TextREC: A Dataset for Referring Expression Comprehension with Reading Comprehension

Chenyang Gao, Biao Yang, Hao Wang, Mingkun Yang, Wenwen Yu, Yuliang Liu(✉), and Xiang Bai

Huazhong University of Science and Technology, Wuhan, China
{m202172425,hust_byang,wanghao4659,yangmingkun,
wenwenyu,ylliu,xbai}@hust.edu.cn

Abstract. Referring expression comprehension (REC) aims at locating a specific object within a scene given a natural language expression. Although referring expression comprehension has achieved tremendous progress, most of today's REC models ignore the scene texts in images. Scene text is ubiquitous in our society, and frequently critical to understand the visual scene. To study how to comprehend scene text in the referring expression comprehension task, we collect a novel dataset, termed TextREC, in which most of the referring expressions are related to scene text. Our TextREC dataset challenges a model to recognize scene text, relate it to the referring expressions, and select the most relevant visual object. We also propose a text-guided adaptive modular network (TAMN) to comprehend scene text associated with objects in images. Experimental results reveal that current state-of-the-art REC methods fall short on the TextREC dataset, while our TAMN gets inspiring results by integrating scene text.

Keywords: Referring expression comprehension · Scene text representation · Multi-modal understanding

1 Introduction

Referring expression comprehension (REC) [17] aims at locating a specific object within a scene given a natural language expression. It is a fundamental issue in the field of human-computer interaction and also a bridge between computer vision and natural language processing. Although referring expression comprehension has achieved tremendous progress, most of today's REC models ignore the scene texts in images. However, scene text is indispensable and more natural for distinguishing different objects. Considering the situation in Fig. 1, it is difficult to detect the target man using basic visual attributes, since the players wear the same uniform and their position is constantly changing during a football match. But the target man can be easily and naturally detected with the guidance of scene text.

Scene text is ubiquitous in our society, which conveys rich information for understanding the visual scene [27]. As the COCO-Text dataset [36] suggests,

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
G. A. Fink et al. (Eds.): ICDAR 2023, LNCS 14189, pp. 402–420, 2023.
https://doi.org/10.1007/978-3-031-41682-8_25

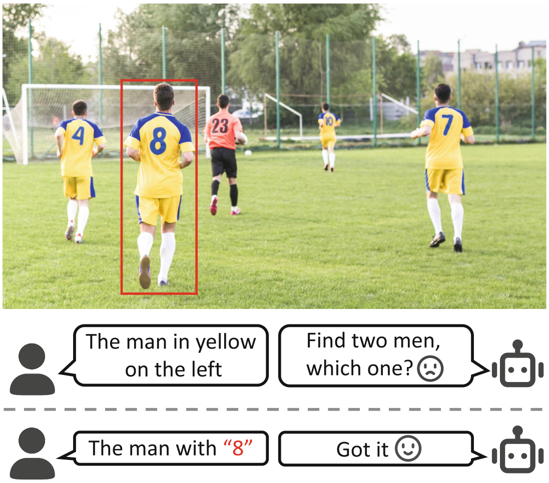


Fig. 1. This paper introduces a novel dataset to study integrating scene text in the referring expression comprehension task. For the above example, the scene text "8" provides crucial information that naturally distinguishes different players.

about 50% of the images contain scene text in large-scale datasets such as MS COCO [24] and the percentage increases sharply in urban environments. To move towards human-oriented referring expression comprehension, it is necessary to integrate scene text in existing REC pipelines. Scene text can provide more discriminative information so that the target object can be more easily specified. For example, "get a bottle of Coca-Cola from the fridge" is more precise for a robot to find the target object and more user-friendly. In literature, there are many studies successfully using scene text for vision-language tasks, *e.g.*, visual question answering [34], image captioning [33], cross-modal retrieval [27,37], and fine-grained image classification [16]. Therefore, explicitly utilizing scene text should be a natural step toward a more reasonable REC model.

To study how to comprehend scene text associated with objects in images, we collect a new dataset named TextREC. It contains 24,352 referring expressions and 36,083 scene text instances on 8,690 images, and most of the referring expressions are related to scene text. Our TextREC dataset challenges a model to recognize scene text, relate it to the referring expressions, and choose the most relevant visual object, requiring semantic and visual reasoning between multiple scene text tokens and visual entities. Besides, we also evaluate the performance of different state-of-the-art REC models, from which we observe the limited performance due to ignoring the scene texts contained in images. To this end, we propose a **T**ext-guided **A**daptive **M**odule **N**etwork (**TAMN**) to address this issue. The contributions of this paper are threefold:

- We introduce a novel dataset (TextREC) in which most of the referring expressions are related to scene text. Our TextREC dataset requires a model to leverage the additional modality provided by scene text so that the relation-

ship between the visual objects in images and the textual semantic referring expression can be identified properly.
- We propose a text-guided adaptive modular network (TAMN) to utilize scene text, relate it to the referring expressions, and select the most relevant visual object.
- Substantial experimental results on the TextREC dataset demonstrate that it is important and meaningful to take into account scene text for locating the target object, meanwhile demonstrating the excellent performance of our TAMN in this task.

2 Related Work

2.1 Referring Expression Comprehension Datasets.

To tackle the REC task, numerous datasets [3,25,28,39,42,45] have been constructed. The first large-scale REC dataset was introduced by Kazemzadeh et al. [17], which is collected by applying a two-player game named ReferIt Game on the ImageCLEF IAPR [8] dataset. Unlike ReferIt Game, RefCOCOg [28] is collected in a non-interactive setting based on the MSCOCO [24] images. RefCOCO [45] and RefCOCO+ [45] are also collected using ReferIt Game on the MSCOCO images. Due to the non-interactive setting, the referring expressions in RefCOCOg are longer and more complex than those in RefCOCO and RefCOCO+. The above datasets are collected in real-world images. While Liu et al. [25] consider using synthesized images and carefully design templates to generate referring expressions, resulting in a synthetic dataset named CLEVR-Ref+. Wang et al. [39] point out that commonsense knowledge is important to identify the objects in the images in our daily life. They also collect a dataset based on Visual Genome [18], named KB-Ref. To answer each referring expression, at least one piece of commonsense knowledge should be included. Chen et al. [3] and Yang et al. [42] adopt the expression template and scene graphs provided in [11,18] to generate referring expressions in the real-world images. Recently, Bu et al. [2] collect a dataset based on various image sources to highlight the importance of scene text.

2.2 Vision-Language Tasks with Text Reading Ability

With the maturity of reading scene text (OCR) [6,19–23,26,31,32,41,46], vision-language tasks with text reading ability become an active research field. Several existing datasets [1,29,30,34,35,40] study the task of Visual Question Answering with Text Reading Ability. These datasets require understanding the scene text in the image when answering the questions. Similarly, to enhance scene text comprehension in an image, a new task named image captioning with reading comprehension and a corresponding dataset called TextCaps [33] is proposed.

Existing works [7,10,15,33,34,38,47] propose various network architectures to utilize scene text information. LoRRA [34] adds an OCR attention branch

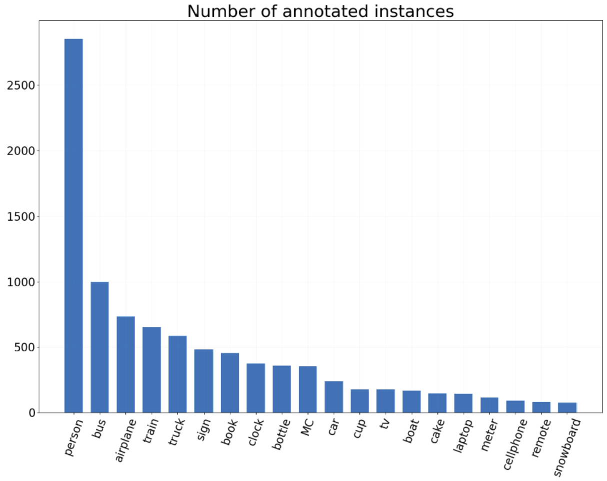


Fig. 2. Number of annotated instances per category.

to the VQA model [13], to select an answer either from a fixed vocabulary or detected OCR tokens. M4C [10] utilizes a multi-modal Transformer encoder to encode the question, image and scene text jointly, then generates answers through a dynamic pointer network. M4C-Captioner [33] directly remove question input in the aforementioned M4C model to solve the text-based image captioning task. SA-M4C [15] proposes a spatially aware self-attention layer that ensures each input focuses on a local context rather than dispersing attention amongst all other entities in a standard self-attention layer. MM-GNN [7] utilizes graph neural networks to build three separate graphs for different modalities. Then the designed aggregators use multi-modal contexts to obtain a better representation for the downstream VQA. SSBaseline [47] designs three simple attention blocks to suppress irrelevant features. LSTM-R [38] constructs the geometrical relationship between OCR tokens through the relation-aware pointer network.

3 TextREC Dataset

Our dataset enables referring expression comprehension models to conduct spatial, semantic, and visual reasoning between multiple scene text tokens and visual objects. In this section, we describe the process of constructing our TextREC dataset. We start by describing how to select the images used in TextREC. We then explain the pipeline for collecting the referring expressions related to scene texts. Finally, we provide statistics and an analysis of our TextREC.

3.1 Images

In order to make full use of the annotations of existing datasets, we rely on the MSCOCO 2014 train images (Creative Commons Attribution 4.0 License). Since

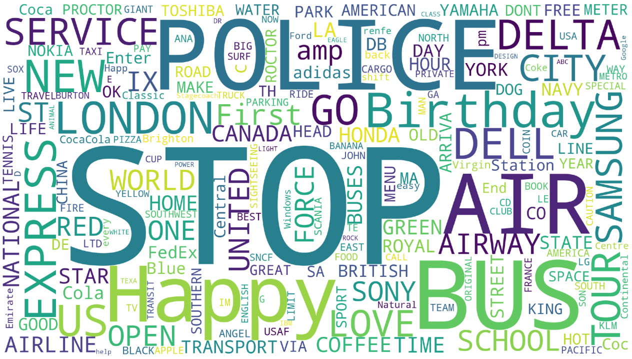


Fig. 3. Wordcloud visualization of most frequent scene text tokens contained in the referring expressions.

the goal of our dataset is to integrate scene text in existing REC pipelines, we are more interested in the images that contain scene texts. To select images containing scene texts, we use COCO-Text [36], which is a scene text detection and recognition dataset based on the MSCOCO dataset. We select images containing at least one non-empty legible scene text instance. Through the visualization of the result images, we notice that some scene text instances are too small and difficult to recognize. So we further add a constraint to the images to filter out the scene text instances with an area smaller than 100 pixels. Filtering these out results in 10,752 images, which form the basis of our TextREC dataset.

3.2 Referring Expressions

In the second stage, we collect referring expressions for objects in the above images. Different from the traditional referring expression comprehension task, in most cases, the target object can be uniquely specified with scene text. For example, if we want to ground NO.13 player in a football match, only using the number 13 on the player's clothes is sufficient. So in the referring expressions, we want to include scene text as much as possible, ignoring appearance information and location information. As a result, we choose some simple templates to generate referring expressions. We can get the bounding box of each object based on MSCOCO annotations. According to the bounding box, we can find the scene text instances contained in this bounding box. For each selected scene text, we generate referring expressions using two templates: "*The object with <OCR string> on it*" and "*The <category name> with <OCR string> on it*". Among these templates, *<OCR string>* will be replaced by the scene text instance in the images, and *<category name>* will be replaced by the category name of the object. However, the referring expressions generated through the two templates may not refer to the corresponding objects. The reason is that the scene text instance is contained in the object's bounding box but irrelevant to the object. As shown in Fig. 7, the scene text instances are contained in the

Table 1. Comparison between standard benchmarks and the proposed TextREC.

Dataset	**Total Images**	**Annotations**	
		Scene Text Related Expressions	Scene Text
ReferItGame [17]	20,000	×	×
RefCOCOg [28]	26,711	×	×
RefCOCO [45]	19,994	×	×
RefCOCO+ [45]	19,992	×	×
Clevr-ref+ [25]	85,000	×	×
KB-Ref [39]	24,453	×	×
Ref-Reasoning [42]	113,000	×	×
Cops-Ref [3]	113,000	×	×
TextREC	**8,690**	✓	✓

red bounding boxes, but irrelevant to the corresponding objects. To address this issue, we develop an annotation tool using Tkinter to check the plausibility of each referring expression. Finally, we manually filter out 48,704 valid referring expressions from 61,000 expressions.

3.3 Statistics and Analysis

Our TextREC dataset contains 8,690 images, 36,083 scene text instances, 10,450 annotated bounding boxes belonging to 50 categories and 48,704 referring expressions (each template has 24,352 referring expressions). We also compare our TextREC dataset with standard benchmarks in the referring expression comprehension task. As shown in Table 1, our dataset is the only benchmark containing both scene text related expressions and scene text annotations.

We also analyze the number of annotated instances per category to see which categories are most likely to contain scene texts. The top-20 categories and their corresponding instance numbers are shown in Fig. 2. It can be observed that the category of person is most likely to contain scene texts. This is not surprising since people usually wear clothing with various logos such as “nike” or “adidas”. The category of the vehicle also tends to contain scene texts. The bus often indicates its route using some characters, and the airplane also indicates which airline it belongs to.

Moreover, we visualize word clouds for the scene text tokens contained in the referring expressions. As shown in Fig. 3, most scene text tokens are meaningful. The most frequent word is “stop” since one category of MSCOCO is stop sign. The second most frequent word is “police” because police vehicles appear frequently in our dataset.

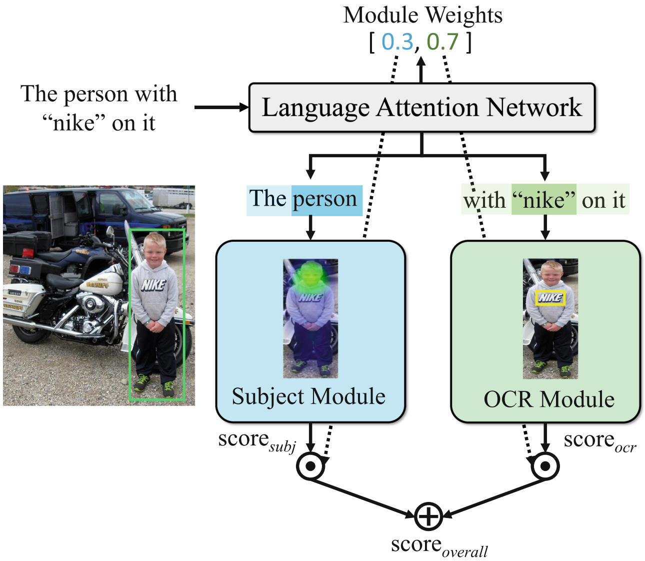


Fig. 4. Our model learns to parse an expression into the subject module and the text-guided matching module using the language attention network. Then computes an individual matching score for each module. For simplicity, we refer to the **text-guided matching module** as the **OCR module** for short.

4 Method

In this section, we introduce our Text-Guided Adaptive Modular Network (TA-MN) to align the referring expressions with the scene texts. The overall framework is shown in Fig. 4. Given a referring expression r and a candidate object o_i as input, where i represents the i-th object in the image, we start with the language attention network to parse the expressions into the subject module and the text-guided matching module. Then we use the text-guided matching module to calculate a matching score for o_i with respect to the weighted referring expression r. Finally, we take this matching score along with the score from the subject module proposed in MAttNet [44]. The overall matching score between o_i and r is the weighted combination of these two scores.

4.1 Language Attention Network

Similar to CMN [9] and MAttNet [44], we utilize the soft attention mechanism over the word sequence to attend to the relevant words automatically. As shown in Fig. 5, given a expression of T words $r = \{m_t\}_{t=1}^{T}$, we first embed each word m_t to a vector e_t using an one-hot word embedding. Then a bi-directional LSTM is applied to encode the context for each word. To get the final representation

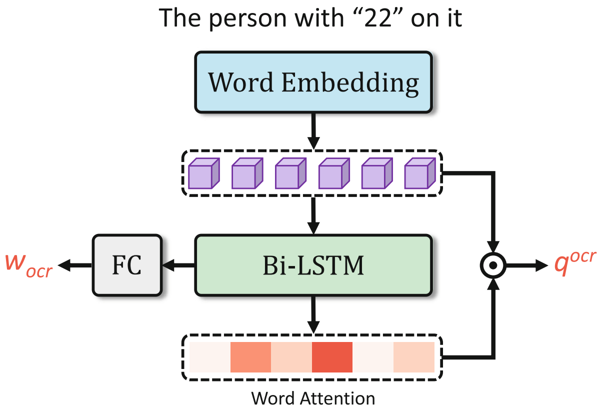


Fig. 5. The illustration of the language attention network.

for each word, we concatenate the hidden state in both directions:

$$\begin{aligned} e_t &= \text{embedding}(m_t) \\ \vec{h}_t &= \overrightarrow{\text{LSTM}}(e_t, \vec{h}_{t-1}) \\ \overleftarrow{h}_t &= \overleftarrow{\text{LSTM}}(e_t, \overleftarrow{h}_{t+1}) \\ h_t &= [\vec{h}_t, \overleftarrow{h}_t]. \end{aligned}$$

The attention weight over each word m_t for the text-guided matching module is obtained through a learned linear prediction over h_t followed by a softmax function:

$$a_t = \frac{\exp\left(\text{FC}(h_t)\right)}{\sum_{k=1}^{T} \exp\left(\text{FC}(h_k)\right)}$$

The language representation of the text-guided matching module is obtained by the weighted sum of word embeddings:

$$q^{ocr} = \sum_{t=1}^{T} a_t e_t$$

Finally, we utilize another two fully-connected layers to get the weights w_{ocr} and w_{subj} for our text-guided matching module and subject module:

$$[w_{ocr}, w_{subj}] = \text{softmax}(\text{FC}([h_0, h_T]))$$

4.2 Text-Guided Matching Module

Our text-guided matching module is illustrated in Fig. 6. Given a candidate o_i and all the ground truth scene text instances $\{p_n\}_{n=1}^{N}$ contained in the bounding box of o_i, we first encode each scene text instance p_n to a vector using the same word embedding layer of the language attention network.

$$u_n = \text{embedding}(p_n)$$

Then we compute the cosine similarity between each word embedding of the scene text instance and q^{ocr}:

$$S(u_n, q^{ocr}) = \frac{u_n^T q^{ocr}}{||u_n||||q^{ocr}||}$$

The similarity score between $\{u_n\}_{n=1}^N$ and q^{ocr} can be obtained by choosing the largest score in $\{S(u_n, q^{ocr})\}_{n=1}^N$:

$$S(u, q^{ocr}) = \max_{1 \leq n \leq N} S(u_n, q^{ocr})$$

This score is not sufficient as the matching score between o_i and r. We will illustrate the reasons with a few specific examples. As shown in Fig. 7, a scene text instance may exist both in the bounding boxes of two different objects. But it only relates to one object (green box). If we use $S(u, q^{ocr})$ as the matching score, another unrelated object (red box) may mismatch with the expression. To address this problem, the algorithm should find the association between the scene text and object. For example, "NIKE" is unlikely to appear on a motorcycle, but can appear on a person. So we further add a confidence score to $S(u, q^{ocr})$:

$$S(f_{obj}, q^{ocr}) = \frac{f_{obj}^T q^{ocr}}{||f_{obj}||||q^{ocr}||} \tag{1}$$

where f_{obj} is the visual representation of the candidate object extracted in the subject module. This confidence score can drive the model to learn the association between the scene text and object.

The final matching score of our text-guided matching module can be obtained by multiplying $S(u, q^{ocr})$ with its confidence score:

$$S(o_i|q^{ocr}) = S(f_{obj}, q^{ocr})S(u, q^{ocr})$$

4.3 Learning Objective

Assume we get $S(o_i|q^{ocr})$ and $S(o_i|q^{subj})$ from our proposed text-guided matching module and the subject module proposed in MAttNet [44]. We also get the module weights w_{ocr} and w_{subj} for the text-guided matching module and the subject module in the language attention network. The overall matching score for candidate object o_i and referring expression r is:

$$S(o_i|r) = w_{ocr}S(o_i|q^{ocr}) + w_{subj}S(o_i|q^{subj})$$

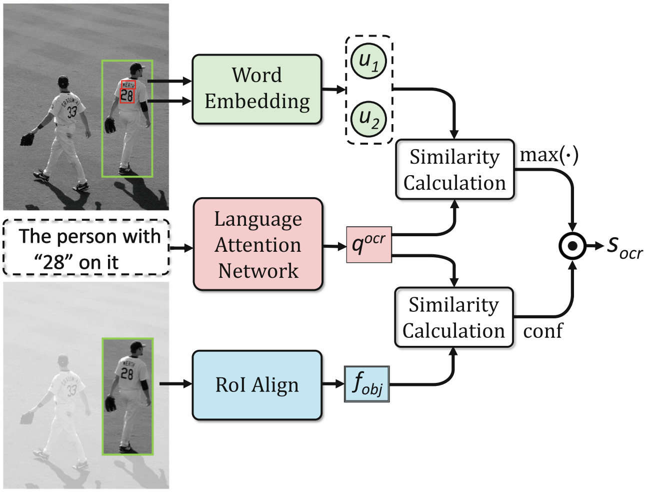


Fig. 6. The illustration of proposed text-guided matching module, "conf" refers to the confidence score calculated in Eq. 1.

Inspired by the triplet loss for the image retrieval task, for each positive pair (o_i, r_i), we randomly sample two negative pairs (o_i, r_j) and (o_k, r_i). r_j is the expression matched with other object in the same image of o_i, and o_k is other object in the same image of r_i. The combined hinge loss is calculated as follows:

$$\begin{aligned} L_{rank}^{overall} &= \sum_i \lambda_1[\delta + S(o_i|r_j) - S(o_i|r_i)]_+ \\ &+ \sum_i \lambda_2[\delta + S(o_k|r_i) - S(o_i|r_i)]_+ \end{aligned}$$

where δ is a margin hyper-parameter and $[\cdot]_+ = \max(\cdot, 0)$. To stabilize the training procedure, we further add a hinge loss to the text-guided matching module:

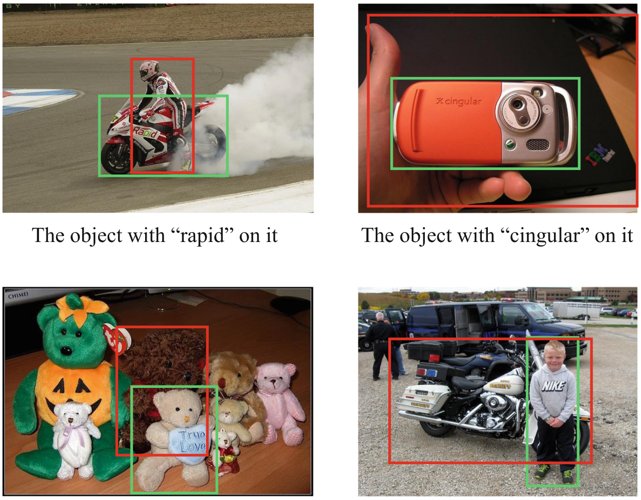
$$\begin{aligned} L_{rank}^{ocr} &= \sum_i \lambda_3[\delta + S(o_i|q_j^{ocr}) - S(o_i|q_i^{ocr})]_+ \\ &+ \sum_i \lambda_4[\delta + S(o_k|q_i^{ocr}) - S(o_i|q_j^{ocr})]_+ \end{aligned}$$

The final loss function is summarized as follows:

$$L = L_{rank}^{ocr} + L_{rank}^{overall}$$

5 Experiment

In this section, we first introduce the experiment setting. Then we evaluate the TAMN and several state-of-the-art REC methods on our TextREC dataset.



The object with “rapid” on it

The object with “cingular” on it

The teddy bear with “love” on it

The object with “nike” on it

Fig. 7. The motivation of adding the confidence score in our OCR module.

Furthermore, we conduct ablation studies to demonstrate the effectiveness of each component in our TAMN. We also explore more templates and a new test setting. Finally, the attention weights for each word in the referring expressions are visualized to demonstrate the effectiveness of the language attention network.

5.1 Dataset and Evaluation Protocol

We evaluate our text-guided adaptive modular network on the TextREC dataset. From Fig. 2, it can be observed that the categories of the dataset follow a long-tailed distribution. To ensure that the test set contains rare categories, we divide our dataset according to the ratio of instances of each category to the total, resulting in train and test splits with image numbers 7,422 and 1,268.

Following the standard evaluation setting [28], we compute the Intersection over Union (IoU) ratio between the ground truth and predicted bounding box. We regard the detection as a true positive If IoU is greater than 0.5, otherwise it is a false positive. For each image, we then compute the precision@1 measure according to the confidence score. The final performance is obtained by averaging these scores over all images.

5.2 Implementation Details

The detection model we adopt is Mask R-CNN. We follow the same implementation as MattNet [44]. The detection model is trained on a union of MSCOCO’s 80k train and 35k subset of val (trainval35k) images excluding the test images in our TextREC dataset. We use the ground truth bounding boxes during training. In the test stage, we utilize the Mask R-CNN mentioned above to generate boxes. Our model is optimized with Adam optimizer and the batch size is set to 15. The initial learning rate is 0.0004. Moreover, the model is trained for 50 epochs with a learning rate decay by a factor of 2 every 16 epochs. The size of the word embedding and the hidden state of the bi-LSTM is set to 512. The size of the word embedding for the scene text is also set to 512. We set the output of all fully-connected layers within our model to be 512-dimensional. For the hyper-parameters in the loss functions, we set $\lambda_1 = 1$ and $\lambda_2 = 1$ in $L_{rank}^{overall}$. In addition, we set $\lambda_3 = 1$ and $\lambda_4 = 1$ in L_{rank}^{ocr}.

Table 2. Performance of the baselines on our TextREC dataset. TAMN significantly benefits from scene text input and achieves the highest precision@1 (%) score, suggesting that it is important to integrate scene text for the referring expression comprehension task.

Model	**Template1**	**Template2**
TransVG [4]	50.1	54.0
MAttNet [44]	52.3	60.5
QRNet [43]	52.7	59.1
Mdetr [14]	54.4	63.3
TAMN (ours)	**77.8**	**80.8**

5.3 Performance of the Baselines on TextREC Dataset

To illustrate the gap between the traditional REC datasets and our TextREC dataset, we conduct experiments with different state-of-the-art REC methods. As shown in Table 2, current state-of-the-art methods [4,14,43,44] fall short on our TextREC dataset. The results indicate that these methods ignore scene text in images, while our TAMN gets inspiring results by integrating scene text. This clearly verifies that it is important and meaningful to take into account scene text for the referring expression comprehension task.

5.4 Ablation Studies

The Subject Module and OCR Module. As shown in Fig. 4, our TAMN consists of two modules: the subject module and OCR module. We test the performance only with each module and the results are shown in Table 3. Compared with the only subject module, adding our OCR module givens 26.4 and 20.2 performance improvement in template1 and template2, respectively. Cooperating with our OCR module, the subject module gives 1.5 and 2.7 performance improvement in template1 and template2, respectively. These verify the effectiveness of the subject module and OCR module. Moreover, for our TAMN,

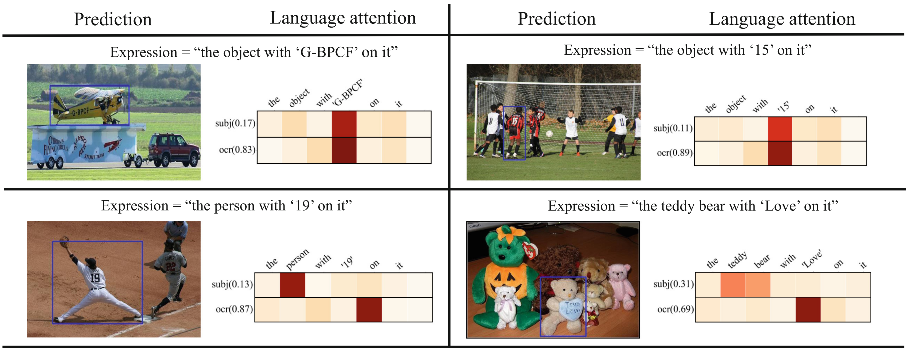


Fig. 8. The visualization results of the word attention in the language attention network.

Table 3. Ablation studies on different modules in our framework. The precision@1 (%) is reported.

Subject	OCR	Template1	Template2
✓	×	51.4	60.6
×	✓	76.3	78.1
✓	✓	**77.8**	**80.8**

Table 4. Ablation studies on different OCR systems. "GT" denotes using grounding truth scene text annotations.

Model	Template1	Template2
PaddleOCR [5]	63.2	69.8
EasyOCR [12]	67.1	72.7
GT	77.8	80.8

we compute the similarity score of each module to the overall score over the whole test set. The experimental results are summarized as follows: in template1, our OCR module makes the dominating contribution (**97.1%**) to the overall score. The contribution (**2.90%**) of the subject module can be ignored. When the expression form transfers to template2, the contribution of our OCR module decreases from **97.1%** to **70.0%**. While the contribution of the subject module increases from **2.90%** to **30.0%**. Our OCR module still accounts for the majority. The reason is that scene texts provide more information than the object categories in most cases. These clearly demonstrate the effectiveness of our proposed OCR module.

The Confidence Score in Our OCR Module. As shown in Fig. 6, we add a confidence score by calculating the similarity between the RoI feature of the candidate object and the scene text embedding. To verify the effectiveness of this confidence score, we conduct ablation experiments which are shown in Table 5. It can be observed that adding the confidence score gains 1.8% and 3.3% performance improvement in template1 and template2 only using the OCR module. We also test the effectiveness of adding the confidence score in our whole framework. It can be observed that adding the confidence score gains 1.9 and 1.3 performance improvement in template1 and template2. These results clearly verify the effectiveness of adding this confidence score.

Table 5. Ablation studies on the confidence score in our OCR module. The precision@1 (%) is reported.

Model	Confidence	Template1	Template2
OCR	×	74.5	74.8
OCR	✓	76.3	78.1
TAMN	×	75.9	79.5
TAMN	✓	77.8	80.8

Different OCR Systems. We conduct ablation studies to see the performance using different OCR systems. Results in Table 4 show that the performance of scene text detection and recognition methods has a great impact on the final results. The reason why EasyOCR has better performance is that the text spotting precision of EasyOCR is 6.8% higher than that of PaddleOCR.

Templates in Different Forms. We conduct ablation studies to see the performance using different templates. As shown in Table 6, it can be observed that the performance is very close with different templates as long as they contain the same amount of information (<category name> or <OCR string>). For example, in row 1, 3, and 5, the performance differences are within 0.3 in terms of precision@1 measure. Similarly, in row 2 and 4, the performance differences are also within 0.3.

Table 6. Ablation studies on the templates in different forms. The precision@1 (%) is reported.

Templates	Pre@1
The object with <OCR string> on it	77.8
The <category name> with <OCR string> on it	**80.8**
Object with <OCR string>	77.6
<category name> with <OCR string>	80.5
<OCR string>	77.5
The object	51.2

New Test Setting. In the traditional referring expression comprehension datasets, one referring expression only has one corresponding bounding box in an image. However, in our TextREC dataset, one referring expression can have multiple corresponding bounding boxes. For example, we may ask “The object with ‘police’ on it”, there can be more than one police car in the image. It is necessary to find all the objects that match the description. Therefore, we propose a new test setting that calculates the precision, recall, and F1 score. This can be done by setting a threshold on the confidence of all detected bounding boxes. We set 0.75 for template1 and 0.35 for template2 due to their different score distributions. Then we take the selected boxes to match the ground truth bounding boxes to get the true positives, false positives, and false negatives. We test our TAMN on this new setting and the results are shown in Table 7. We believe this new setting can offer more comprehensive evaluations on the models.

Table 7. The performance of our TAMN in the new test setting. The precision, recall and F1-Score (%) are reported.

Template	**Threshold**	**Precision**	**Recall**	**F1-Score**
Template1	0.70	76.8	75.2	76.0
	0.75	78.6	73.8	76.1
	0.80	80.1	72.2	75.9
Template2	0.30	73.0	83.8	78.1
	0.35	81.8	76.2	78.9
	0.40	86.7	66.2	75.1

5.5 Visualization Analysis

To verify the effectiveness of the language attention network. We visualize the attention weight for each word in the referring expressions. As shown in Fig. 8, both the subject module and the OCR module focus on the scene texts in template1. When the expression form transfers to template2, the OCR module still focuses on the scene texts. However, the subject module changes to focus on the category name. For example, in the sentence "the object with '15' on it", the subject module focuses on the "15". While, it focuses on the "person" in the sentence "the person with '19' on it". It is reasonable since the only discriminate information is the scene text in template1.

6 Conclusion

In this paper, we point out that most of the existing REC models ignore scene text which is naturally and frequently employed to refer to objects. To address this issue, we construct a new dataset termed TextREC, which studies how to comprehend the scene text associated with objects in an image. We also propose a text-guided adaptive modular network (TAMN) that explicitly utilizes scene text, relates it to the referring expressions, and chooses the most relevant visual object. Experimental results on the TextREC dataset show that the current state-of-the-art REC methods fail to achieve the expected results, but our TAMN achieves excellent results. The ablation studies also show that it is important to take into account scene text for the referring expression comprehension task.

Acknowledgements. This work was supported by the National Science Fund for Distinguished Young Scholars of China (Grant No.62225603), the Young Scientists Fund of the National Natural Science Foundation of China (Grant No.62206103), and the National Natural Science Foundation of China (No.622061 04).

References

1. Biten, A.F., et al.: Scene text visual question answering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4291–4301 (2019)
2. Bu, Y., et al.: Scene-text oriented referring expression comprehension. IEEE Transactions on Multimedia (2022)
3. Chen, Z., Wang, P., Ma, L., Wong, K.Y.K., Wu, Q.: Cops-Ref: a new dataset and task on compositional referring expression comprehension. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10086–10095 (2020)
4. Deng, J., Yang, Z., Chen, T., Zhou, W., Li, H.: TransVG: end-to-end visual grounding with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1769–1779 (2021)
5. Du, Y., et al.: PP-OCR: a practical ultra lightweight OCR system. arXiv preprint arXiv:2009.09941 (2020)
6. Fang, S., Xie, H., Wang, Y., Mao, Z., Zhang, Y.: Read like humans: autonomous, bidirectional and iterative language modeling for scene text recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7098–7107 (2021)
7. Gao, D., Li, K., Wang, R., Shan, S., Chen, X.: Multi-modal graph neural network for joint reasoning on vision and scene text. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12746–12756 (2020)
8. Grubinger, M., Clough, P., Müller, H., Deselaers, T.: The IAPR TC-12 benchmark: a new evaluation resource for visual information systems. In: International Workshop ontoImage, vol. 2 (2006)
9. Hu, R., Rohrbach, M., Andreas, J., Darrell, T., Saenko, K.: Modeling relationships in referential expressions with compositional modular networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1115–1124 (2017)
10. Hu, R., Singh, A., Darrell, T., Rohrbach, M.: Iterative answer prediction with pointer-augmented multimodal transformers for textVQA. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9992–10002 (2020)
11. Hudson, D.A., Manning, C.D.: GQA: a new dataset for compositional question answering over real-world images **3**(8). arXiv preprint arXiv:1902.09506 (2019)
12. JaidedAI: EasyOCR (2022). https://github.com/JaidedAI/EasyOCR
13. Jiang, Y., Natarajan, V., Chen, X., Rohrbach, M., Batra, D., Parikh, D.: Pythia v0. 1: the winning entry to the VQA challenge 2018. arXiv preprint arXiv:1807.09956 (2018)

14. Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., Carion, N.: MDETR-modulated detection for end-to-end multi-modal understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1780–1790 (2021)
15. Kant, Y., et al.: Spatially aware multimodal transformers for TextVQA. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12354, pp. 715–732. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58545-7_41
16. Karaoglu, S., Tao, R., Gemert, J.C.V., Gevers, T.: Con-text: text detection for fine-grained object classification. IEEE Trans. Image Process. **26**, 3965–3980 (2017)
17. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.: ReferitGame: referring to objects in photographs of natural scenes. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, pp. 787–798 (2014)
18. Krishna, R., et al.: Visual genome: connecting language and vision using crowdsourced dense image annotations. Int. J. Comput. Vision **123**(1), 32–73 (2017)
19. Liao, M., Pang, G., Huang, J., Hassner, T., Bai, X.: Mask TextSpotter v3: segmentation proposal network for robust scene text spotting. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12356, pp. 706–722. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58621-8_41
20. Liao, M., Shi, B., Bai, X.: Textboxes++: a single-shot oriented scene text detector. IEEE Trans. Image Process. **27**(8), 3676–3690 (2018)
21. Liao, M., Shi, B., Bai, X., Wang, X., Liu, W.: Textboxes: a fast text detector with a single deep neural network. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 31 (2017)
22. Liao, M., Wan, Z., Yao, C., Chen, K., Bai, X.: Real-time scene text detection with differentiable binarization. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 11474–11481 (2020)
23. Liao, M., Zou, Z., Wan, Z., Yao, C., Bai, X.: Real-time scene text detection with differentiable binarization and adaptive scale fusion. IEEE Trans. Pattern Anal. Mach. Intell. **45**(1), 919–931 (2022)
24. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
25. Liu, R., Liu, C., Bai, Y., Yuille, A.L.: CLEVR-Ref+: diagnosing visual reasoning with referring expressions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4185–4194 (2019)
26. Lyu, P., Liao, M., Yao, C., Wu, W., Bai, X.: Mask TextSpotter: an end-to-end trainable neural network for spotting text with arbitrary shapes. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018. LNCS, vol. 11218, pp. 71–88. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01264-9_5
27. Mafla, A., de Rezende, R.S., G'omez, L., Larlus, D., Karatzas, D.: StacMR: scene-text aware cross-modal retrieval. In: 2021 IEEE Winter Conference on Applications of Computer Vision, pp. 2219–2229 (2021)
28. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A.L., Murphy, K.: Generation and comprehension of unambiguous object descriptions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11–20 (2016)
29. Mathew, M., Karatzas, D., Jawahar, C.: DocVQA: A dataset for VQA on document images. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2200–2209 (2021)

30. Mishra, A., Shekhar, S., Singh, A.K., Chakraborty, A.: OCR-VQA: visual question answering by reading text in images. In: 2019 International Conference on Document Analysis and Recognition, pp. 947–952. IEEE (2019)
31. Shi, B., Bai, X., Yao, C.: An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE Trans. Pattern Anal. Mach. Intell. **39**(11), 2298–2304 (2016)
32. Shi, B., Yang, M., Wang, X., Lyu, P., Yao, C., Bai, X.: ASTER: an attentional scene text recognizer with flexible rectification. IEEE Trans. Pattern Anal. Mach. Intell. **41**(9), 2035–2048 (2018)
33. Sidorov, O., Hu, R., Rohrbach, M., Singh, A.: TextCaps: a dataset for image captioning with reading comprehension. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12347, pp. 742–758. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58536-5_44
34. Singh, A., et al.: Towards VQA models that can read. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8317–8326 (2019)
35. Tanaka, R., Nishida, K., Yoshida, S.: VisualMRC: machine reading comprehension on document images. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 13878–13888 (2021)
36. Veit, A., Matera, T., Neumann, L., Matas, J., Belongie, S.: Coco-text: Dataset and benchmark for text detection and recognition in natural images. arXiv preprint arXiv:1601.07140 (2016)
37. Wang, H., Bai, X., Yang, M., Zhu, S., Wang, J., Liu, W.: Scene text retrieval via joint text detection and similarity learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4556–4565 (2021)
38. Wang, J., Tang, J., Yang, M., Bai, X., Luo, J.: Improving OCR-based image captioning by incorporating geometrical relationship. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1306–1315 (2021)
39. Wang, P., Liu, D., Li, H., Wu, Q.: Give me something to eat: referring expression comprehension with commonsense knowledge. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 28–36 (2020)
40. Wang, X., et al.: On the general value of evidence, and bilingual scene-text visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10126–10135 (2020)
41. Yang, M., et al.: Reading and writing: Discriminative and generative modeling for self-supervised text recognition. In: Proceedings of the 30th ACM International Conference on Multimedia, pp. 4214–4223 (2022)
42. Yang, S., Li, G., Yu, Y.: Graph-structured referring expression reasoning in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9952–9961 (2020)
43. Ye, J., et al.: Shifting more attention to visual backbone: query-modulated refinement networks for end-to-end visual grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15502–15512 (2022)
44. Yu, L., et al.: MattNet: modular attention network for referring expression comprehension. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1307–1315 (2018)

45. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 69–85. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_5
46. Yu, W., Liu, Y., Hua, W., Jiang, D., Ren, B., Bai, X.: Turning a clip model into a scene text detector. arXiv preprint arXiv:2302.14338 (2023)
47. Zhu, Q., Gao, C., Wang, P., Wu, Q.: Simple is not easy: a simple strong baseline for textvqa and textcaps. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 3608–3615 (2021)