Weakly Supervised Localization Using Deep Feature Maps

Archith John Bency¹, Heesung Kwon², Hyungtae Lee²,³, S. Karthikeyan¹ and B. S. Manjunath¹

¹ University of California, Santa Barbara, CA USA, 93106 ² U.S. Army Research Laboratory, Adelphi, MD, USA ³ Booz Allen Hamilton Inc., McLean, VA USA

Abstract

• Object localization aims to recognize and locate interesting objects in an image
• Ground truth image bounding boxes is difficult to obtain for large-scale datasets
• Learning to localize from Image labels (Weak Supervision) is crucial

Overview:

• Train Deep CNN classifiers from Image Labels
• Propose bounding box candidates on the final Convolutional Feature Maps’s spatial grid
• Better localized candidates tend to have higher classification scores
• Rank and prune candidates using beam-search

Experimental Results on Pascal VOC 2007, 2012 and MSCOCO datasets

Introduction

Weak Supervision for Object Localization

• Strong supervision for Object localization requires object-level annotations
• Annotations include bounding boxes, segmentation maps

Ground Truth Image

Image labels are a weaker form of supervision

• Large datasets with Image labels already exist
• ImageNet, PASCAL VOC 2012, MS COCO

Can we learn how to localize objects using a dataset with only Image labels?

Deep Convolutional Neural Networks

• CNNs are state-of-the-art class of techniques for Image classification and object detection
• Unified feature learning and classification

Weakly Supervised Localization Using Deep Feature Maps

CNN classification and localization

• Correlation between CNN localization of object-of-interest and corresponding class scores
• Consequence of local nature of learnt convolutional filters
• Feature Maps: The output obtained by applying learnt convolutional filters and a non-linear function on data from previous layer

The localization algorithm operates on the final conv. layer’s Feature Map

• Alexnet: 6 x 6 x 256, VGG 16: 7 x 7 x 512, in general: L x L x N

• Localization candidates are subsets of feature maps characterized by boxes: b = [x₁, y₁, W₁, H₁]

For the box b, feature map values are re-calculated as f being an interpolation function:

\[ M_{loc}(x, y) = \begin{cases} 1 & \frac{x-x_1}{w_1} \leq x \leq \frac{x+w_1}{w_1} \\ \frac{y-y_1}{h_1} \leq y \leq \frac{y+h_1}{h_1} \\ 0 & \text{otherwise} \end{cases} \]

The candidates are back-projected onto image coordinates and further localization is performed on M_{loc}

Weakly Supervised Localization Using Deep Feature Maps (contd.)

Search Strategy

• Search for the best localization candidate is organized in a search tree
• The root node corresponds to the coarsest candidate, the entire image: b₁ = [0, 0, L, L]
• Children nodes are generated by reducing the width or height by one and ranked by resultant class score

• Beam-search is applied to prune low-ranking candidates
• Number of localization candidates are kept to be tractable
• Averts greedy decisions

Datasets and Metrics

• Datasets:
  - Pascal Visual Object Challenge (VOC) 2007, 2012: 20 object classes
  - Microsoft Common Objects in Context (MS COCO): 80 object classes

• Metrics:
  - Standard IoU detection metric
  - Object localization metric: introduced by Oqba et al., CVPR 2015
  - Correct Localization (CorLoc)

Qualitative Results

Quantitative Results

• Localization metric results on Pascal VOC 2012 validation set:

<table>
<thead>
<tr>
<th>Method</th>
<th>Localization score (mAP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>45.2</td>
</tr>
<tr>
<td>Proposed Method + Aplency</td>
<td>58.3</td>
</tr>
<tr>
<td>Proposed Method + VGG-16</td>
<td>58.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>mAP detection</th>
<th>CorLoc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multi-scale M1</td>
<td>72.4</td>
<td>38.4</td>
</tr>
<tr>
<td>Flcar et al. [26]</td>
<td>70.4</td>
<td>35.4</td>
</tr>
<tr>
<td>LCL-p2ASC [20]</td>
<td>50.9</td>
<td>58.5</td>
</tr>
<tr>
<td>Proposed Method + VGG-16</td>
<td>25.7</td>
<td>30.7</td>
</tr>
</tbody>
</table>

Acknowledgements

Research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-09-2-0053 (the ARL Network Science CTA). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.