Open Vocabulary Object Detection With an Open Corpus: Supplementary Materials

Jiong Wang[1] Huiming Zhang[2] Haiwen Hong[2] Xuan Jin[2]
Yuan He[2] Hui Xue[2] Zhou Zhao[1] *
[1] Zhejiang University, China [2] Alibaba Group
[1]{liubinggunzu, zhaozhou}@zju.edu.cn
[2]{zhm220845, honghaiwen.hhw, jinxuan.jx, heyuan.hy, hui.xueh}@alibaba-inc.com

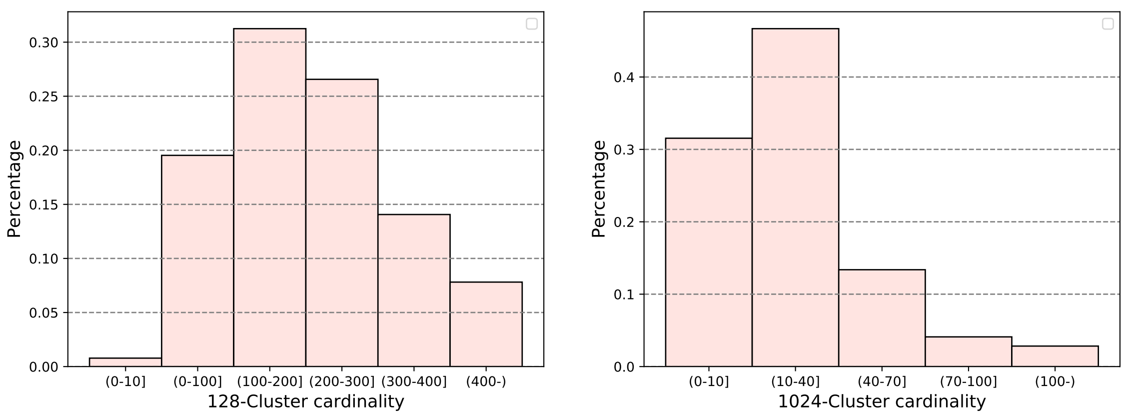

Figure 1: Visualization of the open corpus cluster cardinality distribution. The open corpus is separately clustered into 128 and 1,024 centroids.

GOAT cluster	**OV-COCO**		
	Novel AP	Base AP	All AP
COCO caption	35.82	53.03	48.53
CC3M	35.75	52.84	48.37
ImageNet-21k	35.91	53.12	48.62
All	**36.04**	**53.09**	**48.63**

Table 1: Ablation of generalized objectness assessment with different concept pools on the OV-COCO protocol.

NCE cluster	**OV-COCO**		
	Novel AP	Base AP	All AP
COCO caption	34.97	52.72	48.08
CC3M	35.19	52.89	48.26
ImageNet-21k	35.49	52.91	48.36
All	**35.83**	**53.98**	**48.49**

Table 2: Ablation of negative cluster expanding with different concept pools on the OV-COCO protocol.

1. Statistics of the Open Corpus

As briefly described in the main paper, the open object corpus is constructed with the object categories in the image recognition dataset (ImageNet-21k [1]) and caption datasets (COCO Caption [2] and Conceptual Captions [3]). The ImageNet-21k dataset contains 21,843 object concepts and the COCO caption and conceptual captions (CC3M) datasets contain 4,764 and 6,250 object concepts, sepa-

*corresponding author

Cluster 1: 'balloon' 'frisbee' 'sphere' 'egg' 'quahog' 'ovoid' 'basketball' 'soccer ball' 'soap bubble'

Cluster 2: 'jeep' 'ambulance' 'racing car' 'truck' 'wagon' 'police van' 'cab' 'suv' 'lorry' 'dumpster'

Cluster 4: 'pomelo' 'mango' 'carambola' 'watermelon' 'avocado' 'lemon' 'mandarin orange' 'grapefruit'

Cluster 5: 'salad' 'macaroni' 'soup' 'French dressing' 'burgoo' 'sakiyaki' 'confiture' 'curry' 'bearnaise'

Cluster 6: 'mixer' 'cutter' 'coffee mill' 'chain printer' 'vacuum_cleaner' 'crusher' 'autoloader' 'bulldozer'

Cluster 7: 'birch' 'banyan' 'cucumber tree' 'silk_tree' 'copper beech' 'sweet gum' 'olive_tree' 'swamp oak'

Cluster 8: 'bottle' 'vase' 'coffee_mug' 'barrel oilcan' 'samovar' 'perfume bottle' 'cylinder' 'tumbler'

Figure 2: Visualization of selected object concepts in the open corpus clusters.

Category	**Similar words**				
bench	park bench	garden bench	park bench	street bench	city bench
fork	pitchfork	tuning fork	carving fork	tablefork	plastic fork
bottle	water bottle	smelling bottle	liquor bottle	plastic bottle	glass bottle
suitcase	luggage bag	suit case	luggage case	luggage	luggage suitcase
umbrella	sun umbrella	umbrella hat	table umbrella	paper umbrella	umbrella tent

Table 3: Examples of similar words to the classifier in the open corpus.

rately. After gathering the concepts in three datasets and removing duplicate concepts, we get the open corpus used in this paper with 28,535 object concepts.

As illustrated in Figure 4 of the main paper, the best performance is achieved in the proposed generalized objectness assessment (GOAT) with 128 clusters and in the negative cluster expanding (NCE) with 1024 clusters. We give the distribution of cluster cardinality in Figure 1. It can be seen that most 128-clusters have cardinalities larger than 10, while most 1024-clusters have cardinalities smaller than 40. The GOAT objectness is obtained by equally summarizing the similarities of visual feature to the cluster centroids. The 128-cluster is a better choice because each cluster contains more concepts and is more likely to represent a type of objects. For NCE, 1024-clusters get better performance because more negative samples are expanded in the contrastive loss.

2. Ablation of Open Corpus

The proposed generalized objectness assessment (GOAT) is based on the open corpus clusters, and the comparison of using different datasets as the open corpus is illustrated in Table 1. It can be seen that GOAT with smaller concept pools (COCO caption and CC3M) get similar performance with larger (COCO caption and CC3M). Gathering them together gets slight performance enhancement.

The proposed negative cluster expanding (NCE) expands the negative samples in region-word alignment with the open corpus clusters. We give the comparison of using different datasets as the open corpus in Table 2. It can be seen that NCE with larger concept pools gets better performance and gathering all of them gets best performance.

3. Visualization of Open Corpus Clusters

The visualization of object concepts in several clusters is illustrated in Figure 2. It can be seen that each cluster contains object concepts with similar patterns. For example, clusters 1 and 4 separately contain round-shaped objects and fruits. Similar observation can be drawn in other clusters.

4. Visualization of the Open Corpus Classifier

The open corpus classifier is constructed by similar objects with the original classifier in the open corpus. We give some examples of similar words in Table 3, where we found the top-5 similar words are usually detailed descriptions of the original category. The open corpus classifier acts similarly with the text prompts, but gives more general descriptions to the original category.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[2] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1

[3] 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)*, pages 2556–2565, 2018. 1