Supplementary Materials for "FindIt: Generalized Localization with Natural Language Queries"

Weicheng Kuo, Fred Bertsch, Wei Li, AJ Piergiovanni, Mohammad Saffar, Anelia Angelova

Google Research, Brain Team
{weicheng,fredbertsch,mweili,ajpiergi,msaffar,anelia}@google.com

1 Mixture Ablations

Table 1 studies the mixing ratio of detection and localization with all RefCOCO splits together. Each row in Table 1 is a single unified model able to perform REC, LOC, and DET tasks on all RefCOCO splits. Unlike the other ablations, we compare the mixtures here on the full training schedule. We observe that increasing the detection and localization mixing ratio tends to improve the detection and localization tasks, but at a slight cost to the referring expressions comprehension tasks.

2 Failure Mode Analysis

In Figure 1, we visualize the failure cases of FindIt. The model tends to struggle with confounding objects of similar categories or attributes especially when given a complex query. The rare, long-tail and novel categories also pose challenges. These can be remedied by training on referring expression datasets with confounding objects [2], larger detection datasets [5], or pretraining on large language and vision datasets [1,4,3].

3 FindIt Visualizations

Figure 2 shows more visualizations of FindIt. The model is able to answer many complex expressions correctly and can even localize novel concepts such as "bottom part of a water plane". In addition, the model is able to localize many objects based on the category queries such as "Find the sports ball", and respond accurately to a detection query "Find all the objects".

Table 1. Ablations on FindIt mixtures with all RefCOCO splits. The mixture is given as the ratios between Det:Loc:Ref:Ref+:Ref-g:Ref-umd tasks.

Mixture	DET	LOC	RefCOCO			RefCOCO+			RefCOCOg		
			val	testA	testB	val	testA	testB	val-g	val-u	test-u
2:2:1:1:1:1	38.4	78.7	84.92	85.54	83.44	74.31	76.93	69.91	82.77	83.17	84.11
3:3:1:1:1:1	39.3	79.3	84.27	85.52	83.80	74.56	76.06	68.75	82.56	83.09	83.58
4:4:1:1:1:1	39.7	79.7	83.69	83.45	82.99	72.63	73.68	68.64	80.72	81.03	81.75

Fig. 1. Failure cases of FindIt on the referring expression comprehension and the text-based localization tasks (best viewed in color).

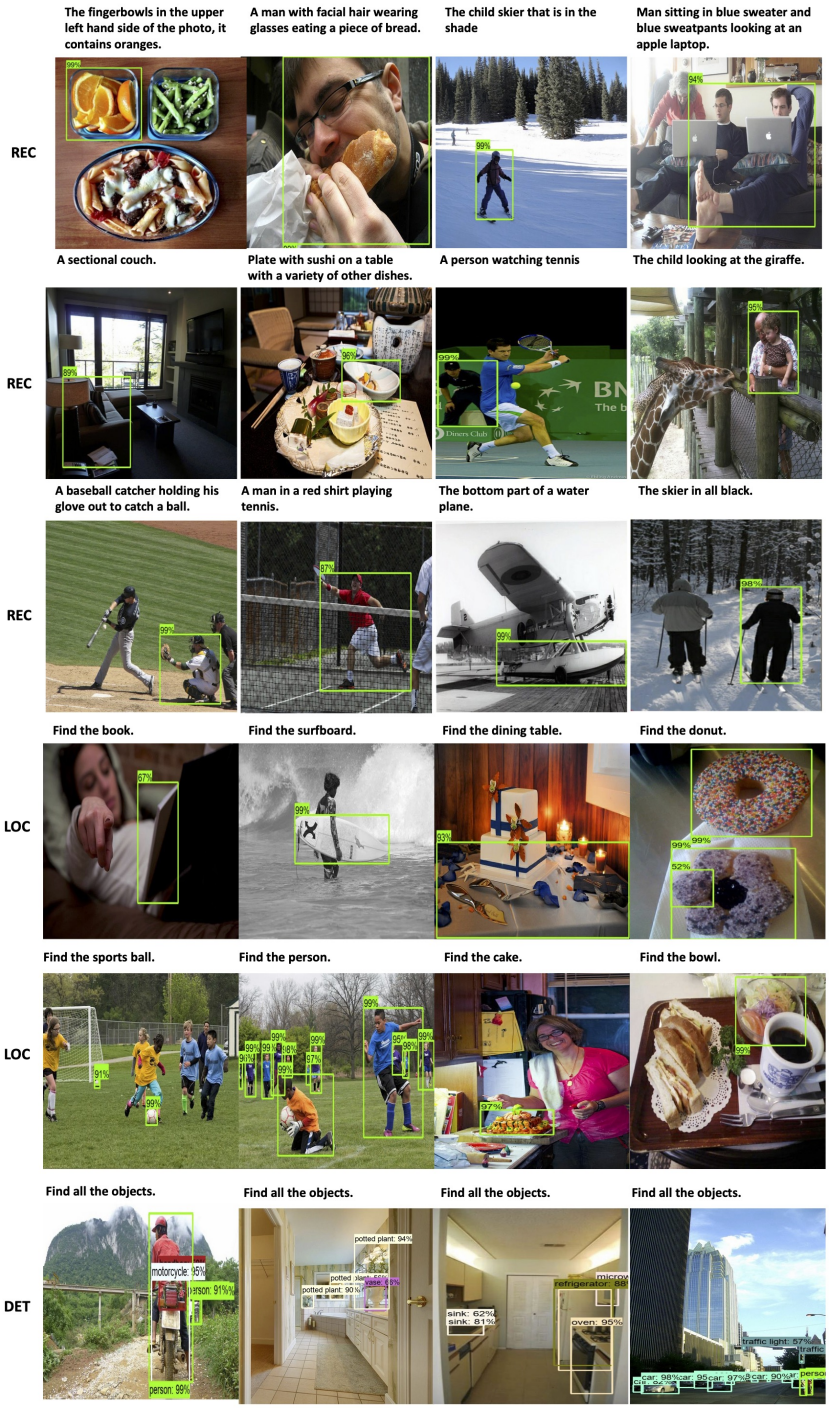

Fig. 2. FindIt visualization on REC, LOC, and DET tasks (best viewed in color).

References

1. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: CVPR (2021) 1
2. 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: CVPR (2020) 1
3. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q.V., Sung, Y., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: ICML (2021) 1
4. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML (2021) 1
5. Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Li, J., Zhang, X., Sun, J.: Objects365: A large-scale, high-quality dataset for object detection. In: ICCV (2019) 1