Bridging the Gap between Expression and Scene Text for Referring Expression Comprehension (Student Abstract)

Yuqi Bu[1,2], Jiayuan Xie[1,2], Liuwu Li[1,2], Qiong Liu[1], Yi Cai[1,2*]

[1] School of Software Engineering, South China University of Technology, Guangzhou, China
[2] Key Laboratory of Big Data and Intelligent Robot (South China University of Technology), Ministry of Education
seyqbu@mail.scut.edu.cn, sexiejiayuan@mail.scut.edu.cn, liuwu.li@outlook.com, liuqiong@scut.edu.cn, ycai@scut.edu.cn

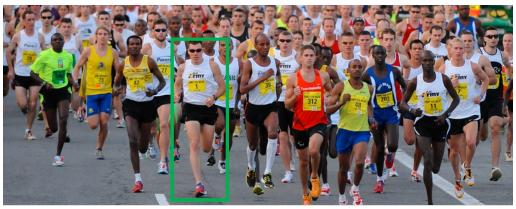
Abstract

Referring expression comprehension aims at grounding the object in an image referred to by the expression. Scene text that serves as an identifier has a natural advantage in referring to objects. However, existing methods only consider the text in the expression, but ignore the text in the image, leading to a mismatch. In this paper, we propose a novel model that can recognize the scene text. We assign the extracted scene text to its corresponding visual region and ground the target object guided by expression. Experimental results on two benchmarks demonstrate the effectiveness of our model.

Introduction

Referring expression comprehension (REC) aims at grounding the object in an image referred to by the natural language expression (Qiao, Deng, and Wu 2020). It is a fundamental step of many vision-language tasks to locate the target object, such as visual question answering and vision-language navigation. Humans often refer to object based on the most significant and unique characteristics. Specifically, the visual scene text on the object is the most straightforward and efficient identification in the man-made environment as shown in Figure 1. It can refer to object straightforward and efficiently, especially in scenes with dense objects, non-grid arrangements and similar appearances. As illustrated in case (a) of Figure 1, humans already number the athletes as identification to distinguish them. Thus "number 1" is the most natural referring way, rather than counting positions or extracting unique features from the crowd.

Existing methods on the REC task align sentence features with image features to find the target object. However, these methods only consider the text in expression unilaterally, but ignore the text in the image. Thus, the text from two modalities cannot map one and the other. As a result, when the referring expression involves scene text, these methods incorrectly align this scene text to basic visual features (such as color, shape and position). As shown in case (b) of Figure 1, existing models regard the text "Burger King" in the image as pixels to extract its color, shape and other visual features, without knowing that it is a store name. It causes



(a) player number 1

(b) Burger King

Figure 1: The cases of referring expression involving scene text. The target object is identified by green box.

the expression "Burger King" to be misaligned with the visual features of this text. Therefore, it is necessary to bridge the gap between text in expressions and text in images.

In this paper, we propose a novel method for referring expression comprehension that can utilize the scene text as a significant feature to ground the target object. It is composed of a feature extractor, scene text binding module and expression-guided alignment module. First, to obtain the information of scene text, the feature extractor employs an OCR mechanism to recognize the text and its position in an image. Second, the scene text binding module maps the features of scene text to the corresponding visual region, to assign text features to image-level features. At last, the expression-guided alignment model iteratively aligns the expression with fused visual features to localize the target object. Experimental results show that our method achieves improvement on RefCOCO and RefCOCOg.

Our Model

Feature Extractor

Given an image, the feature extractor employs a one-stage visual encoder to extract its grid features $G = \{g_i\}_{i=1}^{W \times H}$ of dimension C_v. To recognize the text in image, we use

*Corresponding author: Yi Cai, ycai@scut.edu.cn

the optical character recognition (OCR) approach and language encoder to extract the textual feature of N scene text words $SW = \{sw_n\}_{n=1}^{N}$ and their coordinate positions $SL = \{sl_n\}_{n=1}^{N}$. Each sw is the textual feature of dimension C. Each sl contains the vector $[x, y, w, h]$, i.e., the coordinate of scene text bounding box.

Given a referring expression, the language encoder extracts feature $S = \{s_t\}_{t=1}^{T}$ of T words, where each s_t has a dimension of C_l.

Scene Text Binding Module

In addition to the visual features (e.g., color and shape) obtained by visual encoder, this module attached high-level semantic information of scene text to the visual representation. Specifically, the relatedness of scene text and visual region is calculated according to scene text position and grid position:

$$f_i = \begin{cases} sw_n, & sl_n \in g_i \\ 0, & sl_n \notin g_i \end{cases}, \tag{1}$$

where $F = \{f_i\}_{i=1}^{W \times H}$ is the textual features of scene text in the spatial position of visual features. If the scene text is inside a grid, f_i is the textual feature of this scene text.

Then, F is mapped to dimension C_v and element-wise added to grid feature G. The text-bound visual features are:

$$FG = Mapping(f_i) + g_i. \tag{2}$$

Expression-Guided Alignment Module

Before the alignment, FG and S are mapped to the common dimension C, respectively. Following the multi-step alignment method (Yang et al. 2020), in each step, a group of words is extracted and the modulation factors are learned according to the features of these words. Then the modulation factors are used to refine the FG. The refined features of the last step will be fed to the localization method to ground the bounding box of the target object.

Experiments

Experimental Settings

Datasets. We evaluate the proposed method on two benchmarks including RefCOCO (Yu et al. 2016) and RefCOCOg (Mao et al. 2016). RefCOCO contains 19,994 images and 142,209 expressions. The referring objects are person in testA and other categories in testB. RefCOCOg is characterized by long and complex expressions. We follow the settings of RefCOCOg-umd.

Baselines. We compare our model with the state-of-the-art one-stage methods, including SSG (Chen et al. 2018), One-Stage-BERT (1-BERT) (Yang et al. 2019), ReSC-Base (ReSC) (Yang et al. 2020).

Experimental Results

The performance of accuracy is shown in Table 1. Our model has about 1% improvement on RefCOCO testA and 2% on RefCOCOg, which demonstrates the effectiveness of the proposed method. Our model can better deal with the longer and more complex expressions on the RefCOCOg. The reason is that the long sentences contain more visual information and are more likely to contain scene text information. Our method can align the text in expression with the text in image, which improves the performance. Additionally, on the RefCOCO, our model performs better when referring to people, because it is easier to recognize the scene text on people than on other objects.

Method	RefCOCO			RefCOCOg	
	val	testA	testB	val	test
SSG	-	76.51	67.50	58.80	-
1-BERT	72.05	74.81	67.59	59.03	58.70
ReSC	76.59	78.22	**73.25**	64.87	64.87
Ours	**76.67**	**79.18**	72.80	**66.48**	**66.04**

Table 1: Performance comparisons on the RefCOCO and RefCOCOg datasets. The bold values indicate the best performance, and symbol (-) indicates the unavailable results.

Conclusion

In this work, we propose a method for REC to deal with the expression involving scene text. We extract the text in image and bind its feature to the corresponding visual region according to their position relatedness. The text-bound visual features can better align with the scene text involved in expression. The relationship between multiple scene texts and visual instances will be explored in future works.

Acknowledgments

This work was supported by National Natural Science Foundation of China (62076100 and 61976094), Fundamental Research Funds for the Central Universities, SCUT (D2210010, D2200150, and D2201300), the Science and Technology Planning Project of Guangdong Province (2020B0101100002).

References

Chen, X.; Ma, L.; Chen, J.; Jie, Z.; Liu, W.; and Luo, J. 2018. Real-time referring expression comprehension by single-stage grounding network. *arXiv preprint arXiv:1812.03426*.

Mao, J.; Huang, J.; Toshev, A.; Camburu, O.; Yuille, A. L.; and Murphy, K. 2016. Generation and Comprehension of Unambiguous Object Descriptions. In *CVPR*, 11–20.

Qiao, Y.; Deng, C.; and Wu, Q. 2020. Referring Expression Comprehension: A Survey of Methods and Datasets. *IEEE Transactions on Multimedia*.

Yang, Z.; Chen, T.; Wang, L.; and Luo, J. 2020. Improving One-Stage Visual Grounding by Recursive Sub-query Construction. In *ECCV*, 387–404.

Yang, Z.; Gong, B.; Wang, L.; Huang, W.; Yu, D.; and Luo, J. 2019. A Fast and Accurate One-Stage Approach to Visual Grounding. In *ICCV*, 4682–4692.

Yu, L.; Poirson, P.; Yang, S.; Berg, A. C.; and Berg, T. L. 2016. Modeling Context in Referring Expressions. In *ECCV*, 69–85.