Multiple Instance Detection Network with Online Instance Classifier Refinement

Peng Tang

pengtang@hust.edu.cn

Huazhong University of Science and Technology
Weakly-supervised visual learning (WSVL)

- Weakly-supervised visual learning is a new trend in CVPR

Search keyword “weakly supervised” (14 papers), “weakly-supervised” (5 papers), “multi-instance” (1 paper), and “multiple instance” (3 papers), 23/783 papers in total

http://cvpr2017.thecvf.com/program/main_conference

---

Huazhong University of Science and Technology
**WSVL avoids the expensive human annotations**

<table>
<thead>
<tr>
<th>Tasks</th>
<th>Training info</th>
<th>Testing output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weakly-supervised object detection (WSOD)</td>
<td>Image-level</td>
<td>Bounding box</td>
</tr>
<tr>
<td>Weakly-supervised semantic segmentation</td>
<td>Image-level</td>
<td>Pixel-level</td>
</tr>
<tr>
<td>Weakly-supervised instance segmentation</td>
<td>Image-level/bounding box</td>
<td>Instance pixel-level</td>
</tr>
<tr>
<td>Semi-supervised object detection/segmentation</td>
<td>Partial of fully-labeled data</td>
<td>Bounding box/pixel-level</td>
</tr>
</tbody>
</table>
WSVL avoids the expensive human annotations

<table>
<thead>
<tr>
<th>Tasks</th>
<th>Training info</th>
<th>Testing output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weakly-supervised object detection (WSOD)</td>
<td>Image-level</td>
<td>Bounding box</td>
</tr>
<tr>
<td>Weakly-supervised semantic segmentation</td>
<td>Image-level</td>
<td>Pixel-level</td>
</tr>
<tr>
<td>Weakly-supervised instance segmentation</td>
<td>Image-level/bounding box</td>
<td>Instance pixel-level</td>
</tr>
<tr>
<td>Semi-supervised object detection/segmentation</td>
<td>Partial of fully-labeled data</td>
<td>Bounding box/pixel-level</td>
</tr>
</tbody>
</table>
How to do WSOD

Possible solutions to this problem, clustering based, matching based, co-segmentation based, topic model based, multi-instance learning based method.
How to do WSOD

Possible solutions to this problem, clustering based, matching based, co-segmentation based, topic model based, multi-instance learning based method.

Solving this problem by multiple instance learning
• Image as bag, since image label is given
• Proposals (Selective Search, EdgeBox, Bing) as instances
  • Proposal descriptors: Deep CNN Features, Fisher Vectors
  • Number of proposals: ~2k (SS), ~4k (EB)
What is the core problem in WSOD

- Is it a bird?

The answers are YES!
What is the core problem in WSOD

- Is it a bird?

The answers are YES!

However, only some of them are correct detection results (IoU > 0.5)

Huazhong University of Science and Technology
What is the core problem in WSOD

Is it a bird?

The answers are YES!

However, only some of them are correct detection results (IoU > 0.5)
What is the core problem in WSOD

- **Is it a bird?**

The answers are YES!

However, only some of them are **correct detection results (IoU>0.5)**

Huazhong University of Science and Technology
What is the core problem in WSOD

Previous methods tend to localize parts of objects instead of whole objects.

Result of MIDN/WSDDN [4]
Multiple Instance Detection Network with Online Instance Classifier Refinement
Motivation

Proposals having **high spatial overlaps** with detected parts may cover the **whole object**, or at least contain **larger portion** of the object.

Result of MIDN/WSDDN [4]
Motivation

Propagating the scores to the highly overlapped proposals to alleviate the problem caused by ambiguity

Result of MIDN/WSDDN [4]

Result of OICR

Huazhong University of Science and Technology
The multi-instance detection network (MIDN)
The multi-instance detection network (MIDN)

- The basic network of WSDDN [4] by H. Bilen and A. Vedaldi
- Single network, end to end training
The multi-instance detection network (MIDN)

\[
[s(\mathbf{x}^c)]_{ij} = \frac{e^{x^c_{ij}}}{\sum_{k=1}^{C} e^{x^c_{kj}}}
\]

\[
\phi_c = \sum_{r=1}^{|R|} x_{cr}^R
\]

\[
[s(\mathbf{x}^d)]_{ij} = \frac{e^{x^d_{ij}}}{\sum_{k=1}^{|R|} e^{x^d_{ik}}}
\]

\[
\mathbf{x}^R = s(\mathbf{x}^c) \odot s(\mathbf{x}^d)
\]
The multi-instance detection network (MIDN)

\[ [\sigma(x^c)]_{ij} = \frac{e^{x^c_{ij}}}{\sum_{k=1}^{C} e^{x^c_{kj}}} \]

\[ \phi_c = \sum_{r=1}^{|R|} x^R_{cr} \]

\[ [\sigma(x^d)]_{ij} = \frac{e^{x^d_{ij}}}{\sum_{k=1}^{|R|} e^{x^d_{ik}}} \]

\[ x^R = \sigma(x^c) \odot \sigma(x^d) \]

\[ L_b = -\sum_{c=1}^{C} \{y_c \log \phi_c + (1 - y_c) \log(1 - \phi_c)\} \]
The network for OICR

Multiple instance detection network
- Fc layer
- Softmax over classes
- Softmax over proposals
- Element-wise product
- Sum over proposals
- Image scores

Instance classifier refinement, 1-st time
- Fc layer
- Softmax over classes
- Proposal scores
- Supervision

Instance classifier refinement, K-th time
- Fc layer
- Softmax over classes
- Proposal scores
- Supervision

Conv layers → Conv feature map → SPP layer → Proposal feature vector

Proposals
The network for OICR

- Additional blocks (instance classifiers) for score propagation
- In-network supervision

Huazhong University of Science and Technology
Effective online training/refinement
The top scoring proposal can always detect at least parts of objects.
Effective online training/refinement

- The top scoring proposal can always detect at least parts of objects.
- Proposals having high spatial overlaps with detected parts may cover larger portion of the object.
Effective online training/refinement

- The top scoring proposal can always detect at least parts of objects.
- Proposals having high spatial overlaps with detected parts may cover larger portion of the object.
- Proposals with high spatial overlap could share similar label information.
Effective online training/refinement

Algorithm 1 Online instance classifier refinement

**Input:** Image $X$ and its proposals; image label vector $Y = [y_1, \ldots, y_C]$; refinement times $K$.

**Output:** Loss weights $w_r^k$; proposal label vectors $Y_r^k = [y_r^k, \ldots, y_{(C+1)r}^k]^T$. Where $r \in \{1, \ldots, |R|\}$ and $k \in \{1, \ldots, K\}$.

1. Feed $X$ and its proposals into the network to produce proposal score matrices $x_r^{Rk}$, $k \in \{0, \ldots, K - 1\}$.
2. for $k = 0$ to $K - 1$ do
3.    Set all elements in $I = [I_1, \ldots, I_{|R|}]^T$ to $-\infty$.
4.    Set all $y_c^{k+1} = 0$, $c \in \{1, \ldots, C\}$ and $y_{(C+1)r}^{k+1} = 1$.
5.    for $c = 1$ to $C$ do
6.        if $y_c = 1$ then
7.            Choose the top-scoring proposal $j_c^k$ by Eq. (2).
8.                for $r = 1$ to $|R|$ do
9.                    Compute IoU $I_r'$ between proposal $r$ and $j_c^k$.
10.                   if $I_r' > I_r$ then
11.                       Set $I_r = I_r'$ and $w_r^{k+1} = x_c^{Rk}$.
12.                   if $I_r > I_t$ then
13.                       Set $y_{c'r}^{k+1} = 0$, $c' \neq c$ and $y_{cr}^{k+1} = 1$. 

Huazhong University of Science and Technology
Effective online training/refinement

Algorithm 1 Online instance classifier refinement

**Input:** Image $X$ and its proposals; image label vector $\mathbf{Y} = [y_1, \ldots, y_C]$; refinement times $K$.

**Output:** Loss weights $w_r^k$; proposal label vectors $\mathbf{Y}_r^k = [y_{1r}^k, \ldots, y_{(C+1)r}^k]^T$. Where $r \in \{1, \ldots, |R|\}$ and $k \in \{1, \ldots, K\}$.

1. Feed $X$ and its proposals into the network to produce proposal score matrices $\mathbf{x}^{R_k}$, $k \in \{0, \ldots, K-1\}$.
2. For $k = 0$ to $K - 1$ do
   3. Set all elements in $\mathbf{I} = [I_1, \ldots, I_{|R|}]^T$ to $-\infty$.
   4. Set all $y_{cr}^{k+1} = 0$, $c \in \{1, \ldots, C\}$ and $y_{(C+1)r}^{k+1} = 1$.
   5. For $c = 1$ to $C$ do
      6. If $y_c = 1$ then
         7. Choose the top-scoring proposal $j_c^k$ by Eq. (2).
         8. For $r = 1$ to $|R|$ do
            9. Compute IoU $I'_r$ between proposal $r$ and $j_c^k$.
            10. If $I'_r > I_r$ then
                11. Set $I_r = I'_r$ and $w_r^{k+1} = x_{cr}^{R_k}$.
                12. If $I_r > I_t$ then
                    13. Set $y_{c'r}^{k+1} = 0$, $c' \neq c$ and $y_{cr}^{k+1} = 1$.

The loss weight controls the learning process.

Huazhong University of Science and Technology
Experimental Results

- The influence of refinement times and different refinement strategies

Huazhong University of Science and Technology
## Experimental Results

- **Detection results (mAP) on VOC 2007 test set**

<table>
<thead>
<tr>
<th>Method</th>
<th>aero</th>
<th>bike</th>
<th>bird</th>
<th>boat</th>
<th>bottle</th>
<th>bus</th>
<th>car</th>
<th>cat</th>
<th>chair</th>
<th>cow</th>
<th>table</th>
<th>dog</th>
<th>horse</th>
<th>mbike</th>
<th>person</th>
<th>sheep</th>
<th>sofa</th>
<th>train</th>
<th>tv</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>WSDDN-VGG.F [4]</td>
<td>42.9</td>
<td>56.0</td>
<td>32.0</td>
<td>17.6</td>
<td>10.2</td>
<td>61.8</td>
<td>50.2</td>
<td>29.0</td>
<td>3.8</td>
<td>36.2</td>
<td>18.5</td>
<td>31.1</td>
<td>45.8</td>
<td>54.5</td>
<td>10.2</td>
<td>15.4</td>
<td>36.3</td>
<td>45.2</td>
<td>50.1</td>
<td>43.8</td>
</tr>
<tr>
<td>WSDDN-VGG.M [4]</td>
<td>43.6</td>
<td>50.4</td>
<td>32.2</td>
<td>26.0</td>
<td>9.8</td>
<td>58.5</td>
<td>50.4</td>
<td>30.9</td>
<td>7.9</td>
<td>36.1</td>
<td>18.2</td>
<td>31.7</td>
<td>41.4</td>
<td>52.6</td>
<td>8.8</td>
<td>14.0</td>
<td>37.8</td>
<td>46.9</td>
<td>53.4</td>
<td>47.9</td>
</tr>
<tr>
<td>WSDDN-VGG16 [4]</td>
<td>39.4</td>
<td>50.1</td>
<td>31.5</td>
<td>16.3</td>
<td>12.6</td>
<td>64.5</td>
<td>42.8</td>
<td>42.6</td>
<td>10.1</td>
<td>35.7</td>
<td>24.9</td>
<td>38.2</td>
<td>34.4</td>
<td>55.6</td>
<td>9.4</td>
<td>14.7</td>
<td>30.2</td>
<td>40.7</td>
<td>54.7</td>
<td>46.9</td>
</tr>
<tr>
<td>WSDDN+context [16]</td>
<td>57.1</td>
<td>52.0</td>
<td>31.5</td>
<td>7.6</td>
<td>11.5</td>
<td>55.0</td>
<td>53.1</td>
<td>34.1</td>
<td>1.7</td>
<td>33.1</td>
<td>42.0</td>
<td>42.3</td>
<td>56.6</td>
<td>15.3</td>
<td>12.8</td>
<td>24.8</td>
<td>48.9</td>
<td>44.4</td>
<td>47.8</td>
<td>36.3</td>
</tr>
<tr>
<td>OICR-VGG.M</td>
<td>53.1</td>
<td>57.1</td>
<td>32.4</td>
<td>12.3</td>
<td>15.8</td>
<td>58.2</td>
<td>56.7</td>
<td>39.6</td>
<td>0.9</td>
<td>44.8</td>
<td>39.9</td>
<td>31.0</td>
<td>54.0</td>
<td>62.4</td>
<td>4.5</td>
<td>20.6</td>
<td>39.2</td>
<td>38.1</td>
<td>48.9</td>
<td>48.6</td>
</tr>
<tr>
<td>OICR-VGG16</td>
<td>58.0</td>
<td>62.4</td>
<td>31.1</td>
<td>19.4</td>
<td>13.0</td>
<td>65.1</td>
<td>62.2</td>
<td>28.4</td>
<td>24.8</td>
<td>44.7</td>
<td>30.6</td>
<td>25.3</td>
<td>37.8</td>
<td>65.5</td>
<td>15.7</td>
<td>24.1</td>
<td>41.7</td>
<td>46.9</td>
<td>64.3</td>
<td>62.6</td>
</tr>
<tr>
<td>WSDDN-Ens. [4]</td>
<td>46.4</td>
<td>58.3</td>
<td>35.5</td>
<td>25.9</td>
<td>14.0</td>
<td>66.7</td>
<td>53.0</td>
<td>39.2</td>
<td>8.9</td>
<td>41.8</td>
<td>26.6</td>
<td>38.6</td>
<td>44.7</td>
<td>59.0</td>
<td>10.8</td>
<td>17.3</td>
<td>40.7</td>
<td>49.6</td>
<td>56.9</td>
<td>50.8</td>
</tr>
<tr>
<td>OM+MIL+FRCNN [20]</td>
<td>54.5</td>
<td>47.4</td>
<td>41.3</td>
<td>20.8</td>
<td>17.7</td>
<td>51.9</td>
<td>63.5</td>
<td>46.1</td>
<td>21.8</td>
<td>57.1</td>
<td>22.1</td>
<td>34.4</td>
<td>50.5</td>
<td>61.8</td>
<td>16.2</td>
<td>29.9</td>
<td>40.7</td>
<td>15.9</td>
<td>55.3</td>
<td>40.2</td>
</tr>
<tr>
<td>OICR-Ens.</td>
<td>58.5</td>
<td>63.0</td>
<td>35.1</td>
<td>16.9</td>
<td>17.4</td>
<td>63.2</td>
<td>60.8</td>
<td>34.4</td>
<td>8.2</td>
<td>49.7</td>
<td>41.0</td>
<td>31.3</td>
<td>51.9</td>
<td>64.8</td>
<td>13.6</td>
<td>23.1</td>
<td>41.6</td>
<td>48.4</td>
<td>58.9</td>
<td>58.7</td>
</tr>
<tr>
<td>OICR-Ens.+FRCNN</td>
<td>65.5</td>
<td>67.2</td>
<td>47.2</td>
<td>21.6</td>
<td>22.1</td>
<td>68.0</td>
<td>68.5</td>
<td>35.9</td>
<td>5.7</td>
<td>63.1</td>
<td>49.5</td>
<td>30.3</td>
<td>64.7</td>
<td>66.1</td>
<td>13.0</td>
<td>25.6</td>
<td>50.0</td>
<td>57.1</td>
<td>60.2</td>
<td>59.0</td>
</tr>
</tbody>
</table>

- **Detection results (mAP) on VOC 2012 test set**

<table>
<thead>
<tr>
<th>Method</th>
<th>aero</th>
<th>bike</th>
<th>bird</th>
<th>boat</th>
<th>bottle</th>
<th>bus</th>
<th>car</th>
<th>cat</th>
<th>chair</th>
<th>cow</th>
<th>table</th>
<th>dog</th>
<th>horse</th>
<th>mbike</th>
<th>person</th>
<th>sheep</th>
<th>sofa</th>
<th>train</th>
<th>tv</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>WSDDN+context [16]</td>
<td>64.0</td>
<td>54.9</td>
<td>36.4</td>
<td>8.1</td>
<td>12.6</td>
<td>53.1</td>
<td>40.5</td>
<td>28.4</td>
<td>6.6</td>
<td>35.3</td>
<td>34.4</td>
<td>49.1</td>
<td>42.6</td>
<td>62.4</td>
<td>19.8</td>
<td>15.2</td>
<td>27.0</td>
<td>33.1</td>
<td>33.0</td>
<td>50.0</td>
</tr>
<tr>
<td>OICR-VGG.M</td>
<td>64.4</td>
<td>50.6</td>
<td>34.8</td>
<td>16.7</td>
<td>16.5</td>
<td>49.7</td>
<td>44.8</td>
<td>20.4</td>
<td>5.0</td>
<td>39.0</td>
<td>18.2</td>
<td>46.2</td>
<td>50.3</td>
<td>64.3</td>
<td>4.3</td>
<td>15.1</td>
<td>32.4</td>
<td>38.5</td>
<td>36.3</td>
<td>45.1</td>
</tr>
<tr>
<td>OICR-VGG16</td>
<td>67.7</td>
<td>61.2</td>
<td>41.5</td>
<td>25.6</td>
<td>22.2</td>
<td>54.6</td>
<td>49.7</td>
<td>25.4</td>
<td>19.9</td>
<td>47.0</td>
<td>18.1</td>
<td>26.0</td>
<td>38.9</td>
<td>67.7</td>
<td>2.0</td>
<td>22.6</td>
<td>41.1</td>
<td>34.3</td>
<td>37.9</td>
<td>55.3</td>
</tr>
<tr>
<td>OICR-Ens.</td>
<td>68.4</td>
<td>58.6</td>
<td>39.9</td>
<td>23.4</td>
<td>21.3</td>
<td>52.8</td>
<td>48.7</td>
<td>23.5</td>
<td>13.5</td>
<td>44.8</td>
<td>22.0</td>
<td>36.5</td>
<td>47.6</td>
<td>68.3</td>
<td>2.6</td>
<td>21.7</td>
<td>39.1</td>
<td>39.6</td>
<td>38.2</td>
<td>52.7</td>
</tr>
<tr>
<td>OICR-Ens.+FRCNN</td>
<td>71.4</td>
<td>69.4</td>
<td>55.1</td>
<td>29.8</td>
<td>28.1</td>
<td>55.0</td>
<td>57.9</td>
<td>24.4</td>
<td>17.2</td>
<td>59.1</td>
<td>21.8</td>
<td>26.6</td>
<td>57.8</td>
<td>71.3</td>
<td>1.0</td>
<td>23.1</td>
<td>52.7</td>
<td>37.5</td>
<td>33.5</td>
<td>56.6</td>
</tr>
</tbody>
</table>
Experimental Results

Huazhong University of Science and Technology
Conclusion

- Advantages
  - Our method can improve the detection results a lot through our OICR strategy, especially for rigid objects.
  - The network can be trained in a very efficiently (online) way.
Conclusion

- **Advantages**
  - Our method can improve the detection results a lot through our OICR strategy, especially for rigid objects.
  - The network can be trained in a very efficiently (online) way.

- **Limitations**
  - The performance is poor for non-rigid objects, as these objects are always with great deformation and their representative parts are with much less deformation.
Thank you!


- Codes are available at https://github.com/ppengtang/oicr.