Supplementary Materials for Region-Aware Pretraining for Open-Vocabulary Object Detection with Vision Transformers

Dahun Kim Anelia Angelova Weicheng Kuo
Google Research, Brain Team
{mcahny, anelia, weicheng}@google.com

Appendix

In the supplementary materials, we provide our detection visualizations along with our application on ego-centric data. We also provide more implementation details with used hyper-parameters and discuss the current limitations in the proposed RO-ViT in the hope to inspire more future research.

A. Implementation Details

Table 1 summarizes the hyper-parameters used in the image-text pretraining and open-vocabulary detection finetuning.

configuration	contrastive image-text pretraining	open-vocab. detection finetuning (LVIS)
optimizer	AdamW	SGD
momentum	β=0.9	β=0.9
weight decay	1e-2	1e-4
learning rate	5e-4	0.36
step decay factor	-	0.1$\times$
step decay schedule	-	[0.8, 0.9, 0.95]
backbone lr ratio	N/A	0.1 (ViT-B) / 0.5 (ViT-L)
warmup steps	1e4	1k
total steps	5e5	46.1k
batch size	4096 or 16384	256
image size	224	1024

Table 1. **RO-ViT hyper-parameters** for image-text pretraining and open-vocabulary detection finetuning.

B. Detection Visualization

We visualize our RO-ViT outputs on LVIS novel categories (Sec. 4.2) and transfer detection (Sec. 4.4) onto Objects365 in Fig. 1 and 2, respectively.

We use the ViT-B/16 backbone for visualization. The model was trained on the LVIS base categories following Sec. 4.2 of the paper. On LVIS, we only show the novel categories for clarify. RO-ViT is able to detect many novel objects (*e.g.*, *fishbowl, sombrero, shepherd dog, gargoyle, persimmon, chinaware, gourd, satchel*, and *washbasin*). We also visualize the transfer detection on Objects365 by replacing the vocabulary without finetuning. RO-ViT can detect a wide range of objects in complex scenes (*e.g.*, *power outlet, binocular, glasses, traffic sign*, and *shrimp*).

C. Application on Ego-centric Data

A main advantage of open-vocabulary detection is to deal with out-of-distribution data with categories given by the users on the fly. We test RO-ViT's transfer detection to a real-world ego-centric data, Ego4D [2]. We use the same RO-ViT trained on LVIS base categories with ViiT-B/16 backbone, as in Sec. B, *i.e.*, the model has been never trained on Ego4D.

The categories are provided by the user's visual inspection of the video, and are as follows.

- For the indoor scene: *plate, cabinet, stove, towel, cleaning rag, ventilator, knob, sauce and seasoning, steel lid, window, window blinds, plant, light switch, light, door, carpet, exit sign, doormat, hair, door lock, tree, poster on the wall, sticker on the wall, faucet, recycle bin, rack, hand, can, carton, trash, Christmas tree, plastic container, fridge*.
- For the grocery store scene: *exit sign, poster, chocolate bar, bag of candy, bag of cookies, snack, oreo, soy sauce, apple, pear, orange, grapes, price tag, cereal, instant noodle/ramen, cracker, ATM machine, instant noodle, wooden basket, red ramen bowls, magazine, drugs and medicine, Mayo, Ketchup, Cup noodle, burrito, Lays/Sun chips, seasoning sauce, black carton, salad dressing, canned food*.

Fig. 3 shows our RO-ViT prediction. Despite the large domain shift and heavy camera motions, RO-ViT is able to capture many objects in the ego-centric videos. Specifically, it is able to detect many novel categories never seen during

Figure 1. **LVIS novel category visualization (prediction).** We only show the novel categories for clarity. RO-ViT detects many novel categories (pointed by the red arrows) that it has never seen during detection training (*e.g.*, *fishbowl, sombrero, shepherd dog, gargoyle, persimmon, chinaware, gourd, satchel*, and *washbasin*).

training (*e.g.*, *light switch, exit sign, recycle bin, seasoning sauce, salad dressing, bag of cookies, canned food, and red ramen bowls*).

D. Limitations

RO-ViT leverages the knowledge in pretrained Vision Language Models (VLM). Therefore, the biases of trained VLMs can proparate into the downstream detector. In this paper, we use RO-ViT to demonstrate its capabilities and compare with existing works in open-vocabulary detection.

Figure 2. **Objects365 transfer detection visualization (prediction).** Our trained RO-ViT is able to perform on a new dataset without any finetuning, and captures many challenging categories including novel categories (pointed by red arrows, *e.g.*, *power outlet* and *shrimp*).

We recommend careful analysis of ethical risks before using it for other purposes.

E. Dataset License

- LVIS [3]: CC BY 4.0 + COCO license
- COCO Captions (retrieval) [1]: CC BY
- Flickr30k (retrieval) [4]: Custom (research-only, non-commercial)
- Objects365 [5]: Custom (research-only, non-commercial)
- Ego4D [2]: https://ego4d-data.org/pdfs/Ego4D-Licenses-Draft.pdf

(a) **Indoor scene.**

(b) Grocery store scene.

Figure 3. **Ego4D transfer detection visualization (prediction).** Ego4D [2] is a real-world and out-of-distribution data. Despite large domain shift and heavy camera movement, RO-ViT is able to detect novel, unseen objects (*e.g.*, *light switch, exit sign, recycle bin, seasoning sauce, salad dressing, bag of cookies, canned food, and red ramen bowls*).

References

[1] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. In *https://arxiv.org/abs/1504.00325*, 2015. 3

[2] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*, pages 18995–19012, 2022. 1, 3, 4

[3] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 3

[4] 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. In *ICCV*, pages 2641–2649, 2015. 3

[5] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *ICCV*, 2019. 3