SILCO – See a Few Images, Localize the Common Object

Tao HU, Pascal Mettes, Jia-Hong Huang and Cees G. M. Snoek
University of Amsterdam
ICCV 2019

Presented by Dax Borde
Table of Contents

• Problem Intro
• Related Work
• Method
• Results
• Conclusion
Problem Intro
Problem Intro

Inputs
1. Several Support Images (3 - 9)
2. Query Image

Output
• Bounding Box in Query
Problem Intro

Advantages
1. Less Data
2. Weak Supervision
3. Combination w/ others
Related Work
Related Work

Few Shot Classification

• Find similarities robustly
• Evaluate on novel classes
• Foundation for Similarity Mod
Related Work

**Few Shot Classification**

Siamese Networks
- One shot only
- Fully connected
  - Inserted after convolutions
  - Weights tied together
- Learn similarity

[Link to Related Work](https://www.cs.cmu.edu/~rsalakhu/papers/oneshot1.pdf)
Related Work

Few Shot Classification

Prototypical Networks
- Few shot (or 0 shot)
- Predefined classes
- Embeddings

Related Work

Graph Convolutional Networks

More information in Method section!
Weakly Supervised Detection

- Given image and label
- Localize labeled object
- Requires Lots of weak labels
- Predefined Classes
Weakly Supervised Detection

Related Work

Related Work

Object Co-Detection

- Detect same object in 2 imgs
- Predefined Classes
- CRFs for sequential detection

Related Work

How SILCO is Different

• Variable number of input images
• Less hunger for labeled data & bounding boxes
• No predefined classes
Method

Overview
Method

Backbone & Final Detection

- VGG feature extraction
- Get Query & Support features
- Multi scale box detection
  - Like SSD
Global Average Pooling: Baseline

- Pool features from baseline
  - $q_i$ - from query image
  - $S_i$ - from support images
- Average features together

$$\phi(q_i, S_i) = q_i + \frac{1}{N} \sum_{j=1}^{N} GAP(S_i^{(j)})$$
Spatial Similarity Module

- Find spatial support between query and support images
- Image wise similarity (not global)
- Convolutions & Matrix mult
- Concat with $q_i$ again at end
Feature Reweighting Module

- Some support images may be more useful than others
- More salient features may not be as important
- Uses Graph Convolutional Network
Graph Convolutional Networks

• Useful for graph datasets
  • Social networks, 3d skeletons, protein-interaction networks
• Can model other problems as graph problems
  • Eg. Image clustering

Method: Feature Reweighting Module
Graph Convolutional Networks

- Nodes represent entities (images)
  - Each Node has feature vector
- Edges are relationships
  - Random or initialized
- Start with few target labels
- Subsequent layers relax adjacent edges

Method: Feature Reweighting Module

https://tkipf.github.io/graph-convolutional-networks/
Feature Reweighting Module

- Uses GCN to reweight support images
- Predict adjacency matrix
- Graph convolutions to find important image clusters
Training

• Just one box per sample
  • Support images don’t require bounding boxes

• Losses:
  • Bounding box regression
  • Cross entropy classification

\[
L(x, c, l, g) = \frac{1}{BD} \sum_{i,j=1}^{B,D} (\text{bce}(c_{ij}, x_{ij}) + \ell^s_1(l_{ij}, g_{ij})),
\] (12)
Experimental Setup

• Datasets modified for few-shot learning
• Image datasets are divided in half
  • One for training, other for testing
  • Then swap
• CL-VID used for testing only
  • Evaluate on each frame
Experimental Setup

- Datasets modified for few-shot learning
- Image datasets are divided in half
  - One for training, other for testing
  - Then swap
- CL-VID used for testing only
  - Evaluate on each frame

Results

- Pascal VOC $\rightarrow$ CL-VOC
  - 20 class $\rightarrow$ 2x 10 class
- MS-COCO $\rightarrow$ CL-COCO
  - 80 $\rightarrow$ 2x 40
- ImageNet VID $\rightarrow$ CL-VID
  - Train on CL-VOC-12
  - 30 class, 3,862 clips
Results

Ablation

- Spatial similarity
  - Local over global
- Feature reweighting
  - Catches and removes outliers

Table 2. Ablation of spatial similarity and feature reweighting. The metric is mean average precision (%). As we adopt spatial similarity (SSM) and feature reweighting (FRM) the accuracy gradually increases over a simple global average pooling (GAP), indicating the effectiveness of our proposed modules.

<table>
<thead>
<tr>
<th>dataset</th>
<th>GAP</th>
<th>SSM</th>
<th>FRM</th>
<th>Group 1</th>
<th>Group 2</th>
<th>mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>CL-VOC-07</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>55.17</td>
<td>52.18</td>
<td>53.67</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>56.12</td>
<td>55.52</td>
<td>55.82</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>57.17</td>
<td>56.45</td>
<td>56.82</td>
</tr>
<tr>
<td>CL-VOC-12</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>53.55</td>
<td>54.61</td>
<td>54.08</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>54.04</td>
<td>57.39</td>
<td>55.71</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>55.11</td>
<td>58.62</td>
<td>56.86</td>
</tr>
<tr>
<td>CL-COCO</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>13.37</td>
<td>6.50</td>
<td>9.94</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>18.50</td>
<td>7.70</td>
<td>13.10</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>18.62</td>
<td>8.20</td>
<td>13.40</td>
</tr>
</tbody>
</table>
Ablation

- Spatial similarity
- Local over global
- Feature reweighting
- Catches and removes outliers

Results
Effect of Support Images

- Peak precision reached early
- Gap w/ baseline widens
Results

Effect of Object Size

Table 3. Effect of object size on CL-VOC-12 in mAP(%). Compared to a global average pooling (GAP) baseline, our spatial similarity (SSM) and feature reweighting modules (FRM) especially improves for medium-sized objects.

<table>
<thead>
<tr>
<th>method</th>
<th>small</th>
<th>medium</th>
<th>large</th>
</tr>
</thead>
<tbody>
<tr>
<td>GAP</td>
<td>9.60</td>
<td>10.87</td>
<td>28.64</td>
</tr>
<tr>
<td>SSM, FRM</td>
<td><strong>10.99</strong></td>
<td><strong>13.85</strong></td>
<td><strong>29.24</strong></td>
</tr>
</tbody>
</table>
Results

**Success and Failure Cases**

- Queries with multiple objects
- Multiple instances
- Corruption & bad examples
Results

Success and **Failure** Cases

1. Saliency
2. Object Size
3. Context Information
4. Instance Boundaries
## Results

### Comparative Evaluation

Table 4. **Comparative evaluation** on CL-VOC-07, CL-VOC-12, and CL-COCO. Across all datasets, our approach outperforms the center box and Region Proposal Network (RPN) [39] baselines by a large margin. Furthermore, our Spatial Similarity (SSM) and Feature Reweighting (FRM) modules are preferred over the ConvLSTM of HU et al. [24] for few-shot common-localization.

<table>
<thead>
<tr>
<th></th>
<th>CL-VOC-07</th>
<th></th>
<th>CL-VOC-12</th>
<th></th>
<th>CL-COCO</th>
<th></th>
<th>CL-VID</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>group 1</td>
<td>group 2</td>
<td>mean</td>
<td>group 1</td>
<td>group 2</td>
<td>mean</td>
<td>group 1</td>
</tr>
<tr>
<td>Center box</td>
<td>5.28</td>
<td>4.05</td>
<td>4.67</td>
<td>6.56</td>
<td>6.90</td>
<td>6.73</td>
<td>0.54</td>
</tr>
<tr>
<td>RPN [39]</td>
<td>18.72</td>
<td>15.13</td>
<td>16.93</td>
<td>20.23</td>
<td>18.54</td>
<td>19.39</td>
<td>3.20</td>
</tr>
<tr>
<td>Contrastive RPN</td>
<td>18.89</td>
<td>15.33</td>
<td>17.21</td>
<td>19.21</td>
<td>18.53</td>
<td>18.87</td>
<td>3.54</td>
</tr>
<tr>
<td><strong>This paper</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>w/ ConvLSTM [24]</td>
<td>53.46</td>
<td>54.90</td>
<td>54.18</td>
<td>51.34</td>
<td>57.42</td>
<td>54.38</td>
<td>16.37</td>
</tr>
<tr>
<td>w/ SSM and FRM</td>
<td>57.17</td>
<td>56.10</td>
<td>56.64</td>
<td>55.12</td>
<td>58.62</td>
<td>56.87</td>
<td>18.62</td>
</tr>
</tbody>
</table>

This table shows the evaluation results for different methods on three datasets: CL-VOC-07, CL-VOC-12, and CL-COCO. The methods are compared in terms of accuracy, and our approach outperforms baseline methods significantly.
Summary and Conclusion

- Few-shot common-localization
- Cheap bounding boxes
- Spatial similarity module
- Feature reweighting module
- Beats baselines by large margin
Appendix: Useful Links

• Paper Website: http://taohu.me/SILCO/
• Supplemental:
• Graph Conv Nets: https://tkipf.github.io/graph-convolutional-networks/
• Few Shot w/ Graph Nets: https://arxiv.org/pdf/1711.04043.pdf