Dual-Gradients Localization framework for Weakly Supervised Object Localization

Chuangchuang Tan\textsuperscript{1*}, Tao Ruan\textsuperscript{1*}, Guanghua Gu\textsuperscript{2}, Shikui Wei\textsuperscript{1}, Yao Zhao\textsuperscript{1}

\textsuperscript{1}Beijing Jiaotong University
\textsuperscript{2}Yanshan University
Weakly Supervised Object Localization (WSOL)

- WSOL is understanding an image at pixel level only using image-level annotations
- use much cheaper annotations
Steps of previous works:
- Force classification network to focus on more regions of feature map.
- Produce localization map on the last convolutional layer by applying CAM.

Problem:
- Ignore the localization ability of other layers.
- Both localization and classification tasks are trained online.

---

I can produce WSOL, too.
Dual-Gradients Localization (DGL) framework

Main ideas:

- Utilize gradients of classification loss function to mine entire target object regions.
- Leverage gradients of target class to identify the correlation ratio of pixels to the target class within any convolutional feature maps.

Characteristics:

- Simple, DGL is an offline approach, needn’t to train for localization.
- Effective, achieving localization on any convolutional layer.

Source image             Mixed_6f                    Mixed_6e
Overview of the DGL framework

Feature Maps

Class-aware Enhanced Map Branch

\[
\frac{\partial J(p, \alpha y_c)}{\partial S}
\]

\(l_2\) normalize

Enhanced map

Pixel-level Selection Branch

\[
\frac{\partial y_c}{\partial S}
\]

\(l_2\) normalize

GAP

FC

softmax

Cross-entropy loss

Classification model

Localization Maps

Beijing Jiaotong University
Classification model architecture:
- use a customized InceptionV3, i.e. SPG-plain.
- remove the layers after the second Inception block, i.e., the third Inception block, pooling and linear layer.
- add two convolutional layers
- add a GAP layer and a softmax layer
Class-aware Enhanced Map Branch

- Feature maps predicted to class c only capture the discrimination parts of objects, when the feature maps close the boundary of classification regions.
- The feature maps located at center of classification regions can highlight more object regions.

\[ \frac{\partial \text{cost}(p, \alpha y_c)}{\partial S} \]

\[ l_2 \text{ normalize} \]

Enhanced map A

Feature Maps

Class-aware Enhanced Map Branch

Decision boundary

Enhanced Map

Gradient

Feature map S
Class-aware Enhanced Map Branch

- Our key idea of Class-aware Enhanced Map is pulling the feature maps toward inside of the classification region for specific-class, along with gradients of classification loss function.
**Is gradients or weights?**

- CAM actually achieves localization by employing a weighted sum of feature maps and gradients of target class on the last convolutional layer, instead of weights of the final FC layer.

- Pixel-level Selection is a generalization to CAM.

\[ \overline{M}_c^i = \sum_k \left( \frac{\partial l_c}{\partial S_k^i} \right) \{A_c^i\}_k \]

Enhanced map A

Pixel-level Selection Branch

\[ \frac{\partial y_c}{\partial S} \]

sum and resize

\[ \ldots \]
Results on the Validation Set of LID

MS: Multi-scale inputs during test
MC: Morph close the localization map during test

<table>
<thead>
<tr>
<th>MS</th>
<th>MC</th>
<th>mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>✗</td>
<td>✗</td>
<td>58.23</td>
</tr>
<tr>
<td>✔</td>
<td>✗</td>
<td>61.46</td>
</tr>
<tr>
<td>✔</td>
<td>✔</td>
<td>62.22</td>
</tr>
</tbody>
</table>

- Fusion the localization maps of branch1 and branch2 on Mixed_6e layer.
- Input size 324
Qualitative Results

- Examples of DGL on test set
Thanks