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 fine-tuning.

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

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

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).

D. Limitations

RO-ViT leverages the knowledge in pretrained Vision Language Models (VLM). Therefore, the biases of trained VLMs can propagate 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 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


