Hierarchical Visual Relationship Detection

Xu Sun\(^1\), Yuan Zi\(^1\), Tongwei Ren\(^{1,3}\), Jinhui Tang\(^2\), Gangshan Wu\(^1\)

\(^1\) State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
\(^2\) School of Computer Science, Nanjing University of Science and Technology, Nanjing, China
\(^3\) Shenzhen Research Institute of Nanjing University, Shenzhen, China
Motivation and Solution

• **Hierarchical visual relationship detection (HVRD)** encourages predicting abstract yet compatible relationship triplets when the confidence level of the specific image content is relatively low.

• Our solution
  • **Hierarchical concept embedding**: embed concepts in different abstraction levels with order embedding
  • **Hierarchical object detection**: trade off specificity for accuracy with a vision and knowledge joint model
  • **Hierarchical predicate detection**: combining visual feature and context information
Experiments

• Datasets: H-VRD and H-VG
  • Construct two datasets for HVRD by extending VRD and VG datasets
• Evaluation criterion: \( \text{recall}@N \ (k=\alpha) \)

\[
\varphi^S(g, r) = \begin{cases} 
\frac{d_{S'}}{d_{gS}}, & r^S \in T_{gS}, \\
0, & \text{otherwise},
\end{cases}
\]

triplet score
\[
\varphi(g, r) = \begin{cases} 
0, & \varphi^S(g, r) \cdot \varphi^P(g, r) \cdot \varphi^O(g, r) = 0 \\
\frac{1}{3} (\varphi^S(g, r) + \varphi^P(g, r) + \varphi^O(g, r)), & \text{otherwise}
\end{cases}
\]

component score

• Comparison
  • Task: HPD and HVRD
  • Result: our method is superior to the state-of-the-art baselines on all the criteria

<table>
<thead>
<tr>
<th>Method</th>
<th>HR@50</th>
<th>HR@100</th>
<th>BR@50</th>
<th>BR@100</th>
<th>HR@50</th>
<th>HR@100</th>
<th>BR@50</th>
<th>BR@100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lin’s</td>
<td>50.32</td>
<td>50.32</td>
<td>50.75</td>
<td>50.75</td>
<td>13.81</td>
<td>14.92</td>
<td>13.84</td>
<td>15.26</td>
</tr>
<tr>
<td>VTS</td>
<td>50.08</td>
<td>50.08</td>
<td>50.59</td>
<td>50.59</td>
<td>11.84</td>
<td>13.95</td>
<td>12.04</td>
<td>15.15</td>
</tr>
<tr>
<td>DR-net</td>
<td>53.62</td>
<td>53.62</td>
<td>54.02</td>
<td>54.02</td>
<td>14.80</td>
<td>16.90</td>
<td>14.84</td>
<td>17.50</td>
</tr>
<tr>
<td>DSR</td>
<td>54.19</td>
<td>54.23</td>
<td>54.71</td>
<td>54.79</td>
<td>14.64</td>
<td>16.82</td>
<td>14.68</td>
<td>17.46</td>
</tr>
<tr>
<td>Ours</td>
<td>60.28</td>
<td>60.28</td>
<td>66.20</td>
<td>66.20</td>
<td>15.94</td>
<td>18.66</td>
<td>17.03</td>
<td>19.94</td>
</tr>
</tbody>
</table>

H-VRD dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>HR@50</th>
<th>HR@100</th>
<th>BR@50</th>
<th>BR@100</th>
<th>HR@50</th>
<th>HR@100</th>
<th>BR@50</th>
<th>BR@100</th>
</tr>
</thead>
<tbody>
<tr>
<td>VTS</td>
<td>64.44</td>
<td>64.66</td>
<td>65.24</td>
<td>65.47</td>
<td>6.19</td>
<td>8.17</td>
<td>6.21</td>
<td>8.63</td>
</tr>
<tr>
<td>DSR</td>
<td>64.27</td>
<td>68.56</td>
<td>65.12</td>
<td>69.47</td>
<td>0.31</td>
<td>0.57</td>
<td>0.32</td>
<td>0.57</td>
</tr>
<tr>
<td>Ours</td>
<td>73.89</td>
<td>73.99</td>
<td>76.11</td>
<td>76.25</td>
<td>9.40</td>
<td>11.29</td>
<td>9.77</td>
<td>11.74</td>
</tr>
</tbody>
</table>

H-VG dataset