Exploiting Unlabeled Data with Vision and Language Models for Object Detection

Shiyu Zhao, Zhixing Zhang, Samuel Schulter, Long Zhao, Vijay Kumar B.G, Anastasis Stathopoulos, Manmohan Chandraker, Dimitris Metaxas

ECCV 2022

Code: https://github.com/xiaofeng94/VL-PLM
Motivation

*Traditional way to train a detector*

Costly human annotations on limited categories

柴油 training data

Test data

Unlabeled data

 Detector

Detection results

้ำผล

Categories are limited due to costly human annotations  ?? Easy-to-access unlabeled images are not leveraged
Two solutions to leverage the unlabeled

Semi-supervised object detection (SSOD)

- Task categories: N categories
- Data: Fully labeled, Unlabeled
- Ground truth: Fully labeled
- Pseudo labels: Unlabeled

Categories are fixed and limited
Unlabeled images are used

Zero-shot/Open vocabulary object detection (OVD)

- Task categories: Base, Novel
- Data: Partially labeled
- Ground truth: Base, Novel

Categories can be any or infinite
Unlabeled images are not leveraged
Motivation

Two solutions to leverage the unlabeled

Semi-supervised object detection (SSOD)

Task categories

Ground truth

N categories

Fully labeled

Unlabeled

Pseudo labels
Two solutions to leverage the unlabeled

Semi-supervised object detection (SSOD)
- Task categories: N categories
- Data: Fully labeled, Unlabeled

Zero-shot/Open vocabulary object detection (OVD)
- Task categories: Base, Novel
- Data: Partially labeled

😊 Categories are fixed and limited
😊 Unlabeled images are used

😢 Categories can be any or infinite
😢 Unlabeled images are not leveraged
Our solution

Our V&L-guided Pseudo-Label Mining (VL-PLM)

Motivation

😄 Unlabeled images are used

😊 Categories can be any or infinite
Overview of our pseudo label generation

Two steps:

• Generate region proposals using a pretrained two-stage class-agnostic proposal generator

• Classify region proposals into categories of interests with pretrained V&L model (e.g. CLIP)
Two-stage region proposal network (RPN) matters: A single-stage RPN introduces many noisy boxes. But those noisy boxes may converge to one location by repeating the box regression of RoI head (RoI refinement)
Classifying region proposals

Improving CLIP’s localization ability: RPN scores indicate localization quality of boxes for various objects. Thus, average CLIP and RPN scores as the final prediction score (RPN fusion)
Experiments

Quantitative results

VL-PLM achieves state-of-the-art performance on OVD and benefits SSOD

| OVD on LVIS          | Supervised | Supervised+PLs | Supervised+VL-PLM | ViLD [16] | Base + Novel | Base |
|----------------------|------------|----------------|-------------------|-----------|--------------|------|
| Method               | Training data | AP_r | AP_c | AP_f | mAP     | AP_r | AP_c | AP_f | mAP   | AP_r | AP_c | AP_f | mAP   | AP_r | AP_c | AP_f | mAP   | AP_r | AP_c | AP_f | mAP   | AP_r | AP_c | AP_f | mAP   |
| Supervised           | Base + Novel | 12.3 | 24.3 | 32.4 | 25.4   | 16.6 | 21.1 | 31.6 | 24.4   | 17.2 | 23.7 | 35.1 | 27.0   | 17.2 | 23.7 | 35.1 | 27.0   | 17.2 | 23.7 | 35.1 | 27.0   | 17.2 | 23.7 | 35.1 | 27.0   |
| ViLD [16]            | Base        |       |      |      |        | 16.6 | 21.1 | 31.6 | 24.4   |       |      |      |        |       |      |      |      |        |       |      |      |      |        |
| VL-PLM (Ours)        | Base        | 17.2 | 23.7 | 35.1 | 27.0   |      |      |      |        |       |      |      |      |        |       |      |      |      |        |       |      |      |      |        |

| SSOD on COCO          | Supervised | Supervised+PLs | Supervised+VL-PLM | STAC [46] | STAC+VL-PLM | STAC | STAC+VL-PLM | STAC | STAC+VL-PLM | STAC | STAC+VL-PLM | STAC | STAC+VL-PLM | STAC | STAC+VL-PLM | STAC | STAC+VL-PLM | STAC | STAC+VL-PLM | STAC | STAC+VL-PLM |
|-----------------------|------------|----------------|-------------------|-----------|-------------|------|-------------|------|-------------|------|-------------|------|-------------|------|-------------|------|-------------|------|-------------|------|-------------|
| Method                | Training data | 1% COCO | 2% COCO | 5% COCO | 10% COCO | Novel | Base | Overall | Novel | Base | Overall | Novel | Base | Overall | Novel | Base | Overall | Novel | Base | Overall |
| Supervised            | Base + Novel | 9.25 | 12.70 | 17.71 | 22.10 | 0.31 | 29.2 | 24.9   | 0.31 | 29.2 | 24.9   | 0.31 | 29.2 | 24.9   | 0.31 | 29.2 | 24.9   | 0.31 | 29.2 | 24.9   |
| Supervised+PLs        | Base        | 11.18 | 14.88 | 21.20 | 25.98 | 3.41 | 13.8 | 13.0   | 3.41 | 13.8 | 13.0   | 3.41 | 13.8 | 13.0   | 3.41 | 13.8 | 13.0   | 3.41 | 13.8 | 13.0   |
| Supervised+VL-PLM     | Base        | 15.35 | 18.60 | 23.70 | 27.23 | 4.12 | 35.9 | 27.9   | 4.12 | 35.9 | 27.9   | 4.12 | 35.9 | 27.9   | 4.12 | 35.9 | 27.9   | 4.12 | 35.9 | 27.9   |
| STAC [46]             |             | 13.97 | 18.25 | 24.38 | 28.64 |       |      |        |       |      |        |       |      |        |       |      |        |       |      |        |       |      |        |       |      |        |
| STAC+VL-PLM           |             | 17.71 | 21.20 | 26.21 | 29.61 |       |      |        |       |      |        |       |      |        |       |      |        |       |      |        |       |      |        |       |      |        |

Zero-shot/OVD on COCO

| Method               | Training Source | Novel AP | Base AP | Overall AP |
|----------------------|-----------------|----------|---------|------------|
| Bansal et al. [4]    | instance-level labels in $S_B$ | 0.31    | 29.2    | 24.9       |
| Zhu et al. [63]      | instance-level labels in $S_B$ | 3.41    | 13.8    | 13.0       |
| Rahman et al. [40]   | instance-level labels in $S_B$ | 4.12    | 35.9    | 27.9       |
| OVR-CNN [56]         | image-caption pairs in $S_B \cup S_N$ | 22.8    | 46.0    | 39.9       |
| Gao et al. [14]      | image-caption pairs in $S_B \cup S_N$ | 30.8    | 46.1    | 42.1       |
| RegionCLIP [59]      | instance-level labels in $S_B$ | 31.4    | 57.1    | 50.4       |
| RegionCLIP* [59]     | raw image-text pairs via Internet | 14.2    | 52.8    | 42.7       |
| ViLD [16]            | instance-level labels in $S_B$ | 27.6    | 59.5    | 51.3       |
| VL-PLM (Ours)        | raw image-text pairs via Internet | 34.4    | 60.2    | 53.5       |

[ViLD] Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Open-vocabulary Object Detection via Vision and Language Knowledge Distillation. In ICLR 2022

[STAC] Sohn, K., Zhang, Z., Li, C.L., Zhang, H., Lee, C.Y., Pfister, T.: A simple semisupervised learning framework for object detection. arXiv 2020

[RegionCLIP] Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L.H., Zhou, L., Dai, X., Yuan, L., Li, Y., Gao, J.: RegionCLIP: Region-based language-image pretraining. In CVPR 2022
Experiments

Understanding pseudo label quality

Visualization of good pseudo labels
Experiments

Understanding pseudo label quality

Visualization of bad pseudo labels
Exploiting Unlabeled Data with Vision and Language Models for Object Detection

Shiyu Zhao, Zhixing Zhang, Samuel Schulter, Long Zhao, Vijay Kumar B.G, Anastasis Stathopoulos, Manmohan Chandraker, Dimitris Metaxas

ECCV 2022

Website: https://www.nec-labs.com/~mas/VL-PLM/
Code: https://github.com/xiaofeng94/VL-PLM
Paper link: https://arxiv.org/abs/2207.08954