Visual Relationship Detection with Multiple Cues

Arun Mallya, Bryan A. Plummer, Svetlana Lazebnik
University of Illinois at Urbana-Champaign

Overview

Visual Relationship Detection

Proposed Visual and Language Cues

1. Image-Class Compatibility Scores (6-dim)
   Measures compatibility of regions with proposed classes and relationships using normalized Canonical Correlation Analysis (CCA).

2. Subject/Object Size (2-dim)
   Captures tendency of objects to be of certain size and scale in images.

3. Subject/Object Position (2-dim)
   Captures tendency of objects to be in certain positions in images.

4. Relative Subject/Object Position (1-dim)
   Measures compatibility of relative subject-object positions with predicate.

Final Model:
Rank-SVM on concatenated features \( \Phi = [\Phi_{\text{CCA}}, \Phi_{\text{size}}, \Phi_{\text{pos}}, \Phi_{\text{rel-pos}}] \)
- Negative Proposals from EdgeBoxes
- Test-time proposals provided by detectors trained in [1]

Results

Correctly detected relationships

Logically correct detected relationships, penalized as false positives

Incorrectly detected relationships

Phrase and Relationship detection recall at different thresholds on the VRD Dataset [1].

<table>
<thead>
<tr>
<th>Method</th>
<th>Phrase Detection</th>
<th>Relationship Detection</th>
<th>Zero-shot Phrase Detection</th>
<th>Zero-shot Relationship Detection</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Recall @50</td>
<td>Recall @100</td>
<td>Recall @50</td>
<td>Recall @100</td>
</tr>
<tr>
<td>Visual-Only Model [1]</td>
<td>2.24</td>
<td>2.61</td>
<td>1.58</td>
<td>1.85</td>
</tr>
<tr>
<td>CCA + Size</td>
<td>11.72</td>
<td>15.85</td>
<td>10.36</td>
<td>14.05</td>
</tr>
<tr>
<td>CCA + Size + Position</td>
<td>16.89</td>
<td>20.70</td>
<td>15.08</td>
<td>18.37</td>
</tr>
</tbody>
</table>

Key Takeaways:
- Competitive or better performance than previous methods, without the use of end-to-end training and complicated models.
- Position cues greatly help improve performance as relationships constrain mutual position.
- Continuous embeddings and simpler CCA models help generalization resulting in improved performance for the zero-shot setting.

References