Table 1 shows all the annotation box pictures in the top-1 precision scene candidate set. The highest precision in all cases includes non-informative results where random guesses are used. The results show that our complete SCRC model achieved the highest pre-1 accuracy. As can be seen in Table 1, pre-training the image captions, adding spatial configurations, and adding scene-level contexts all improve performance, because the spatial configuration is not only so beneficial in the query if spatial relationships are directly involved, but also There are locations of objects that enable the network to learn prior assignments.

5. Conclusion

In this article, we discuss natural language object retrieval and spatial context recursion. The recursive neural network model for recursive candidate boxes is based on local image descriptors, spatial configuration, and global context-level contexts. We show that natural language object retrieval is significantly more efficient in merging spatial configurations and global environments to improve performance. The cyclic network model used in our approach led to an end-to-end training scoreability feature, which is significantly better than the benchmark method.

References

[1] R. Arandjelovic and A. Zisserman. Multiple queries for large scale specific object retrieval. In *BMVC*, 2012. 2

[2] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 1

[3] 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 *CVPR*, 2015. 2

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

[5] A. Karpathy and F. Li. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 2

[6] B. Klein, G. Lev, G. Sadeh, and L. Wolf. Fisher vectors derived from hybrid gaussian-laplacian mixture models for image annotation. *arXiv preprint arXiv:1411.7399*, 2014. 2

[7] J. Mao, J. Huang, A. Toshev, O. Camburu, A. Yuille, and K. Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 2016. 2

[8] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015. 2