Heterogeneous Learning for Scene Graph Generation

Yunqing He¹
Tongwei Ren¹
Jinhui Tang²
Gangshan Wu¹

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
Scene Graph Generation

Goals
- localize holistic object instances
- recognize their relationships

Challenges
- long-tail data distribution
- sparse samples on triplet categories
- large intra-class variation and high inter-class similarity

Application
- captioning
- retrieval
- visual question answering
- multi-modal dialog

An example of scene graph
Motivation

- Heterogeneity between objects and relationships has not been discussed yet.
- Heterogeneous objects and relation feature spaces can alleviate the large intra-class variation and inter-class ambiguity problem.
Related Work

Plug-and-play methods can be easily attached to any other type of method and strengthen their effectiveness.

- **RNN-based methods**
  - boy
  - skateboard
  - ground
  - jump
  - on

- **GNN-based methods**
  - boy
  - on
  - jump
  - ground
  - skateboard

- **SGG Models**
  - Training Plugin
  - Inference Plugin

- **Knowledge Graph**
  - Knowledge Models
  - Priors

- **Knowledge-based methods**
  - <boy, jump, skateboard>
  - <boy, on, ground>

- **Plug-and-play methods**
  - <boy, on, skateboard>
  - <boy, jump, skateboard>

- **Heavy-feature and weak-network methods**
  - boy
  - skateboard
  - ground
  - jump
  - on
Our Method – Framework

- Initialize relation representation with Feature Transformation Module
- Find possible relation proposals with Link Prediction Module
- Construct heterogeneous object and relation features spaces with Object Prediction Confusion Module
- Propagate the heterogeneity to arbitrary SGG relation predictors with Auto Encoder Module

HLB only works as training branch: It means NO additional inference cost
Our Method – Feature Transformation

- GNN usually suffers from over-smooth problem
- Severe long-tail problem in VG dataset exacerbates the over-smooth problem
- Alleviate over-smooth problem by enhancing each node’s original feature

\[ x'_i = \sigma(\omega_1 \cdot F_{j \in N(i)}(x_j)) \]

\[ x'_i = \sigma(\omega_1 \cdot F_{j \in N(i)}(x_j) + \omega_2 \cdot x_i) \]

<table>
<thead>
<tr>
<th>Method</th>
<th>mR@20</th>
<th>mR@50</th>
<th>mR@100</th>
<th>R@20</th>
<th>R@50</th>
<th>R@100</th>
</tr>
</thead>
<tbody>
<tr>
<td>GCN</td>
<td>3.14</td>
<td>4.17</td>
<td>4.83</td>
<td>23.05</td>
<td>29.58</td>
<td>33.54</td>
</tr>
<tr>
<td>GCN+</td>
<td>4.02</td>
<td>5.45</td>
<td>6.32</td>
<td>24.53</td>
<td>31.47</td>
<td>35.52</td>
</tr>
<tr>
<td>GAT</td>
<td>3.96</td>
<td>5.35</td>
<td>6.21</td>
<td>24.55</td>
<td>31.41</td>
<td>35.50</td>
</tr>
<tr>
<td>GAT+</td>
<td>4.06</td>
<td>5.50</td>
<td>6.40</td>
<td>24.58</td>
<td>31.28</td>
<td>35.29</td>
</tr>
<tr>
<td>HLB</td>
<td>3.96</td>
<td>5.27</td>
<td>6.09</td>
<td>24.36</td>
<td>31.08</td>
<td>35.13</td>
</tr>
<tr>
<td>HLB+</td>
<td>4.34</td>
<td>5.87</td>
<td>6.84</td>
<td>24.78</td>
<td>31.79</td>
<td>35.91</td>
</tr>
</tbody>
</table>
Our Method – Link Prediction

- **Hierarchical Link Prediction Module**
  - Probability between two isolated objects \( P_{(i,j)} \)
  - Probability between two objects with consideration of context \( P_{(i,j)|\text{context}} \)
  - Probability of all possible existing relations \( P_{R|\text{context}} \)

What is the relationship between these two object?

In current scene, what is the relationship between these two object?

What could happen in current scene?
Our Method – Object Prediction Confusion

- make the object and relation features less relative
  Problem \(\downarrow\) Reduction

- make relation features contain minimum object information
  Problem \(\downarrow\) Reduction

- make relation features cannot be used in object recognition

Relevance between high-dimension tensors is difficult in quantification
Our Method – Auto Encoder

• Defect of a classifier (only Encoder):
  • Since training two different classifiers simultaneously (classifier from conventional SGG method and that from our HLB method) is difficult, a possible dilemma is both the classifiers tend to generate logits that close to zero (sparse logits) to make the loss value seems to decease.

• Advantage of an Auto Encoder (Encoder + Decoder):
  • The better the prediction logits can be reconstructed, the more information is preserved in the logits. It means that the classifier/Encoder tend to generate dense/non-zero logits.

![Diagram of Auto Encoder]

Reconstruction

Encoder

Decoder

relation logits A

relation logits B

Conventional SGG Predictors

Optimization Goal: minimize(A-B)

Ensure that relation feature could be reconstructed, so as to ensure that more information can be preserved
Datasets: Visual Genome (VG-150)
- 108,077 images
- 1,366,673 object instances
- 1,531,448 relation instances
- 108,249 isolated scene graphs
- 150 object categories
- 50 relation categories

Tasks
- Scene Graph Detection
- Scene Graph Classification
- Predicate Classification

Evaluation metrics
- R@N (recall in top-N results)
- mR@N (mean recall over classes in top-N results)
- ng-R@N (no graph-constraint recall in top-N results)
- zR@N (zero-shot recall in top-N results)
Experimental Results – Toy Experiments

Restricted fusion between object and relation features can significantly improve the performance, i.e., heterogeneity between object and relation features is worth noting.

\[
\text{obj\_logits} = \text{predictor(feature[0])}
\]

\[
\text{obj\_logits} = \text{predictor(feature[-1])}
\]

<table>
<thead>
<tr>
<th>SGDet</th>
<th>mR</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMP</td>
<td>3.8</td>
<td>4.8</td>
</tr>
<tr>
<td>IMP-H</td>
<td>5.37</td>
<td>6.30</td>
</tr>
</tbody>
</table>
### Experimental Results – Comparison Results

<table>
<thead>
<tr>
<th>Model</th>
<th>PredCls</th>
<th></th>
<th></th>
<th>SGCls</th>
<th></th>
<th></th>
<th>SGDet</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mR@50</td>
<td>R@50</td>
<td>mR@100</td>
<td>R@100</td>
<td>mR@50</td>
<td>R@50</td>
<td>mR@100</td>
<td>R@100</td>
<td>mR@50</td>
</tr>
<tr>
<td>GBNet-β [34]</td>
<td>22.1</td>
<td>66.6</td>
<td>24.0</td>
<td>68.2</td>
<td>12.7</td>
<td>37.3</td>
<td>13.4</td>
<td>38.0</td>
<td>7.1</td>
</tr>
<tr>
<td>Graph R-CNN [32]</td>
<td>16.4</td>
<td>54.2</td>
<td>17.2</td>
<td>59.1</td>
<td>9.0</td>
<td>29.6</td>
<td>9.5</td>
<td>31.6</td>
<td>5.8</td>
</tr>
<tr>
<td>ReIDN [38]</td>
<td>15.8</td>
<td>68.7</td>
<td>17.2</td>
<td>68.8</td>
<td>9.3</td>
<td>38.9</td>
<td>9.6</td>
<td>38.9</td>
<td>6.0</td>
</tr>
<tr>
<td>FCSGG [17]</td>
<td>6.3</td>
<td>41.0</td>
<td>7.1</td>
<td>45.0</td>
<td>3.7</td>
<td>23.5</td>
<td>4.1</td>
<td>25.7</td>
<td>3.6</td>
</tr>
<tr>
<td>GPS-Net [31]</td>
<td>19.2</td>
<td>69.7</td>
<td>21.4</td>
<td>69.7</td>
<td>11.7</td>
<td>42.3</td>
<td>12.5</td>
<td>42.3</td>
<td>7.4</td>
</tr>
<tr>
<td>IMP [29]</td>
<td>9.8</td>
<td>59.3</td>
<td>10.5</td>
<td>61.3</td>
<td>5.8</td>
<td>34.6</td>
<td>6.0</td>
<td>35.4</td>
<td>3.8</td>
</tr>
<tr>
<td>IMP + HLB</td>
<td>10.63</td>
<td><strong>60.91</strong></td>
<td><strong>11.37</strong></td>
<td><strong>62.95</strong></td>
<td>6.62</td>
<td><strong>38.10</strong></td>
<td><strong>6.98</strong></td>
<td><strong>39.01</strong></td>
<td>4.19</td>
</tr>
<tr>
<td>IMP-H</td>
<td>10.17</td>
<td>58.89</td>
<td>10.97</td>
<td>61.31</td>
<td>6.05</td>
<td>34.89</td>
<td>6.47</td>
<td>36.59</td>
<td>5.37</td>
</tr>
<tr>
<td>IMP-H + HLB</td>
<td>10.44</td>
<td>59.43</td>
<td>11.17</td>
<td>61.52</td>
<td>7.07</td>
<td>38.21</td>
<td>7.47</td>
<td>39.09</td>
<td>5.87</td>
</tr>
<tr>
<td>VTransE [37]</td>
<td>14.7</td>
<td>65.7</td>
<td>15.8</td>
<td>67.6</td>
<td>8.2</td>
<td>38.6</td>
<td>8.7</td>
<td>39.4</td>
<td>5.0</td>
</tr>
<tr>
<td>VTransE + HLB</td>
<td>15.26</td>
<td>65.68</td>
<td><strong>16.40</strong></td>
<td>67.60</td>
<td>8.24</td>
<td><strong>39.72</strong></td>
<td>8.74</td>
<td><strong>40.61</strong></td>
<td>5.14</td>
</tr>
<tr>
<td>KERN [3]</td>
<td>17.7</td>
<td><strong>65.8</strong></td>
<td>19.2</td>
<td>67.6</td>
<td>9.4</td>
<td>36.7</td>
<td>10.0</td>
<td>37.4</td>
<td>6.4</td>
</tr>
<tr>
<td>KERN + HLB</td>
<td>15.89</td>
<td>61.17</td>
<td>17.15</td>
<td>64.17</td>
<td>9.01</td>
<td><strong>38.16</strong></td>
<td>9.69</td>
<td>39.37</td>
<td>7.11</td>
</tr>
<tr>
<td>MOTIFS [36]</td>
<td>14.0</td>
<td>65.2</td>
<td>15.3</td>
<td>67.1</td>
<td>7.7</td>
<td>35.8</td>
<td>8.2</td>
<td>36.5</td>
<td>5.7</td>
</tr>
<tr>
<td>MOTIFS + HLB</td>
<td>15.39</td>
<td>64.91</td>
<td><strong>16.74</strong></td>
<td>66.80</td>
<td>8.90</td>
<td><strong>39.48</strong></td>
<td>9.44</td>
<td><strong>40.32</strong></td>
<td>7.19</td>
</tr>
<tr>
<td>VCTree-SL [24]</td>
<td>17.0</td>
<td>66.2</td>
<td>18.5</td>
<td>67.9</td>
<td>9.8</td>
<td>37.9</td>
<td>10.5</td>
<td>38.6</td>
<td>6.7</td>
</tr>
<tr>
<td>VCTree-SL + HLB</td>
<td>17.47</td>
<td>65.73</td>
<td><strong>18.79</strong></td>
<td>67.35</td>
<td><strong>11.98</strong></td>
<td><strong>36.95</strong></td>
<td>12.73</td>
<td><strong>38.50</strong></td>
<td><strong>7.46</strong></td>
</tr>
<tr>
<td>BGNN [13]</td>
<td>30.4</td>
<td>59.2</td>
<td>32.9</td>
<td>61.3</td>
<td>14.3</td>
<td>37.4</td>
<td>16.5</td>
<td>38.5</td>
<td>10.7</td>
</tr>
<tr>
<td>BGNN + HLB</td>
<td>28.20</td>
<td>61.06</td>
<td>30.43</td>
<td>63.22</td>
<td><strong>16.72</strong></td>
<td><strong>35.27</strong></td>
<td>18.09</td>
<td><strong>36.64</strong></td>
<td><strong>12.57</strong></td>
</tr>
<tr>
<td><strong>On Average</strong></td>
<td><strong>-1.45%</strong></td>
<td><strong>-0.21%</strong></td>
<td><strong>+0.96%</strong></td>
<td><strong>-0.02%</strong></td>
<td><strong>+11.73%</strong></td>
<td><strong>+4.08%</strong></td>
<td><strong>+10.74%</strong></td>
<td><strong>+4.39%</strong></td>
<td><strong>+12.63%</strong></td>
</tr>
</tbody>
</table>

**Comparison with the state-of-the-arts methods**
Experimental Results – Quantitative Analysis

• Component Analysis
  • AD: remove decoder from Auto-Encoder
  • LP: remove Link Prediction Module
  • GE: remove over-smooth-proof item from GNN

• Feature Representation Analysis

<table>
<thead>
<tr>
<th></th>
<th>PredCls mR @20</th>
<th>PredCls R @20</th>
<th>SGCls mR @20</th>
<th>SGCls R @20</th>
<th>SGDet mR @20</th>
<th>SGDet R @20</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMP-H</td>
<td>8.94</td>
<td>51.47</td>
<td>5.01</td>
<td>30.37</td>
<td>3.98</td>
<td>24.45</td>
</tr>
<tr>
<td></td>
<td>10.17</td>
<td>58.89</td>
<td>6.05</td>
<td>34.89</td>
<td>5.37</td>
<td>31.21</td>
</tr>
<tr>
<td></td>
<td>10.97</td>
<td>61.31</td>
<td>6.47</td>
<td>36.59</td>
<td>6.30</td>
<td>35.36</td>
</tr>
<tr>
<td>IMP-H-AD</td>
<td>7.78</td>
<td>51.67</td>
<td>4.83</td>
<td>30.49</td>
<td>3.90</td>
<td>24.49</td>
</tr>
<tr>
<td></td>
<td>9.67</td>
<td>58.95</td>
<td>5.80</td>
<td>35.00</td>
<td>5.25</td>
<td>31.32</td>
</tr>
<tr>
<td></td>
<td>10.43</td>
<td>61.38</td>
<td>6.21</td>
<td>36.68</td>
<td>6.14</td>
<td>35.49</td>
</tr>
<tr>
<td>IMP-H-LP</td>
<td>7.72</td>
<td>51.61</td>
<td>4.76</td>
<td>30.42</td>
<td>4.02</td>
<td>24.44</td>
</tr>
<tr>
<td></td>
<td>9.54</td>
<td>58.94</td>
<td>5.69</td>
<td>34.91</td>
<td>5.38</td>
<td>31.26</td>
</tr>
<tr>
<td></td>
<td>10.24</td>
<td>61.36</td>
<td>6.09</td>
<td>36.56</td>
<td>6.23</td>
<td>35.39</td>
</tr>
<tr>
<td>IMP-H-GE</td>
<td>7.76</td>
<td>50.74</td>
<td>4.83</td>
<td>29.85</td>
<td>3.87</td>
<td>23.17</td>
</tr>
<tr>
<td></td>
<td>9.82</td>
<td>58.28</td>
<td>5.85</td>
<td>34.20</td>
<td>5.27</td>
<td>30.03</td>
</tr>
<tr>
<td></td>
<td>10.66</td>
<td>60.90</td>
<td>6.27</td>
<td>35.80</td>
<td>6.23</td>
<td>34.36</td>
</tr>
<tr>
<td>IMP-H-HLB</td>
<td>8.50</td>
<td>52.73</td>
<td>5.84</td>
<td>34.89</td>
<td>4.34</td>
<td>24.78</td>
</tr>
<tr>
<td></td>
<td>10.44</td>
<td>59.43</td>
<td>7.07</td>
<td>38.21</td>
<td>5.87</td>
<td>31.79</td>
</tr>
<tr>
<td></td>
<td>11.17</td>
<td>61.52</td>
<td>7.47</td>
<td>39.09</td>
<td>6.81</td>
<td>35.91</td>
</tr>
</tbody>
</table>

Higher intra-class similarity
Less inter-class ambiguity

Component analysis

Feature representation analysis
Experimental Results – Qualitative Analysis

- The words marked with green denote the correctly detected objects and relations.
- The red words and lines represent the wrongly predicted ones with notated labels in brackets.
- The words marked with black color refer to the predicted relations which are considered positive but unlabeled.

Qualitative results of the proposed method.
Thank You

heyq@smail.nju.edu.cn