Convolutional STN for Weakly Supervised Object Localization

Akhil Meethal
Marco Pedersoli
Soufiane Belharbi
Eric Granger
Overview

- Weakly supervised Object Localization (WSOL)
- Convolutional STN (CSTN)
- Empirical Study
- Conclusion
Weakly supervised Object Localization (WSOL)
Why WSOL?

Annotation time for different types of supervision [Bilen].
Standard WSOL Technique

CAM(Class activation Mapping)[Zhou-2016]
Convolutional STN

(a) Standard convolution with a fixed receptive field.

(b) CSTN with a varying receptive field.

STN Module and its action [Jaderberg]
Joint Probability and Multi-scale Localization

Replacing the GAP and following layers with CSTN followed by a 1x1 conv($\#\text{channels} = \#\text{classes}$) layer gives a joint probability over location and classes.

Applying CSTN and 1x1 conv over multiple layers of a feature pyramid gives a joint pdf over scale, location and classes. We used FPN.

a. Basic network  

b. Joint probability  

c. Multiscale  

d. CSTN
Overall System

Initial bounding box \(\text{argmax}_{s,l} \ p(s, l, c^*)\)

Refined bounding box by CSTN:
\[\theta_{s,l}\]

\[\sum_{s,l} p(s, l, c) = p(c) \text{ for training}\]
Additional Regularization Terms

Avoid Discriminative Boxes

\[ L_\theta = \sum_{s \in S} \sum_{i=1}^{h_s \times w_s} \| \theta_{ref} - \theta_i \|^2 \]

Reduce Localization Bias

\[ L_{scale}(x) = \max \left( 0, \max_i p(s = s_1, l, c = c^* | x) - \max_i p(s = s_2, l, c = c^* | x) \right) \]
Convolutional STN Localization - Qualitative Results
## CSTN Localization - CUB-200-2011 Test Set

<table>
<thead>
<tr>
<th>Method</th>
<th>Top-1 Loc</th>
<th>GT-Known Loc</th>
<th>Top-1 Classif</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAM [Zhou-2016]</td>
<td>41.00</td>
<td>71.13</td>
<td>-</td>
</tr>
<tr>
<td>HAS [Singh-2017]</td>
<td>44.67</td>
<td>73.32</td>
<td>76.64</td>
</tr>
<tr>
<td>ACoL [Zhang-2018]</td>
<td>45.92</td>
<td>75.30</td>
<td>71.90</td>
</tr>
<tr>
<td>SPG [Zhang-2018]</td>
<td>46.64</td>
<td>74.11</td>
<td>-</td>
</tr>
<tr>
<td>ADL [Choe-2019]</td>
<td><strong>62.29</strong></td>
<td>78.62</td>
<td>80.34</td>
</tr>
<tr>
<td>CSTN</td>
<td>49.04</td>
<td>76.06</td>
<td>78.46</td>
</tr>
<tr>
<td>CSTN Top-5(GT-Known)</td>
<td>-</td>
<td><strong>81.14</strong></td>
<td>-</td>
</tr>
</tbody>
</table>
## CSTN Localization - ILSVRC Validation Set

<table>
<thead>
<tr>
<th>Method</th>
<th>Top-1 Loc</th>
<th>GT-Known Loc</th>
<th>Top-1 Classif</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAM [Zhou-2016]</td>
<td>42.80</td>
<td>61.10</td>
<td>66.60</td>
</tr>
<tr>
<td>HAS [Singh-2017]</td>
<td>45.21</td>
<td>63.12</td>
<td>70.70</td>
</tr>
<tr>
<td>ACoL [Zhang-2018]</td>
<td>45.83</td>
<td>62.73</td>
<td>67.50</td>
</tr>
<tr>
<td>SPG [Zhang-2018]</td>
<td>48.60</td>
<td>64.24</td>
<td>-</td>
</tr>
<tr>
<td>ADL [Choe-2019]</td>
<td>48.43</td>
<td>63.72</td>
<td>75.85</td>
</tr>
<tr>
<td>CSTN</td>
<td>42.38</td>
<td>60.48</td>
<td>69.48</td>
</tr>
<tr>
<td>CSTN Top-5(GT-Known)</td>
<td>-</td>
<td>63.45</td>
<td>-</td>
</tr>
</tbody>
</table>
Conclusion

- We proposed Convolutional STN, a WSOL method with an explicit component for learning Localization.
- Additional regularizations to prevent the localization from shrinking to discriminative areas and introduce competition among different levels of the pyramid.
- Validated the proposed system on the benchmark datasets - CUB-200-2011 and ImageNet Localization.

Further improvements:

- Incorporating more layers in the pyramid so that small objects can be accurately localized; also improve the competition among levels.
- Different strategies for selecting the localization box apart from the max-scoring one; Top-5 Localization shows we have better localization boxes learned by convolutional STN.
Questions?