LEARNING OBJECT-LANGUAGE ALIGNMENTS FOR OPEN-VOCABULARY OBJECT DETECTION

Anonymous authors
Paper under double-blind review

A APPENDIX

1. Detailed Framework

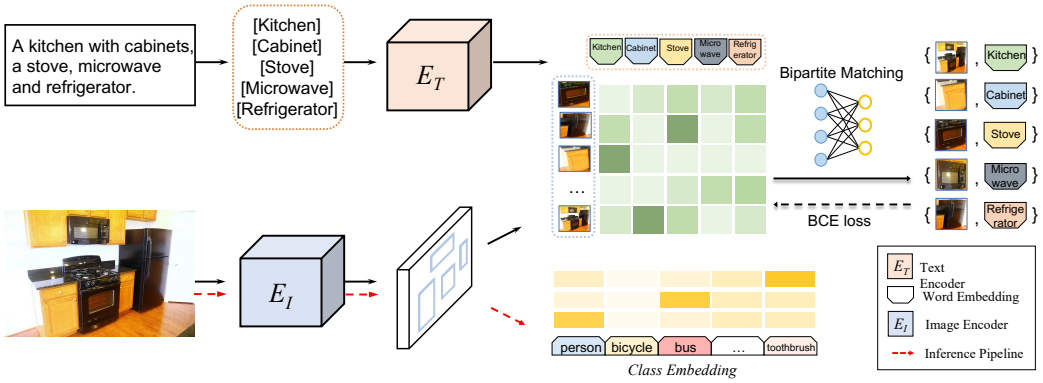

Figure 1: Overview of our proposed **VLDet** framework, which consists of two training parts. The bottom part (red arrows), also the inference path, is a classical two-stage Faster R-CNN pipeline trained with well annotated base-class data. The top part (black arrows) is the one to learn fine-grained object-word alignments with the corpus of image-text pairs, which is formulated as a bipartite matching problem, i.e., matching two sets of region and word candidates. For easy illustration, we did not draw the regression heads of Faster R-CNN here.

2. The Standard LVIS Benchmark

To investigate how helpful the image-text pairs data is to the fully supervised setting, we conducted experiments on the standard LVIS benchmark. In this setting, we use all the annotations including rare classes and base classes in the training set for Box-Supervised method. Detic (Zhou et al., 2022) and our method use additional image-text pairs in CC3M (Sharma et al., 2018). As the Table 1 shown, compared with the sate-of-the-art method Detic, our model further improves 0.8% mAP on rare classes and archives comparable results on the common and frequency classes.

Table 1: Results on standard LVIS benchmark. We evaluate Detic and our method using all classes annotations in the LVIS training set and additional images-text pairs in CC3M.

Method	mAP_r^{mask}	mAP_c^{mask}	mAP_f^{mask}	mAP_{all}^{mask}
Box-Supervised	25.6	30.4	35.2	31.5
Detic Zhou et al. (2022)	27.6	**31.0**	**35.4**	**32.2**
Our	**28.4**	**31.0**	35.1	32.1

3. Visualization of OV-COCO

As shown in Figure 2, we visualize some cases of matching results of open vocabulary COCO setting. As we can see, the model can extract promising region-words pairs from image-text data. The results of the matching include a variety of objects, such as “ostrich”, “duffle bag”, and “adult bear” which significantly expands the vocabulary for object detection. The results demonstrate generalization ability of our method to novel class.

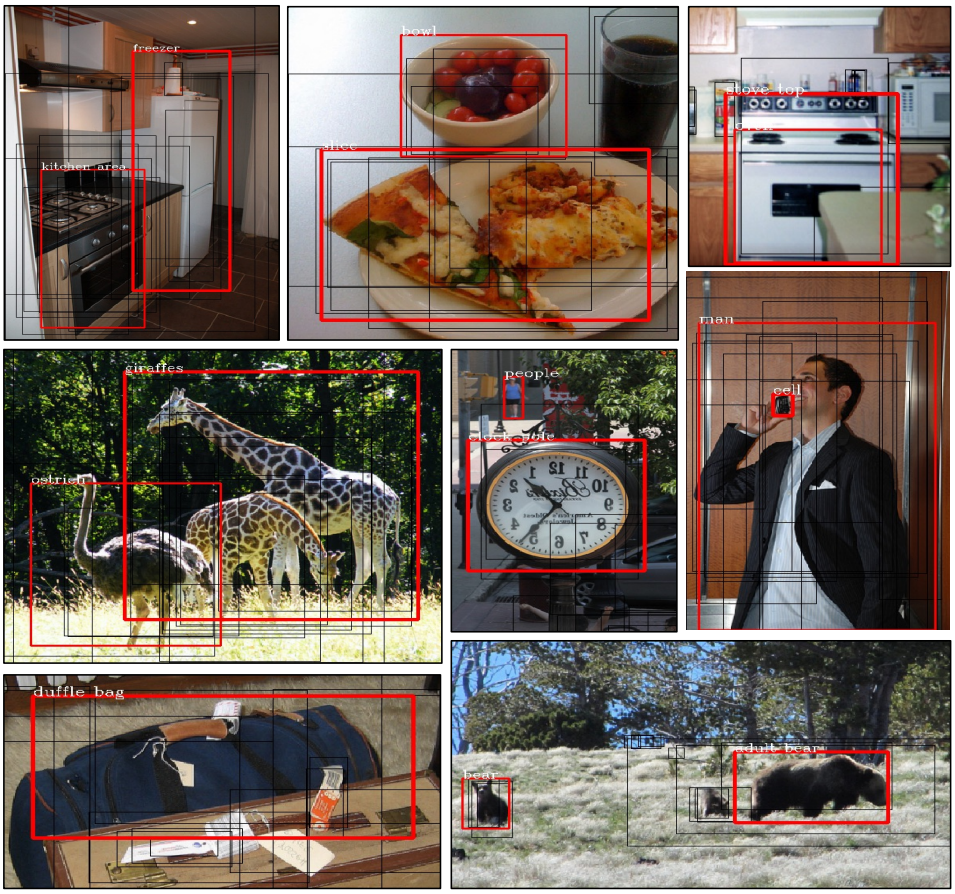

Figure 2: Visualization of the bipartite matching results in open-vocabulary COCO setting.

4. Failure Cases

We provide some failure cases of VLDet in Figure 3. These images is training images from CC3M in the open-vocabulary LVIS setting. We noticed that the region proposal network fails to provide the candidate region for "TV character", resulting in erroneous region-word pairs in sub-figure (1). This observation indicates that there is a scope of improve if a strong region proposal network could extensively providing the candidate regions. Besides, as sub-figure (2) shown, our method cannot handle the case where the target object "port" mentioned in caption but not appeared in the image. We leave these problems for the future study.

Figure 3: Failure cases of our method.

5. Vision and Language Representation Learning

Vision and language tasks, such as visual question answering and natural language for visual reasoning, require a unified understanding of both images and language. Pioneering works of multi-modal representation learning leverage the region-based visual features from object detection for better visual semantics. For example, Oscar (Li et al., 2020) introduces object tags and region features to learn the cross-domain semantics with a universal Transformer, significantly improving visual language understanding and generation tasks. ImageBERT (Radford et al., 2021) utilizes the masked object classification and region feature regression as pre-training tasks. UNITER (Chen et al., 2020) learns generalizable contextualized embeddings by word-region alignments to encourage similar output distributions across multiple modalities. Different from these works that aim to learn a universal semantic space for cross-modal understanding, we focus on increasing the generalization ability of object detection by the weak language supervision.

REFERENCES

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pp. 104–120. Springer, 2020.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pp. 121–137. Springer, 2020.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565, 2018.

Xingyi Zhou, Rohit Girdhar, Armand Joulin, Phillip Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. *arXiv preprint arXiv:2201.02605*, 2022.