Weakly Supervised Semantic Segmentation Using Web-Crawled Videos

Seunghoon Hong¹,²,³ Donghun Yeo¹ Suha Kwak² Honglak Lee³ Bohyung Han¹
(maga33, hanulbog)@postech.ac.kr skwak@dgist.ac.kr honglak@umich.edu bhhan@postech.ac.kr

Motivation
Improving weakly-supervised semantic segmentation by generating synthetic segmentations from web-crawled videos

- Videos from YouTube
- Video segmentation

Generate segmentation

Train classifier and attention model with image

- Class Activation Mapping
  \[ F x = \sigma(\sum_{i=0}^{L} W_i \cdot \phi_i) \]

\[ F(x) = \text{last conv. output} \quad W: \text{weights in fc layer} \quad y^c: \text{onehot vector for class c} \]

Step 1: Learning encoder with images
Train classifier and attention model with image

Classifier: CNN with global average pooling (GAP) [Jou et al. 2016]
Attention: Class Activation Mapping [Jou et al. 2016]

- Superpixel attention
- Motion likelihood
- Appearance likelihood
- Appearance similarity
- Spatial similarity
- Temporal connectivity

\[ E_b(L) = -\lambda_b \sum \log p_b(l_i) \]

\[ E_p(L) = \sum_{(l_i, l_j) \in \mathcal{E}} \big[ \lambda_r \phi_r(l_i, l_j) + \lambda_a \phi_a(l_i, l_j) \big] \]

Step 1: Learning encoder with images

Step 2: Video segmentation

Motion in videos is helpful to distinguish object from background

- Substantial noises in web-crawled videos

- Exploit both weakly labeled images and videos to compensate segmentation challenge in one data from the other

Step 2: Video segmentation with encoder outputs

Localizing object in a video using encoder outputs

- Temporal localization by filtering out irrelevant frames based on classification score
- Spatial localization by computing attention map that discriminates the target from surroundings

Video segmentation by energy minimization on spatio-temporal graph

\[ L^* = \arg\min L(L) + \sum_{c} E_p(L_c) \]

Step 3: Learning decoder with video segmentation results

- Train a decoder to map coarse attention map to dense binary mask

- Class attention as input allows to ignore objects irrespective of the labeled class.
- Class-agnostic property is useful to improve segmentation quality of static objects.

Step 3: Learning decoder with video segmentation results

- Decoder: Deconvolution network with shared pooling switch [Park et al. 2015]

Benefits:

- Motion in videos is helpful to distinguish object from background
- Videos are collected automatically by web search results

Challenges:

- Substantial noises in web-crawled videos

Our approach:

Exploit both weakly labeled images and videos to compensate segmentation challenge in one data from the other

Experiments

Semantic segmentation on PASCAL VOC 2012 dataset

- Training data:
  - Image: PASCAL VOC 2012
  - Videos: 4.6K YouTube videos collected for 20 PASCAL VOC classes

Ablation study

- Separate training with images and videos improves performance
- Collecting more videos improves performance, although obtained videos are noisy and unannotated

Comparison to SOA weakly-supervised approaches

- Substantial improvement over approaches based on image-level class labels
- Competitive performance to approaches based on heavier annotations (point, bounding box)

- Comparison to SOA weakly-supervised approaches

Supervision mlIoU (Val)

<table>
<thead>
<tr>
<th>Method</th>
<th>Supervision</th>
<th>mlIoU (Val)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEC</td>
<td>Class label</td>
<td>50.7</td>
</tr>
<tr>
<td>What’s a Point [Bearman et al. 2016]</td>
<td>Point</td>
<td>46.0</td>
</tr>
<tr>
<td>BoxSup [Dua et al. 2016]</td>
<td>Bounding box</td>
<td>62.0</td>
</tr>
<tr>
<td>Scribbleup [Dua et al. 2016]</td>
<td>Scribble</td>
<td>63.1</td>
</tr>
<tr>
<td>Ours [Jou et al. 2016]</td>
<td>Class label + Video</td>
<td>38.1</td>
</tr>
<tr>
<td>Ours</td>
<td>Class label + Video</td>
<td>58.1</td>
</tr>
</tbody>
</table>

Video segmentation on YouTube-Object dataset

<table>
<thead>
<tr>
<th>Unsupervised</th>
<th>Bounding box</th>
<th>Ours (Class label)</th>
</tr>
</thead>
<tbody>
<tr>
<td>46.8</td>
<td>56.2</td>
<td>58.6</td>
</tr>
</tbody>
</table>