Scaling Open-Vocabulary Object Detection

Authors: Matthias Minderer, Alexey Gritsenko, and Neil Houlsby
NeurIPS 2023, 23 citations
Presented by Group 7: Daniel Cisneros, Suranadi Dodampagamage, Andrew El-Kommos, Bradley Racey, Salem Long
Outline

1. Background
2. Method
3. Results
4. Limitations
5. Conclusion
Background
Open Vocabulary Detection

Cake with cherries on top
OWL-ViT Architecture

Image-level contrastive pre-training
OWL-ViT Architecture

Transfer to open-vocabulary detection

- Text Transformer encoder
- Vision Transformer encoder
- MLP head

Set prediction loss over objects in an image.
Self-Training, Label Spaces, and N-Grams

• How can the scarcity of detection data can be addressed?
• Self-training, uses an existing detector to predict bounding boxes on unlabeled images to generate data for training better detectors.
• Label spaces, refers to the set of all possible labels that can be assigned to instances in a dataset.
Self-Training, Label Spaces, and N-Grams

• Expanded label spaces using N-Grams

  Unigram: Apple

  Bigram: Red apple

  Trigram: Juicy red apple
Method
Method

1. Annotation
   - Caption: "Monarch on a Zinnia"
   - N-grams:
     - Monarch
     - Monarch on
     - Monarch on a
     - Monarch on a Zinnia
   - Open-vocab detector (OWL-ViT L14)

2. Self-training
   - Pseudo-annotated WebLI
   - OWLv2
   - OWL-ViT detection loss

3. Fine-tuning (optional)
   - LVIS_{base}
   - OWLv2
   - OWL-ViT detection loss
1. Annotation

Caption: "Monarch on a Zinnia"

N-grams:
- Monarch
- Monarch on
- Monarch on a
- Monarch on a Zinnia

Open-vocab detector (OWL-ViT L/14)

2. Self-training

Pseudo-annotated WebLI

OWLv2

OWL-ViT detection loss

3. Fine-tuning (optional)

LVIS

base

OWLv2

OWL-ViT detection loss
Annotations

Dataset:
- WebLI (Image-Text Dataset)
- ~10 B images

Label Spaces:
- Human-curated label space
- Machine-generated label space
Self-training

1. Annotation

Caption: "Monarch on a Zinnia"

N-grams:
- Monarch
- Monarch on
- Monarch on a Zinnia

Open-vocab detector (OWL-ViT L/14)

2. Self-training

Pseudo-annotated WebLI

OWLv2

OWL-ViT detection loss

3. Fine-tuning (optional)

LVIS base

OWLv2

OWL-ViT detection loss
Self-training

• Self-train a new detector on the pseudo-annotations

• Enhance training efficiency with:
  • Token dropping
  • Instance selection
  • Mosaics
Token dropping

• Images contains low pixel variance areas
• Less informative
• Drop 50% tokens lower than mean pixel variance
Token dropping

- Dropping up to 50% of tokens does not significantly affect performance
- It remains within one standard deviation of the full performance

<table>
<thead>
<tr>
<th>Metric</th>
<th>Token drop rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.00</td>
</tr>
<tr>
<td>LVIS AP\textsubscript{val} \textsubscript{all}</td>
<td>33.3 ±0.33</td>
</tr>
<tr>
<td>LVIS AP\textsubscript{val} \textsubscript{rare}</td>
<td>31.8 ±1.16</td>
</tr>
</tbody>
</table>
Instance selection

- OWL-ViT predicts one bounding box per encoder token
- Most output tokens do not represent objects
- Objectness head: predicts likelihood that an output token represents an object
- Select 10% of instances by top objectness during training
Mosaics

4 x 4 mosaic before and after 50% of patches dropped
Fine-tuning

1. Annotation
   - Caption: "Monarch on a Zinnia"
   - N-grams: Monarch, Monarch on a Zinnia...
   - Open-vocab detector (OWL-ViT L/14)

2. Self-training
   - Pseudo-annotated WebLI
   - OWLv2
   - OWL-ViT detection loss

3. Fine-tuning (optional)
   - LVIS$_{base}$
   - OWLv2
   - OWL-ViT detection loss
Fine-Tuning

• Further improves detection performance.
• Trade-off between open vocabulary performance and fine-tuned classes.
• Create an ensemble of the model by averaging model weights.
Results
### Experiments – results

<table>
<thead>
<tr>
<th>Method</th>
<th>Backbone</th>
<th>Self-training data</th>
<th>Self-training vocabulary</th>
<th>Human box annotations</th>
<th>ODinW 13</th>
<th>LVIS $AP_{val}$</th>
<th>LVIS $AP_{min}^{val}$</th>
<th>LVIS $AP_{min}^{val}$</th>
<th>LVIS $AP_{val}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. RegionCLIP [40]</td>
<td>R50x4</td>
<td>CC3M</td>
<td>6k concepts</td>
<td>LVIS$_{base}$</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>32.3</td>
<td>22.0</td>
</tr>
<tr>
<td>2. OWL [21]</td>
<td>CLIP B/16</td>
<td>–</td>
<td>–</td>
<td>O365+VG</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>27.2</td>
<td>20.6</td>
</tr>
<tr>
<td>3. OWL [21]</td>
<td>CLIP L/14</td>
<td>–</td>
<td>–</td>
<td>O365+VG</td>
<td>48.4</td>
<td>–</td>
<td>–</td>
<td>34.6</td>
<td>31.2</td>
</tr>
<tr>
<td>4. GLIPv2 [39]</td>
<td>Swin-T</td>
<td>Cap4M</td>
<td>tokens</td>
<td>O365+GoldG</td>
<td>48.5</td>
<td>29.0</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>5. GLIPv2 [39]</td>
<td>Swin-B</td>
<td>CC15M</td>
<td>tokens</td>
<td>FiveODs+GoldG</td>
<td>54.2</td>
<td>48.5</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>6. GLIPv2 [39]</td>
<td>Swin-H</td>
<td>CC15M</td>
<td>tokens</td>
<td>FiveODs+GoldG</td>
<td>55.5</td>
<td>50.1</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>7. F-VLM [14]</td>
<td>R50x4</td>
<td>–</td>
<td>–</td>
<td>LVIS$_{base}$</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>28.5</td>
<td>26.3</td>
</tr>
<tr>
<td>8. F-VLM [14]</td>
<td>R50x64</td>
<td>–</td>
<td>–</td>
<td>LVIS$_{base}$</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>34.9</td>
<td>32.8</td>
</tr>
<tr>
<td>9. 3Ways [1]</td>
<td>NFNet-F0</td>
<td>TODO</td>
<td>captions</td>
<td>LVIS$_{base}$</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>35.7</td>
<td>25.6</td>
</tr>
<tr>
<td>10. 3Ways [1]</td>
<td>NFNet-F6</td>
<td>TODO</td>
<td>captions</td>
<td>LVIS$_{base}$</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>44.6</td>
<td>30.1</td>
</tr>
<tr>
<td>11. OWL-ST</td>
<td>CLIP B/16</td>
<td>WebLI</td>
<td>N-grams</td>
<td>O+VG</td>
<td>48.8</td>
<td>31.8</td>
<td>35.4</td>
<td>27.0</td>
<td>29.6</td>
</tr>
<tr>
<td>12. OWL-ST</td>
<td>CLIP L/14</td>
<td>WebLI</td>
<td>N-grams</td>
<td>O+VG</td>
<td>53.0</td>
<td>38.1</td>
<td>39.0</td>
<td>33.5</td>
<td>34.9</td>
</tr>
<tr>
<td>13. OWL-ST</td>
<td>SigLIP G/14</td>
<td>WebLI</td>
<td>N-grams</td>
<td>O+VG</td>
<td>49.9</td>
<td>37.8</td>
<td>40.9</td>
<td>33.7</td>
<td>37.5</td>
</tr>
</tbody>
</table>

9 point improvement for OWL-ST compared with previous OWL-ViT even without fine-tuning!
Experiments – results

<table>
<thead>
<tr>
<th>Method</th>
<th>Backbone</th>
<th>Self-training data</th>
<th>Self-training vocabulary</th>
<th>Human box annotations</th>
<th>ODinW AP 13</th>
<th>LVIS AP mini all</th>
<th>LVIS AP mini rare</th>
<th>LVIS AP val all</th>
<th>LVIS AP val rare</th>
</tr>
</thead>
<tbody>
<tr>
<td>11 OWL-ST</td>
<td>CLIP B/16</td>
<td>WebLI</td>
<td>N-grams</td>
<td>O+VG</td>
<td>48.8</td>
<td>31.8</td>
<td>35.4</td>
<td>27.0</td>
<td>29.6</td>
</tr>
<tr>
<td>12 OWL-ST</td>
<td>CLIP L/14</td>
<td>WebLI</td>
<td>N-grams</td>
<td>O+VG</td>
<td>53.0</td>
<td>38.1</td>
<td>39.0</td>
<td>33.5</td>
<td>34.9</td>
</tr>
<tr>
<td>13 OWL-ST</td>
<td>SigLIP G/14</td>
<td>WebLI</td>
<td>N-grams</td>
<td>O+VG</td>
<td>49.9</td>
<td>37.8</td>
<td>40.9</td>
<td>33.7</td>
<td>37.5</td>
</tr>
<tr>
<td>14 OWL-ST+FT</td>
<td>CLIP B/16</td>
<td>WebLI</td>
<td>N-grams</td>
<td>O+VG, LVISbase</td>
<td>48.6</td>
<td>47.2</td>
<td>37.8</td>
<td>41.8</td>
<td>36.2</td>
</tr>
<tr>
<td>15 OWL-ST+FT</td>
<td>CLIP L/14</td>
<td>WebLI</td>
<td>N-grams</td>
<td>O+VG, LVISbase</td>
<td>50.1</td>
<td>54.1</td>
<td>46.1</td>
<td>49.4</td>
<td>44.6</td>
</tr>
<tr>
<td>16 OWL-ST+FT</td>
<td>SigLIP G/14</td>
<td>WebLI</td>
<td>N-grams</td>
<td>O+VG, LVISbase</td>
<td>50.1</td>
<td>51.3</td>
<td>50.9</td>
<td>47.0</td>
<td>47.2</td>
</tr>
</tbody>
</table>

- Why does the model perform better on LVIS\textsubscript{rare} than on LVIS\textsubscript{all}?
### Experiments – results

<table>
<thead>
<tr>
<th>Method</th>
<th>Backbone</th>
<th>Self-training data</th>
<th>Self-training vocabulary</th>
<th>Human box annotations</th>
<th>ODinW 13</th>
<th>LVIS AP mini all</th>
<th>LVIS AP mini rare</th>
<th>LVIS AP val all</th>
<th>LVIS AP val rare</th>
</tr>
</thead>
<tbody>
<tr>
<td>7 F-VLM [14]</td>
<td>R50x4</td>
<td>–</td>
<td>–</td>
<td>LVIS base</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>28.5</td>
<td>26.3</td>
</tr>
<tr>
<td>8 F-VLM [14]</td>
<td>R50x64</td>
<td>–</td>
<td>–</td>
<td>LVIS base</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>34.9</td>
<td>32.8</td>
</tr>
<tr>
<td>9 3Ways [1]</td>
<td>NFNNet-F0</td>
<td>TODO</td>
<td>captions</td>
<td>LVIS base</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>35.7</td>
<td>25.6</td>
</tr>
<tr>
<td>10 3Ways [1]</td>
<td>NFNNet-F6</td>
<td>TODO</td>
<td>captions</td>
<td>LVIS base</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>44.6</td>
<td>30.1</td>
</tr>
<tr>
<td>11 OWL-ST</td>
<td>CLIP B/16</td>
<td>WebLI</td>
<td>N-grams</td>
<td>O+VG</td>
<td>48.8</td>
<td>31.8</td>
<td>35.4</td>
<td>27.0</td>
<td>29.6 -3.2</td>
</tr>
<tr>
<td>12 OWL-ST</td>
<td>CLIP L/14</td>
<td>WebLI</td>
<td>N-grams</td>
<td>O+VG</td>
<td>53.0</td>
<td>38.1</td>
<td>39.0</td>
<td>33.5</td>
<td>34.9 +2.1</td>
</tr>
<tr>
<td>13 OWL-ST</td>
<td>SigLIP G/14</td>
<td>WebLI</td>
<td>N-grams</td>
<td>O+VG</td>
<td>49.9</td>
<td>37.8</td>
<td>40.9</td>
<td>33.7</td>
<td>37.5 +4.7</td>
</tr>
<tr>
<td>14 OWL-ST+FT</td>
<td>CLIP B/16</td>
<td>WebLI</td>
<td>N-grams</td>
<td>O+VG, LVIS base</td>
<td>48.6</td>
<td>47.2</td>
<td>37.8</td>
<td>41.8</td>
<td>36.2 +3.4</td>
</tr>
<tr>
<td>15 OWL-ST+FT</td>
<td>CLIP L/14</td>
<td>WebLI</td>
<td>N-grams</td>
<td>O+VG, LVIS base</td>
<td>50.1</td>
<td>54.1</td>
<td>46.1</td>
<td>49.4</td>
<td>44.6 +11.8</td>
</tr>
<tr>
<td>16 OWL-ST+FT</td>
<td>SigLIP G/14</td>
<td>WebLI</td>
<td>N-grams</td>
<td>O+VG, LVIS base</td>
<td>50.1</td>
<td>51.3</td>
<td>50.9</td>
<td>47.0</td>
<td>47.2 +14.4</td>
</tr>
</tbody>
</table>

With fine-tuning, OWL-ST is 14.4 points higher than the next best model from literature!
Experiments – pseudo-annotation label space

![Graphs showing LVIS AP metrics for fine-tuned, unseen classes, and "In the Wild" datasets with different label spaces and methods.]

- Fine-tuned classes
- Unseen classes
- "In the Wild" datasets

Legend:
- Curated vocabulary
- N-grams
- N-grams+curated
Experiments – pseudo-annotations filtering

- Fine-tuned classes
- Unseen classes
- "In the Wild" datasets

Confidence threshold
- 0.1
- 0.3
- 0.5
- 0.7

LVIS AP_frequent (%)

Total examples seen (including repetitions)

LVIS AP_rece (%)

Total examples seen (including repetitions)

ODinW13 mean AP (%)

Total examples seen (including repetitions)
Experiments – scaling
Experiments – effects of fine-tuning
Experiments – effects of fine-tuning
Experiments – effects of fine-tuning
Experiments – effects of fine-tuning
Limitations

• Massive amount of computational resources and data required for self-training
• Trade-off between fine-tuned and open-vocabulary performance (discussed previously)
Conclusion

• Self-training can be scaled up to overcome dependency on human annotations

• OWL-ST shows significant improvements in detection performance using weak supervision from web data
Thank you!