Supplementry Material for: Region-Centric Image-Text Pretraining for Open-Vocabulary Detection

Dahun Kim Anelia Angelova Weicheng Kuo

Google DeepMind

1 Additional Implementation Details

1.1 Region-centric Pretraining

Loss function. As mentioned in Sec 3.2, our Region-centric Pretraining (REP) employs the multi-level image-text supervision and RPN-objectness training. The multi-level image-text supervision consists in the standard image-text contrastive loss (L_{con}) applied at each i-th feature pyramid level. During training, we employ the multi-level visual-text similarity as a supervisory signal for training the RPN's objectness map (see Fig. 3). The target RPN score is computed as the cosine similarity between each RoI embeddings and the text, where any negative dot product value is mapped to zero to keep the target score in range [0, 1]. We use L1 regression loss between the target scores and the corresponding RoIs' center locations on the objectness map. In sum, the total loss objectives of our region-centric pretraining is $L_{REP} = \sum_{i=2}^{5} L_{con}^{i} + \lambda L_{reg}$, where $\lambda = 1$.

Hyperparameters. Table 1 summarizes the hyperparameters used in our Region-centric Pretraining.

	baseline CLIP (Sec 3.1)	Region-centric Pretraining (Sec 3.2)
optimizer	AdamW	AdamW
momentum	β=0.9	β=0.9
weight decay	0.01	0.01
learning rate	0.001	0.0001
warmup steps	5k	5k
total steps	500k	30k
batch size	16384	4096
image size	224	256

Table 1: Hyperparameters for Region-centric Pretraining.

1.2 Open-vocabulary detection finetuning

Loss function. We follow the same objective functions of Mask R-CNN (for LVIS) and Faster R-CNN (for COCO), except that we have an additional frozen

backbone distillation loss (Sec. 3.4). We use a cosine distance loss that aligns the RoI-Align embeddings extracted from the feature maps of finetuned vs frozen backbones. The cosine distance is computed for each RoI then averaged over all RoIs. In sum, our detection loss objectives is $L_{Det} = L_{Rpn\text{-}obj} + L_{Rpn\text{-}box} + L_{Frcnn\text{-}class} + L_{Frcnn\text{-}box} + L_{mask} + \gamma L_{distill}$, where $\gamma = 1$.

Hyperparameters. Table 2 summarizes the hyperparameters used in our open-vocabulary detection finetuning. We use the same open-vocabulary detector design of RO-ViT [4] which adopts the ViTDet architecture [7] and the centerness-based RPN [5] that uses a single anchor per location.

OVD finetuning	ViT-L (LVIS / COCO)	ViT-B and S (LVIS / COCO)
optimizer	SGD	SGD
momentum	β=0.9	β=0.9
weight decay	0.0001	0.0001
learning rate	0.18 / 0.02	0.36 / 0.02
backbone lr ratio	0.6× / 0.2×	0.1× / 0.1×
step decay factor	0.1×	0.1×
step decay schedule	[0.8, 0.9, 0.95]	[0.8, 0.9, 0.95]
warmup steps	1k	1k
total steps	36.8k / 11.3k	46.1k / 11.3k
batch size	128	256
image size	1024	1024

Table 2: Hyperparameters for open-vocabulary detection finetuning.

2 Additional Ablations

Region-centric Pretraining (REP). Table 3a provides more ablations on the RoI sampling and pooling methods in the FPN within the REP training. We investigate whether pooling over random boxes are more effective than global pooling over pixels or blockwise pooling on a regular grid. Table 3a shows that both global avg- and max-pooling are sub-optimal due to the lack of saliency map (avg-pool), and the limited capacity of a single pixel to represent semantic concepts for contrastive learning (max-pool), respectively. Our approach combines the best of both worlds, by first avg-pooling within each RoI, and then max-pooling over these RoI embeddings. Each embedding represents a proper RoI and the global representation captures the saliency map through max-pooling. Despite the absence of explicit region-level supervision in our REP training, the max pooling over random regions encourages the more salient, text-aligned RoI features to contribute more to the whole image representation in the contrastive loss. To study the need of randomness in our method, we divide the feature map into a $N \times N$ grid and treat each grid cell as an RoI (block-wise RoI). The random RoI is superior due to the greater variety in RoI scales and locations.

Table 3b ablates contrastive batch size in our REP pretraining, where we choose batch size 4k, as larger batch does not result in improvements.

RoI sampling	pool	AP_r	AP
global (pixel-wise RoI)	avg	34.0	33.9
global (pixel-wise RoI)	max	33.7	33.5
block-wise (8×8 grid RoI)	max	33.9	33.8
block-wise (4×4 grid RoI)	max	34.2	34.3
block-wise (2×2 grid RoI)	max	33.1	33.8
random RoI	max	34.8	34.9

(a) **RoI sampling and pooling:** The pooling is applied per pyramid level for all methods. Using multiple random RoIs followed by max pooling performs the best, outperforming the whole-image RoI and pixel-wise RoIs methods.

batch	AP_r	AP
1k	33.9	34.1
2k	34.3	34.6
4k	34.8	34.9
16k	34.7	34.8

(b) **Contrastive batch size:** We choose batch size 4k, as larger batch doe not show improvement.

Table 3: More ablations for Region-centric Pretraining (Sec. 3.1). (a) RoI sampling and pooling in the FPN. (b) Contrastive batch size for our region-centric pretraining. Best setting is in gray.

backbone	mask AP_r	mask AP_c	mask AP_f
baseline win. attn.	32.2	33.6	33.1
SWL (ours)	35.0 (+2.8)	35.5 (+1.9)	34.9 (+1.8)

(a) *LVIS OVD benchmark.*

backbone	AP (pretrained)	AP (random init.)
baseline win. attn.	48.0	39.5
SWL (ours)	49.0 (+1.0)	40.3 (+0.8)

(b) ***Fully-supervised*** *detection on COCO (ViT-B).*

Table 4: Generality of Shifted Window Learning (SWL)

Shifted-Window Learning for Detection (SWL). The proposed SWL is beneficial for both OVD and fully supervised detection. In Table 4a, we show that the gain from SWL is 50% larger for *rare* classes (+2.8 AP) than frequent and common classes (+1.9 AP) in the LVIS OVD benchmark. For standard detection, Tab. 4b shows SWL improves performance using both pretrained or randomly initialized backbone. The results show that SWL can improve detection in general, as well.

3 Limitations

Our models utilize the rich image-text information acquired through pretraining, which may reinforce deficiencies and biases in the raw web data and expose potentially harmful biases or stereotypes. The models we trained are designed for academic research purposes and need more rigorous fairness studies or data cleaning before serving product applications.

4 Future direction

As our Region-centric Pretraining only utilizes image-text paired data, without any box annotations, it primarily focuses on the region-recognition pathway of a detector, encompassing components such as the backbone, FPN, RoI-Align, RPN-objectness, and Faster RCNN-classifier. Consequently, the learning of box regression layers occurs during the detection finetuning stage, similarly to other open-vocabulary detection works [2, 4, 6, 11]. Exploring the learning of box regression solely through image-text paired data during CLIP pretraining would be an interesting direction for future research.

5 Dataset license

- LAION [9]: MIT License
- COCO [8]: Creative Commons Attribution 4.0 License
- LVIS [3]: CC BY 4.0 + COCO license
- Objects365 [10]: Custom (research-only, non-commercial)
- Ego4D [1]: https://ego4d-data.org/pdfs/Ego4D-Licenses-Draft.pdf

References

1. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al.: Ego4d: Around the world in 3,000 hours of egocentric video. In: CVPR. pp. 18995–19012 (2022) 4
2. Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Open-vocabulary object detection via vision and language knowledge distillation. In: ICLR (2022) 4
3. Gupta, A., Dollar, P., Girshick, R.: Lvis: A dataset for large vocabulary instance segmentation. In: CVPR (2019) 4
4. Kim, D., Angelova, A., Kuo, W.: Region-aware pretraining for open-vocabulary object detection with vision transformers. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2023) 2, 4
5. Kim, D., Lin, T.Y., Angelova, A., Kweon, I.S., Kuo, W.: Learning open-world object proposals without learning to classify. IEEE Robotics and Automation Letters **7**(2), 5453–5460 (2022) 2
6. Kuo, W., Cui, Y., Gu, X., Piergiovanni, A., Angelova, A.: F-vlm: Open-vocabulary object detection upon frozen vision and language models. ICLR (2023) 4
7. Li, Y., Mao, H., Girshick, R., He, K.: Exploring plain vision transformer backbones for object detection. In: ECCV (2022) 2
8. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014) 4
9. Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A.: Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv preprint arXiv:2111.02114 (2021) 4
10. Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., Sun, J.: Objects365: A large-scale, high-quality dataset for object detection. In: ICCV (2019) 4
11. Zareian, A., Rosa, K.D., Hu, D.H., Chang, S.F.: Open-vocabulary object detection using captions. In: CVPR (2021) 4