Iterative Robust Visual Grounding with Masked Reference based Centerpoint Supervision

Menghao Li[1] * Chunlei Wang[1] * Wenquan Feng[1] Shuchang Lyu[1]

Guangliang Cheng[2] ✉ Xiangtai Li[3] Binghao Liu[1] Qi Zhao[1] ✉

[1] Beihang University [2] University of Liverpool [3] S-Lab, Nanyang Technological University

{sy2102227, wcl_buaa, buaafwq, lyushuchang, liubinghao, zhaoqi}@buaa.edu.cn

{Guangliang.Cheng}@liverpool.ac.uk {xiangtai.li}@ntu.edu.sg

Methods	RefCOCO			RefCOCO+			RefCOCOg	
	val	testA	testB	val	testA	testB	val-u	test-u
Baseline	85.92	88.41	81.77	75.27	80.06	66.33	77.10	76.06
Wo masked	86.57	88.52	82.26	75.84	80.41	67.02	77.57	75.93
Ours	**86.82**	**88.75**	**82.60**	**76.22**	**80.75**	**67.33**	**77.86**	**76.24**

Table 1. Ablation study on multiple masked strategies. "Baseline" denotes the experiment with one full text without centerpoint supervision, "Wo masked" denotes the result with one full text and centerpoint supervision, and "Ours" represents the experiment with MRCS.

1. Detailed rules for masking words in MRCS module.

When we mask lexical words, we prioritize them differently. We first mask prepositions, conjunctions, and qualifiers because they usually do not significantly impact the sentence's meaning. If these types of words are not present, the module then masks auxiliaries, pronouns, and numbers, which can partly affect the sentence's semantics. Finally, the module masks adjectives and verbs, which are critical for the sentence's meaning. If there is only one non-noun word remaining or only nouns remain in the sentence, no further masking is performed. However, even with this priority order, some important words may still get masked, introducing noise into the training. Nevertheless, we empirically demonstrate that the language comprehension improvement from masking operations outweighs the negative effects of introducing noise (shown in Table 1). In all datasets, the number of words exceeds 3, and through three masking operations, we find that the majority of the masked words are prepositions, conjunctions, and qualifiers. Therefore, in most cases, this operation will not affect the meaning of the sentence.

2. Details of Dataset

To comprehensively verify the effectiveness of the proposed robust VG approach, we evaluate it on two types of datasets: the regular VG datasets and the robust VG datasets.

2.1. Regular VG Datasets

We evaluate our proposed approach on five regular VG datasets, including the RefCOCO [6], RefCOCO+ [6], RefCOCOg [2], ReferItGame [1], and Flickr30k [3]. The RefCOCO datasets series, including RefCOCO, RefCOCO+, and RefCOCOg, are three commonly used benchmarks for visual grounding, the images used in these datasets are collected from the train2014 set of MSCOCO dataset. Specifically, the RefCOCO dataset contains 19,994 images, 50,000 reference objects, and a total of 142,210 reference expressions. Among them, 120,624 reference expressions are used as the training set, 10,834 as the validation set, 5657 and 5095 expressions for test A and test B, respectively. The RefCOCO+ dataset provides 19,992 images with 49,856 reference objects and 141,564 reference expressions. Similar to RefCOCO, RefCOCO+ is also divided into training, validation, test A, and test B sets, with 120,191, 10,758, 5,726, and 4,889 reference expressions in these datasets. RefCOCOg contains a total of 25,799 images, 49,822 objects, and 95,010 reference expressions. Compared to the first two datasets, most of the expressions in RefCOCOg have longer sentences and more complex statement structures. RefCOCOg contains two sub-datasets, RefCOCOg-google and RefCOCOg-umd. Since the former dataset does not provide a test set, we mainly use the RefCOCOg-umd dataset. ReferItGame contains 20,000 images, which are collected from the SAIAPR-12 dataset. This dataset has a total of 120,072 reference expressions and is divided into

*Contribute Equally.

a training set with 54,127 reference expressions, a validation set with 5,842 reference expressions, and a test set with 60,103 reference expressions. Flickr30k contains 31,783 images and 427,000 reference expressions. We divide the training, validation, and test sets using the same ratio as the previous work.

2.2. Robust VG Datasets

We construct two robust VG datasets based on the existing benchmarks RefCOCOg and ReferItGame, termed RefCOCOg_F and ReferItGame_F. The train set of our robust VG datasets contains two parts of data, the first part is the train set of the original dataset, while the second part is a random matching dataset, which destroys the correspondence between the image information and the language descriptions. Specifically, for each target on the image, we select one description that is different from its original one among all the text descriptions in the dataset, thus building a dataset where the image is with irrelevant or inaccurate descriptions. During training, the ratio of these two parts of data is 1:1. The test set of our robust VG datasets also consists of two parts of data, the first part is the test set of the original dataset while the second part is the manually modified robust VG dataset, which requires manual intervention to modify some keywords in the descriptions, thus modifying the semantics of the descriptions and building a more difficult dataset. For instance, we manually modify the expression “The man in white T-shirt is riding a bike” to “The man in blue T-shirt is riding a bike”. Specifically, the test set of the RefCOCOg_F dataset contains 2000 pairs of false-alarm data and 9602 pairs of regular data that are from the original RefCOCOg test set. The test set of the ReferItGame_F dataset contains 1000 pairs of false-alarm data and 9000 pairs of regular data that are randomly sampled from the test set of the original ReferItGame dataset.

Specifically, the data combination method of the random matching dataset is to randomly replace the description in each group of data in the training set with a random other description in the dataset to construct false-alarm data. Of course, the description of the same image will not be selected to avoid the existence of the target corresponding to the ran71 dom description on the image. It can be observed that the probability of the existence of the target corresponding to the description on the image is very low for the false alarm data formed by this random selection description method.

We build the manually modified robust VG dataset by manually modifying some keywords in the description. In general, we mainly modify words from the following perspectives. First, modifying key nouns can greatly change the semantics of words, thus generating false alarm data. For example, modify ”Two men on a horse” to ”Two men on a car” (as shown in the first row of Fig.1). Second, modifying key adjectives can also change the description semantics. For example, modify ”A man with a bat wearing a red helmet” to ”A man with a bat wearing a yellow helmet” (as shown in the second row of Fig.1). Third, modify words in the text that relate to spatial location can mismatch the original target with the newly generated text. For example, modify ”An elephant trainer standing beside an elephant walking down the street” to ”An elephant trainer standing far away from an elephant walking down the street” (as shown in the third row of Fig.1). Fourth, changing the words corresponding to some fine-grained features can generate false-alarm data. For example, modify ”A man wearing glasses” to ”A man without glasses” (as shown in the fourth row of Fig.1). Experiments show that our pro95 posed IR-VG is effective for all four types of false alarm data.

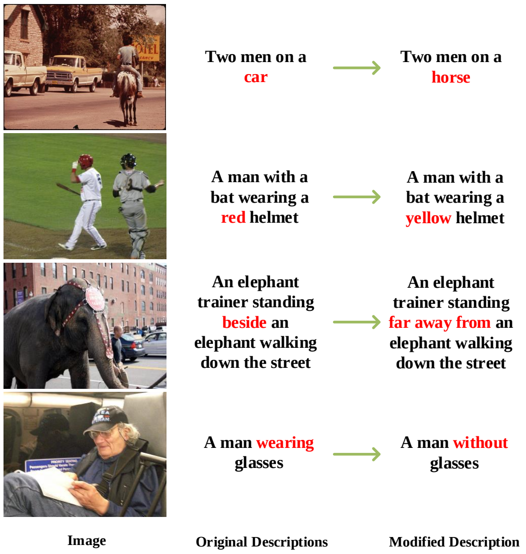
Two men on a car
Two men on a horse
A man with a bat wearing a red helmet
A man with a bat wearing a yellow helmet
An elephant trainer standing beside an elephant walking down the street
An elephant trainer standing far away from an elephant walking down the street
A man wearing glasses
A man without glasses
Image
Original Descriptions
Modified Description

Figure 1. Example of manually modified false-alarm data.

2.3. Training Loss

In the training stage, the proposed VG framework is trained end-to-end using the aforementioned losses. The overall loss function for the proposed framework is $\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}} + \lambda_{\text{key}}\mathcal{L}_{\text{key}}$ as follows, where $\mathcal{L}_{\text{cls}}$, $\mathcal{L}_{\text{reg}}$, and $\mathcal{L}_{\text{key}}$ denote the classification loss, regression loss and centerpoint loss, respectively. λ_{reg} and λ_{key} are introduced to balance the above losses.We empirically set λ_{reg} and λ_{key} as 2 and 5 by default.

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}} + \lambda_{\text{key}}\mathcal{L}_{\text{key}}, \tag{1}$$

Specifically the classification loss $\mathcal{L}_{\text{cls}}$ and the regression loss $\mathcal{L}_{\text{reg}}$ are defined as,

$$\mathcal{L}_{\text{cls}} = \sum_{t=1}^{N}\sum_{i=1}^{K} \text{CELoss}(y^t, \hat{y}_i^t), \tag{2}$$

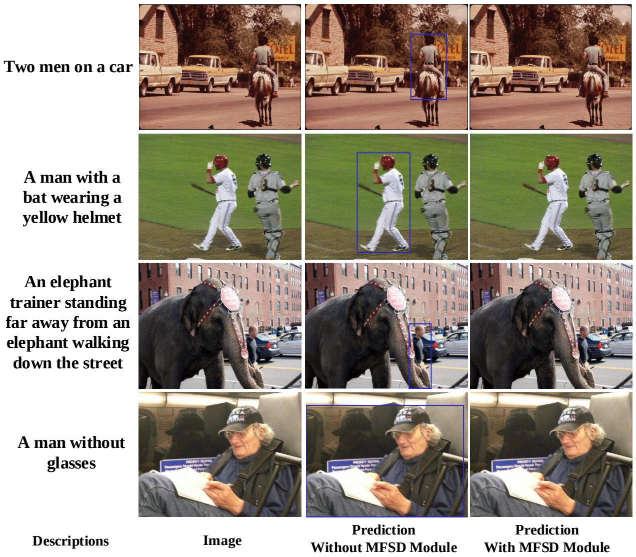
Two men on a car
A man with a bat wearing a yellow helmet
An elephant trainer standing far away from an elephant walking down the street
A man without glasses
Descriptions
Image
Prediction Without MFSD Module
Prediction With MFSD Module

Figure 2. Visualization of the MFSD module.

$$\mathcal{L}_{\text{reg}} = \sum_{t=1}^{N}\sum_{i=1}^{K} \lambda_{\text{GIOU}}\mathcal{L}_{\text{GIOU}}(b^t,\hat{b}_i^t) + \lambda_{\text{L1}}\mathcal{L}_{\text{L1}}(b^t,\hat{b}_i^t), \quad (3)$$

where $\text{CELoss}(\cdot,\cdot)$, $\mathcal{L}_{\text{GIOU}}(\cdot,\cdot)$ and $\mathcal{L}_{\text{L1}}(\cdot,\cdot)$ are the cross entropy loss, GIOU loss [4] and L1 loss, respectively. y^t and $\hat{y}_i^t$ denote the ground truth label and predicted result in t-th iteration. Similarly, b^t and $\hat{b}_i^t$ denote the ground truth bbox and predicted bbox. t denotes the t-th iteration, and i represents the i-th bbox. λ_{GIOU} and λ_{L1} are empirically adjusted, here we set them as 3 and 7 by default for all the following experiments.

3. Qualitative Analysis of MFSD

Fig. 2 illustrates the visualization of prediction results with or without the MFSD module on the robust VG datasets. It shows that the MFSD module enables the model to efficiently identify the presence or absence of targets described in the text on the image. The first row of the figure shows the false alarm data generated by the key nouns in the description being changed, the second row shows the false alarm data generated by the modification of key adjectives (e.g., color). The third line of the figure shows the spatial location relations in the description being modified and the fourth row of the figure shows the fine-grained features in the description being modified. Our MFSD module can effectively identify the false alarm data generated by all the above modification methods.

4. Details of determining false-alarm detection of the previous method

In the previous methods, we follow the same rules as ours to obtain the false alarm. Firstly, we achieve the top1 scoring box as the final prediction box. Then, we calculate the IOU value with the ground truth box. If the IOU value is greater than 0.5, we consider it a true positive, otherwise, we treat it as a false positive. However, the proposed method differs in that it combines the top1 scoring box and its existing result (exist or non-exist) to achieve the final prediction box. During our experiments, we attempted to add an irrelevant text reference head to some previous networks, such as VLTVG [5] but the results were inferior to their baselines. It may not be fair to compare these results in the paper, thus we do not show these results.

References

[1] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L. Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014. 1

[2] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 2016. 1

[3] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *CVPR*, 2017. 1

[4] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian D. Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *CVPR*, 2019. 3

[5] Li Yang, Yan Xu, Chunfeng Yuan, and et al. Improving visual grounding with visual-linguistic verification and iterative reasoning. In *CVPR*, 2022. 3

[6] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions. In *ECCV*, 2016. 1