Cascade RPN: Delving into High-Quality Region Proposal Network with Adaptive Convolution

Thang Vu  Hyunjun Jang  Pham X. Trung  Chang D. Yoo

Korea Advanced Institute of Science and Technology
The proposed method aims to improve the RPN in stage 1.
Region proposal network

• I: Input image
• Backbone: Feature extractor
• H: Head (shared)
• C: Classifier
• A: Anchor regressor

Region proposal network [1]

Alignment in RPN

Extractor feature

Refine anchor box

CNN

Image Space

Feature space

Correspondence = Alignment
Iterative RPN

RPN [1]

Iterative RPN [2]

Misalignment
Anchor shape and position change after being refined

Iterative RPN+ and GA-RPN

**RPN [1]**
- I → Backbone
- C Conv → H Conv
- A Conv → H Conv

**Iterative RPN [2]**
- I → Backbone
- C Conv → H1 Conv
- A Conv → H1 Conv
- C2 Conv → H2 Conv
- A2 Conv → H2 Conv

**Iterative RPN+ [3]**
- I → Backbone
- C Conv → H1 Conv
- A Conv → H1 Conv
- C2 Conv → H2 Conv
- A2 Conv → H2 DefConv
- Offset Conv → H2 DefConv

**GA-RPN [4]**
- I → Backbone
- C Conv → H1 Conv
- A Conv → H1 Conv
- C Conv → H2 DefConv

**Misalignment**
- Arbitrary feature transform
- No constrains for alignment

References:
Proposed Cascade RPN

- RPN [1]
- Iterative RPN [2]
- Iterative RPN+ [3]
- GA-RPN [4]
- Cascade RPN (ours)

Adaptive Convolution

- Standard Convolution
  - Sample at regular grid \( \mathbb{R} \)
    \[
    y[p] = \sum_{r \in \mathbb{R}} w[r] \cdot x[p + r]
    \]
    \[
    \mathbb{R} = \{(-1,-1), (-1,0), \ldots, (0,1), (1,1)\}
    \]

- Adaptive Convolution
  - Sample at offset grid \( \emptyset \), guided by anchor
    \[
    y[p] = \sum_{o \in \emptyset} w[o] \cdot x[p + o]
    \]
    \[
    o = o_{\text{ctr}} + o_{\text{shp}}
    \]

Adaptive conv systematically maintain alignment between features and anchors!
Sampling location

Experiments

- Dataset: COCO2017 [1]
  - Train: 115k images
  - Val: 5k images
  - Test-dev: 20k images
- Evaluation metric:
  - Average Recall (AR) for Region Proposal performance
  - Average Precision (AP) for Detection performance
  - Runtime is measured on a single V100

<table>
<thead>
<tr>
<th>Method</th>
<th>Backbone</th>
<th>AR$_{100}$</th>
<th>AR$_{300}$</th>
<th>AR$_{1000}$</th>
<th>AR$_S$</th>
<th>AR$_M$</th>
<th>AR$_L$</th>
<th>Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SharpMask [1]</td>
<td>ResNet-50</td>
<td>36.4</td>
<td>-</td>
<td>48.2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.76</td>
</tr>
<tr>
<td>GCN-NS [2]</td>
<td>VGG-16</td>
<td>31.6</td>
<td>-</td>
<td>60.7</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.10</td>
</tr>
<tr>
<td>AttractioNet [3]</td>
<td>VGG-16</td>
<td>53.3</td>
<td>-</td>
<td>66.2</td>
<td>31.5</td>
<td>62.2</td>
<td>77.7</td>
<td>4.00</td>
</tr>
<tr>
<td>ZIP [4]</td>
<td>BN-Inception</td>
<td>53.9</td>
<td>-</td>
<td>76.0</td>
<td>31.9</td>
<td>63.0</td>
<td>78.5</td>
<td>1.13</td>
</tr>
</tbody>
</table>

**RPN [5]**
- Iterative RPN       | 48.5          | 55.4       | 58.8       | 32.1       | 56.9   | 65.4   | 0.05   |
- Iterative RPN+      | ResNet-50     | 54.0       | 60.4       | 63.0       | 35.6   | 62.7   | 73.9   | 0.06     |
| GA-RPN [6]           | ResNet-50     | 59.1       | 65.1       | 68.5       | 40.7   | 68.2   | 78.4   | 0.06     |
| Cascade RPN          | **61.1**      | **67.6**   | **71.7**   | **42.1**   | **69.3** | **82.8** | **0.06** |

## Region Proposal Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Backbone</th>
<th>AR&lt;sub&gt;100&lt;/sub&gt;</th>
<th>AR&lt;sub&gt;300&lt;/sub&gt;</th>
<th>AR&lt;sub&gt;1000&lt;/sub&gt;</th>
<th>AR&lt;sub&gt;S&lt;/sub&gt;</th>
<th>AR&lt;sub&gt;M&lt;/sub&gt;</th>
<th>AR&lt;sub&gt;L&lt;/sub&gt;</th>
<th>Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SharpMask [1]</td>
<td>ResNet-50</td>
<td>36.4</td>
<td>-</td>
<td>48.2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.76</td>
</tr>
<tr>
<td>GCN-NS [2]</td>
<td>VGG-16</td>
<td>31.6</td>
<td>-</td>
<td>60.7</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.10</td>
</tr>
<tr>
<td>AttractioNet [3]</td>
<td>VGG-16</td>
<td>53.3</td>
<td>-</td>
<td>66.2</td>
<td>31.5</td>
<td>62.2</td>
<td>77.7</td>
<td>4.00</td>
</tr>
<tr>
<td>ZIP [4]</td>
<td>BN-inception</td>
<td>53.9</td>
<td>-</td>
<td>76.0</td>
<td>31.9</td>
<td>63.0</td>
<td>78.5</td>
<td>1.13</td>
</tr>
<tr>
<td>RPN [5]</td>
<td></td>
<td>44.6</td>
<td>52.9</td>
<td>58.3</td>
<td>29.5</td>
<td>51.7</td>
<td>61.4</td>
<td>0.04</td>
</tr>
<tr>
<td>Iterative RPN</td>
<td></td>
<td>48.5</td>
<td>55.4</td>
<td>58.8</td>
<td>32.1</td>
<td>56.9</td>
<td>65.4</td>
<td>0.05</td>
</tr>
<tr>
<td>Iterative RPN+</td>
<td>ResNet-50</td>
<td>54.0</td>
<td>60.4</td>
<td>63.0</td>
<td>35.6</td>
<td>62.7</td>
<td>73.9</td>
<td>0.06</td>
</tr>
<tr>
<td>GA-RPN [6]</td>
<td></td>
<td>59.1</td>
<td>65.1</td>
<td>68.5</td>
<td>40.7</td>
<td>68.2</td>
<td>78.4</td>
<td>0.06</td>
</tr>
<tr>
<td>Cascade RPN</td>
<td></td>
<td><strong>61.1 (+2.0)</strong></td>
<td><strong>67.6 (+2.5)</strong></td>
<td><strong>71.7 (+3.2)</strong></td>
<td><strong>42.1 (+1.4)</strong></td>
<td><strong>69.3 (+1.1)</strong></td>
<td><strong>82.8 (+4.4)</strong></td>
<td><strong>0.06 (+0.0)</strong></td>
</tr>
</tbody>
</table>

Qualitative Results

Stage 1

Stage 2
Qualitative Results

Stage 1

Stage 2
## Detection Results

<table>
<thead>
<tr>
<th>Detector</th>
<th>Proposal method</th>
<th>AP</th>
<th>AP&lt;sub&gt;50&lt;/sub&gt;</th>
<th>AP&lt;sub&gt;75&lt;/sub&gt;</th>
<th>AP&lt;sub&gt;S&lt;/sub&gt;</th>
<th>AP&lt;sub&gt;M&lt;/sub&gt;</th>
<th>AP&lt;sub&gt;L&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fast R-CNN [1]</strong></td>
<td>RPN [2]</td>
<td>36.6</td>
<td>58.6</td>
<td>39.5</td>
<td>20.3</td>
<td>39.1</td>
<td>47.0</td>
</tr>
<tr>
<td></td>
<td>Iterative RPN+</td>
<td>38.8</td>
<td>58.8</td>
<td>42.2</td>
<td>21.1</td>
<td>41.5</td>
<td>50.0</td>
</tr>
<tr>
<td></td>
<td>GA-RPN [3]</td>
<td>39.5</td>
<td>59.3</td>
<td>43.2</td>
<td>21.8</td>
<td>42.0</td>
<td>50.7</td>
</tr>
<tr>
<td></td>
<td>Cascade RPN</td>
<td>40.1</td>
<td>59.4</td>
<td>43.8</td>
<td>22.1</td>
<td>42.4</td>
<td>51.6</td>
</tr>
<tr>
<td><strong>Faster R-CNN [2]</strong></td>
<td>RPN [2]</td>
<td>36.9</td>
<td>58.9</td>
<td>39.9</td>
<td>21.1</td>
<td>39.6</td>
<td>46.5</td>
</tr>
<tr>
<td></td>
<td>Iterative RPN+</td>
<td>39.2</td>
<td>58.2</td>
<td>43.0</td>
<td>21.5</td>
<td>42.0</td>
<td>50.4</td>
</tr>
<tr>
<td></td>
<td>GA-RPN [3]</td>
<td>39.9</td>
<td><strong>59.4</strong></td>
<td>43.6</td>
<td>22.0</td>
<td>42.6</td>
<td>50.9</td>
</tr>
<tr>
<td></td>
<td>Cascade RPN</td>
<td>40.6</td>
<td>58.9</td>
<td><strong>44.5</strong></td>
<td><strong>22.0</strong></td>
<td><strong>42.8</strong></td>
<td><strong>52.6</strong></td>
</tr>
</tbody>
</table>

Summary

• Alignment is not well persevered in existing multi-stage RPN.
• Cascade RPN systematically ensures alignment by Adaptive Convolution.
• Cascade RPN achieves state-of-the-art proposal performance on COCO dataset.

Code is available at: https://github.com/thangvubk/Cascade-RPN

Poster #86 at East Exhibition Hall B + C

Thank you!