Improving object localization using multiple layer features and scale optimization

Heungwoo Han, Seongmin Kang, Seonghoon Kim, Kibum Bae, Vitaly Lavrukhin
{hwoo.han, cheese.kang, sh0221.kim, kibum85.bae, v.lavrukhin}@samsung.com

The bounding box network is based on the Inception architecture and features from mid and last conv layers are used for bounding box regression.

To improve localization performance, multi-scale predictions with two differently trained models are fused and an optimal scale is chosen.

Procedure

① Regression network: The network is designed to regress on bounding box locations and the final outputs are normalized by sigmoid function. Data augmentation of random crop which involves more than 50% of width and height of GT is used for training to increase performance.

② Modified greedy prediction: Greedy prediction is applied on each scale. The sum of coordinate distances is used as a match score function.

B: bounding box set, b_n: bounding box in set B
1. Repeat merging
2. (b_n', b_n'') = argmin_{b_n, b_n'} Coordinate_Distance(b_n, b_n')
3. if Coordinate_Distance < threshold t
4. stop
5. else
6. remove b_n', b_n'' from B
7. b_n = avg(b_n', b_n'')
8. add b_n to B
9. assign confidence of b_n as sum of confidence b_n', b_n''
10. Choose b_n which has max confidence

③ Scale selection: Optimal scale for bounding box size is chosen.

L: bounding box list sorted by scale, w: sliding window size
1. For each b in L
2. occupation ratio = max(b_n.height / w, b_n.width / w)
3. if occupation rate > threshold r
4. choose b

④ Ensemble: If IOU of 1st bounding box between base model and other models, 4th and 5th bounding box are replaced by 1st bounding box of other models.

Result

<table>
<thead>
<tr>
<th>Models</th>
<th>Training Scale</th>
<th>Test Scale</th>
<th>Localization Error (Classification Error: 0.0539)</th>
</tr>
</thead>
<tbody>
<tr>
<td>224 x 224 PCR (M1)</td>
<td>256 x 256</td>
<td>224 x 224</td>
<td>0.2987</td>
</tr>
<tr>
<td>224 x 224 PCR (M2)</td>
<td>256 x N</td>
<td>224 x 224</td>
<td>0.2989</td>
</tr>
<tr>
<td>M1 + Greedy Prediction + Scale Selection</td>
<td>-</td>
<td>(224 ~ 608) x N</td>
<td>0.2815</td>
</tr>
<tr>
<td>M1, M2 Fusioning + Greedy Prediction + Scale Selection</td>
<td>-</td>
<td>(224 x 224: M1 (256 ~ 608) x N: M2)</td>
<td>0.2715</td>
</tr>
<tr>
<td>Ensemble</td>
<td>-</td>
<td>-</td>
<td>0.2312</td>
</tr>
</tbody>
</table>

References

1. Szegedy et al., Going Deeper with Convolutions. CVPR, 2015.