Revisiting Counterfactual Problems in Referring Expression Comprehension

Supplementary Material

Table 1. Acc-Cls (%) on seven attributes in C-RefCOCO/+/g. “0” indicates that no counterfactual samples are generated based on the corresponding attribute.

Attribute		A1	A2	A3	A4	A5	A6	A7
C-RefCOCO	*val*	93.05	85.07	65.82	93.51	0	0	78.99
	testA	92.14	87.55	75.00	94.11	35.56	31.11	75.36
	testB	91.67	87.05	81.97	93.57	55.81	55.81	78.13
C-RefCOCO+	*val*	91.43	92.60	80.89	86.96	0	0	81.84
	testA	88.18	92.09	78.67	79.86	0	0	75.00
	testB	85.20	89.14	84.44	88.69	0	0	85.58
C-RefCOCOg	*val*	89.78	87.57	77.19	90.45	66.67	44.44	69.23
	test	88.42	88.63	78.15	85.86	71.43	80.95	76.74

Table 2. Acc-Cls (%) on positive samples in C-RefCOCO/+/g.

Model	C-RefCOCO			C-RefCOCO+			C-RefCOCOg	
	val	*testA*	*testB*	*val*	*testA*	*testB*	*val*	*test*
Ours	94.33	94.37	93.19	93.46	94.14	92.18	90.42	90.82

1. More Results on C-RefCOCO/+/g

To better demonstrate the characteristic of our datasets C-RefCOCO/+/g, we report the counterfactual classification performance of our model in details, shown as follows.

Table 1 reports the metric Acc-Cls on seven pre-defined attributes of C-RefCOCO/+/g. Note that there are some splits that do not contain certain categories of attribute words, such as A5 (relative location relation) and A6 (relative location object). Thus, their Acc-Cls are 0. We observe that there is a performance difference among these attributes. Usually A1 (head noun) achieves the highest accuracy while A6 achieves the lowest accuracy. This result indicates that head noun is more prominent for detecting fine-grained counterfactual samples, and counterfactual relative location is more confusing for C-REC models to detect.

Table 2 reports the metric Acc-Cls of our model on the positive samples in C-RefCOCO/+/g. Our model achieves over 90% accuracy on all splits of C-RefCOCO/+/g. This shows that our model keeps a strong perception on normal REC samples.

In addition, the word cloud of attribute words in C-RefCOCO/+/g is shown in Figure 1.

2. Results on Coarse-grained and Fine-grained Counterfactual Samples

2.1. Results on GRES datasets

Ref-ZOM [1] and gRefCOCO [2] are two GRES datasets. Generalized referring expression segmentation (GRES) is another subtask of RES, which considers multi-target, one-target and no-target settings. Among them, no-target setting is actually same as the counterfactual setting we study on, although C-REC is based on REC task. For counterfactual setting, gRefCOCO either manually annotates or selects samples from deceptive expressions in RefCOCO, while Ref-ZOM randomly selects images and expressions from different REC datasets and manually double checks them to guarantee no-target. On the image-text relevancy, gRefCOCO and our datasets C-RefCOCO/+/g are fine-grained, while Ref-ZOM is coarse-grained.

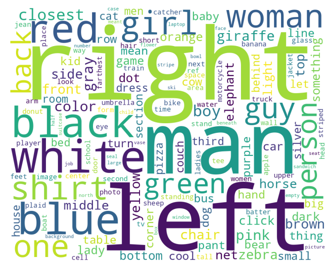

Figure 1. Word cloud of attribute words in C-RefCOCO/+/g.

We conduct experiments on the no-target part of gRefCOCO and Ref-ZOM. We follow their metrics for fair comparison. The results are shown in Table 3 and Table 4. Our model achieves competitive performance on gRefCOCO and a new state-of-the-art performance on Ref-ZOM. This indicates our fine-grained counterfactual resilient framework can easily detect coarse-grained counterfactual queries, but there is still room for improvements on manually annotated fine-grained samples.

Table 3. Performance comparison on gRefCOCO.

Model	*val*		*testA*		*testB*	
	N-acc.	T-acc.	N-acc.	T-acc.	N-acc.	T-acc.
MattNet	41.15	96.13	44.04	97.56	41.32	95.32
VLT	47.17	95.72	48.74	95.86	47.82	94.66
LAVT	49.32	96.18	49.25	95.08	48.46	95.34
ReLA	56.37	96.32	59.02	**97.68**	58.40	**95.44**
Ours($\alpha = 1$)	56.89	**96.50**	59.73	95.81	57.34	92.01
Ours ($\alpha = 0.5$)	**60.58**	95.23	**62.93**	94.80	**59.55**	90.40

Table 4. Performance comparison on Ref-ZOM.

Model	MCN	CMPC	VLT	LAVT	DMMI	Ours
test	75.81	77.01	79.26	83.11	87.02	**93.21**

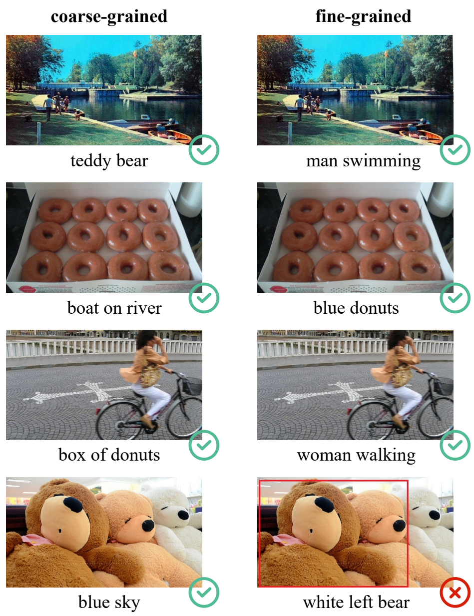

Figure 2. Qualitative examples on coarse-grained and fine-grained counterfactual samples.

2.2. Qualitative Analysis

We provide more coarse-grained and fine-grained qualitative examples, visualized in Figure 2. These images are from MS-COCO and have been trained by our model, while the text queries are newly annotated. We show our model's classification performance on coarse-grained and fine-grained counterfactual samples in two columns, respectively.

Specifically, the left samples are random image-text pairs without any semantic connections. The right samples are matched image-text pairs with the attribute words in texts changed into counterfactual words. Our model successfully identifies all the coarse-grained samples and most fine-grained samples. This result indicates that training a C-REC model on fine-grained samples also contributes to detecting coarse-grained samples, thus covering almost all counterfactual situations.

References

[1] Yutao Hu, Qixiong Wang, Wenqi Shao, Enze Xie, Zhenguo Li, Jungong Han, and Ping Luo. Beyond one-to-one: Rethinking the referring image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4067–4077, 2023. 1

[2] Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23592–23601, 2023. 1