Scene Graph Generation by Iterative Message Passing

Danfei Xu, Yuke Zhu, Christopher B. Choy, Li Fei-Fei (2017)

https://arxiv.org/abs/1701.02426
Scene Graph

In every image, there's more than meets the object detector
Problem statement

Given an image $I$ and a set of boxes $B$ from pretrained Region Proposal Network, we want to identify:

- **Object classes**
  For each box, the object class

- **Box offsets**
  For each box, the offset w.r.t. the proposed box coordinates

- **Pairwise relationships**
  For each pair of boxes, the most likely relationship between their objects
The baseline

Visual Relationship Detection with Language Priors
Cewu Lu, Ranjay Krishna, Michael Bernstein, Li Fei-Fei (2016)

- Uses visual features from the region containing 2 objects
- Uses language priors to cluster relationships together
- Similar to the approach of this paper, without message passing
Primal Dual Graph

- **Nodes**: Represent objects in the scene
- **Edges**: Represent object relationships
Primal Dual Graph

Object nodes
Represent objects in the scene

Relationship nodes
Represent object relationships

Edges
Represent messages exchanged between object and relationship nodes

Object nodes and relationship nodes form a bipartite graph
Gated Recurrent Unit

Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation

Relationship node A ➔ B

\[
m_{i \rightarrow j} = \sigma(w_1^T[h_i, h_{i \rightarrow j}])h_i + \sigma(w_2^T[h_j, h_{i \rightarrow j}])h_j
\]

\[
h_{i \rightarrow j} = GRU(m_{i \rightarrow j}, h_{i \rightarrow j})
\]
Object node B

\[
m_i = \sum_{j: i \rightarrow j} \sigma(v_1^T [h_i, h_{i \rightarrow j}])h_{i \rightarrow j} + \sum_{j: j \rightarrow i} \sigma(v_2^T [h_i, h_{j \rightarrow i}])h_{j \rightarrow i}
\]

\[
h_i = GRU(m_i, h_i)
\]
Training procedure

- Pretrained VGG-16 for region proposals and visual feature
- 512-dimensional vectors for state and messages
- For each image, 128 boxes are randomly selected from the top 2,000 proposed boxes
- For each image, 128 labeled relationships are randomly selected from the 8,128 possible object pairs
- For inference, only the top 50 boxes and all their pairs are considered
Visual Genome

- 100k images
- Top 150 object classes (avg. 25 per image)
- Top 50 relationships (avg. 6.2 per image)

Note
This is a cleaned version of the VG dataset, because the original annotations were of poor quality.
## Visual Genome

<table>
<thead>
<tr>
<th>Model</th>
<th>R@50</th>
<th>R@100</th>
<th>[26]</th>
<th>avg. pool</th>
<th>max pool</th>
<th>final</th>
</tr>
</thead>
<tbody>
<tr>
<td>PredCls</td>
<td>27.88</td>
<td>35.04</td>
<td>32.39</td>
<td>34.33</td>
<td>41.99</td>
<td>44.75</td>
</tr>
<tr>
<td></td>
<td>35.04</td>
<td>41.99</td>
<td>39.63</td>
<td>41.99</td>
<td></td>
<td>53.08</td>
</tr>
<tr>
<td>SGCls</td>
<td>11.79</td>
<td>14.11</td>
<td>15.65</td>
<td>16.31</td>
<td>18.70</td>
<td>21.72</td>
</tr>
<tr>
<td></td>
<td>14.11</td>
<td>24.38</td>
<td>18.27</td>
<td>18.70</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SGGen</td>
<td>0.32</td>
<td>0.47</td>
<td>2.70</td>
<td>3.03</td>
<td>3.71</td>
<td>3.44</td>
</tr>
<tr>
<td></td>
<td>0.47</td>
<td>4.24</td>
<td>3.42</td>
<td>3.71</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Visual Genome
<table>
<thead>
<tr>
<th>predicate</th>
<th>[26]</th>
<th>ours</th>
<th>predicate</th>
<th>[26]</th>
<th>ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>on</td>
<td>99.71</td>
<td>99.25</td>
<td>under</td>
<td>28.64</td>
<td>52.73</td>
</tr>
<tr>
<td>has</td>
<td>98.03</td>
<td>97.25</td>
<td>sitting on</td>
<td>31.74</td>
<td>50.17</td>
</tr>
<tr>
<td>in</td>
<td>80.38</td>
<td>88.30</td>
<td>standing on</td>
<td>44.44</td>
<td>61.90</td>
</tr>
<tr>
<td>of</td>
<td>82.47</td>
<td>96.75</td>
<td>in front of</td>
<td>26.09</td>
<td>59.63</td>
</tr>
<tr>
<td>wearing</td>
<td>98.47</td>
<td>98.23</td>
<td>attached to</td>
<td>8.45</td>
<td>29.58</td>
</tr>
<tr>
<td>near</td>
<td>85.16</td>
<td>96.81</td>
<td>at</td>
<td>54.08</td>
<td>70.41</td>
</tr>
<tr>
<td>with</td>
<td>31.85</td>
<td>88.10</td>
<td>hanging from</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>above</td>
<td>49.19</td>
<td>79.73</td>
<td>over</td>
<td>9.26</td>
<td>0.00</td>
</tr>
<tr>
<td>holding</td>
<td>61.50</td>
<td>80.67</td>
<td>for</td>
<td>12.20</td>
<td>31.71</td>
</tr>
<tr>
<td>behind</td>
<td>79.35</td>
<td>92.32</td>
<td>riding</td>
<td>72.43</td>
<td>89.72</td>
</tr>
</tbody>
</table>
NYU Depth v2

- 1,449 RGB-D images
- 4 object classes (floor, structure, furniture, prop)
- 3 support relationships (behind, below, hidden)

Notes

The depth channel is not used during the experiments.

Ground-truth object locations are provided as inputs, not predicted.
### NYU Depth v2

<table>
<thead>
<tr>
<th></th>
<th>Support Accuracy</th>
<th>PREDCLS</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>t-ag</td>
<td>t-aw</td>
<td>R@50</td>
<td>R@100</td>
</tr>
<tr>
<td>Silberman <em>et al.</em> [28]</td>
<td>75.9</td>
<td>72.6</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Liao <em>et al.</em> [24]</td>
<td>88.4</td>
<td>82.1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Baseline [26]</td>
<td>87.7</td>
<td>85.3</td>
<td>34.1</td>
<td>50.3</td>
</tr>
<tr>
<td>Final model (ours)</td>
<td>91.2</td>
<td>89.0</td>
<td>41.8</td>
<td>55.5</td>
</tr>
</tbody>
</table>
Discussion