Improving Visual Relationship Detection using Semantic Modeling of Scene Descriptions

Stephan Baier, Yunpu Ma, and Volker Tresp
October 24, 2017
Motivation
Motivation

person-on-motorcycle
Motivation

person-on-motorcycle

person-wear-helmet
Motivation

person-on-motorcycle

person-wear-helmet

motorcycle-has-wheel
Motivation

person-on-motorcycle

motorcycle-has-wheel

person-wear-helmet

person-wear-jacket
## Motivation

<table>
<thead>
<tr>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image_url" alt="Image of a person on a motorcycle" /></td>
<td>person-on-motorcycle</td>
</tr>
<tr>
<td></td>
<td>motorcycle-has-wheel</td>
</tr>
<tr>
<td></td>
<td>person-wear-helmet</td>
</tr>
<tr>
<td></td>
<td>person-wear-jacket</td>
</tr>
</tbody>
</table>

- Deriving triples from images
- Triples describe the objects in an image and their relationship
- The relationships among objects capture important semantics
Challenges

frequent observation

person-ride-horse

infrequent observation

person-ride-elefant

• Object detectors are imperfect and provide probabilities
• Combinatorial complexity ($R \times E \times E$ possible triples)
• Not all combinations have been observed during training
Approach

• Building a probabilistic model which captures prior knowledge about the semantics
• Building a computer vision pipeline for detecting objects and relationships (Lu et al. 2016)
• Combine both in a probabilistic graphical model
Probabilistic knowledge graph is serving as semantic prior
Adjust probabilities based on graph structure
Generalization to missing links
Construct a s-p-o tensor including the counts for each triple from the training dataset

Train a low rank approximation to the tensor (RESCAL, DistMult, ComplEx, Multiway Neural network)

Poisson cost function for count data

Transfer the factorization score into probability using a Boltzmann distribution

Output: $P(s, p, o)$
Find object regions using Region-CNN.
Classify objects and relations in the proposed regions using VGG-16.
Output:
\[ P(s|i_s), P(p|i_p), P(o|i_o) \]
Find object regions using Region-CNN
Classify objects and relations in the proposed regions using VGG-16
Output: $P(s|i_s), P(o|i_o), P(p|i_p)$
• Find object regions using Region-CNN
• Classify objects and relations in the proposed regions using VGG-16
• **Output:** $P(s|i_s), P(p|i_p), P(o|i_o)$
Combination

- **Combine**: prior $P(s, p, o)$ with $P(s|i_s)$, $P(p|i_p)$, and $P(o|i_o)$
- We assume a joint model with latent variables $s, p, o$ and the visual regions as observed variables
- Conditional independence between the visual regions
- Convert using Bayes’ Rule
- **Output**: $P(s, p, o|i_s, i_p, i_o) \propto P(s, p, o) \frac{P(s|i_s) \cdot P(p|i_p) \cdot P(o|i_o)}{P(s) \cdot P(p) \cdot P(o)}$
Experiment Setting

- Experiments on the Stanford visual relationship detection dataset; Lu et al. 2016
- 4000 training images (fine-tuned vision models from image-net)
- 100 objects, 70 relationships
- Various types of relations
  - attributes (e.g. has part, wear)
  - spatial relations (e.g. on, next to)
  - comparative relations (e.g. taller than)
  - actions (e.g. ride, kick)
- Multiple evaluation settings (including and not including object detection)
## Visual Relationship Detection

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R@100</td>
<td>R@50</td>
<td>R@100</td>
<td>R@50</td>
</tr>
<tr>
<td>Evaluation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lu et al. V</td>
<td>2.61</td>
<td>1.85</td>
<td>7.11</td>
<td>2.68</td>
</tr>
<tr>
<td></td>
<td>2.24</td>
<td>1.58</td>
<td>7.11</td>
<td>2.30</td>
</tr>
<tr>
<td>Lu et al. full</td>
<td>17.03</td>
<td>14.70</td>
<td>47.87</td>
<td>18.11</td>
</tr>
<tr>
<td></td>
<td>16.17</td>
<td>13.86</td>
<td>47.87</td>
<td>17.11</td>
</tr>
<tr>
<td>RESCAL</td>
<td>19.17</td>
<td>16.88</td>
<td>52.71</td>
<td>20.23</td>
</tr>
<tr>
<td>MultiwayNN</td>
<td>18.88</td>
<td>16.65</td>
<td>51.82</td>
<td>19.76</td>
</tr>
<tr>
<td></td>
<td>17.75</td>
<td>15.57</td>
<td>51.82</td>
<td>18.53</td>
</tr>
<tr>
<td>ComplEx</td>
<td>19.36</td>
<td>17.12</td>
<td>53.14</td>
<td>20.23</td>
</tr>
<tr>
<td></td>
<td>18.25</td>
<td>16.03</td>
<td>53.14</td>
<td>19.06</td>
</tr>
<tr>
<td>DistMult</td>
<td>15.42</td>
<td>13.64</td>
<td>42.18</td>
<td>16.14</td>
</tr>
<tr>
<td></td>
<td>14.27</td>
<td>12.54</td>
<td>42.18</td>
<td>14.94</td>
</tr>
</tbody>
</table>

- All decompositions except for DistMult improve over the language prior used by Lu et al.
- Correct object detection is a bottleneck
- ComplEx factorization model gives best results
Experiment Examples

😊 person-next to-person

😊 truck-on-road
Experiment Examples

- person-next to-person
- lamp-on-box
- truck-on-road
- motorcycle-has-wheel
## Zero-shot Results

### Zero-shot learning

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R@100</td>
<td>R@50</td>
<td>R@100</td>
<td>R@50</td>
<td>R@100</td>
<td>R@50</td>
<td>R@100</td>
<td>R@50</td>
</tr>
<tr>
<td>Evaluation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lu et al. V</td>
<td>1.12</td>
<td>0.95</td>
<td>0.78</td>
<td>0.67</td>
<td>3.52</td>
<td>3.52</td>
<td>1.20</td>
<td>1.03</td>
</tr>
<tr>
<td>Lu et al. full</td>
<td>3.75</td>
<td>3.36</td>
<td>3.52</td>
<td>3.13</td>
<td>8.45</td>
<td>8.45</td>
<td>5.39</td>
<td>4.79</td>
</tr>
<tr>
<td>RESCAL</td>
<td>6.59</td>
<td>5.82</td>
<td>6.07</td>
<td>5.30</td>
<td>16.34</td>
<td>16.34</td>
<td>6.07</td>
<td>5.30</td>
</tr>
<tr>
<td>MultiwayNN</td>
<td>6.93</td>
<td>5.73</td>
<td>6.24</td>
<td>5.22</td>
<td>16.60</td>
<td>16.60</td>
<td>6.24</td>
<td>5.22</td>
</tr>
<tr>
<td>ComplEx</td>
<td>6.50</td>
<td>5.73</td>
<td>5.82</td>
<td>5.05</td>
<td>15.74</td>
<td>15.74</td>
<td>5.82</td>
<td>5.05</td>
</tr>
<tr>
<td>DistMult</td>
<td>4.19</td>
<td>3.34</td>
<td>3.85</td>
<td>3.08</td>
<td>12.40</td>
<td>12.40</td>
<td>3.85</td>
<td>3.08</td>
</tr>
</tbody>
</table>

- Tensor decompositions especially improve generalization to unseen triples
- Generalization is particularly important when the number of entity and relations grows
- Multiway Neural Networks give the best result
Zero-shot Examples

- bus-next to-bus
- tree-behind-bear
- laptop-on-stove
- bear-ride-motorcycle
Zero-shot Examples

- bus-next to-bus
- tree-behind-bear
- laptop-on-stove
- bear-ride-motorcycle
Zero-shot Examples

- **bus**-next to-bus
- **tree**-behind-bear
- **laptop**-on-stove
- **bear**-ride-motorcycle
Zero-shot Examples

- **bus-next to-bus**
- **tree-behind-bear**
- **laptop-on-stove**
- **bear-ride-motorcycle**

V: 😊
L: 😊
S: 😊
Conclusion

- Extracting semantic triples from images
- Semantic triples capture complex semantics and are well structured
- Significant improvement with probabilistic prior using relational model
- Detection and classification of multiple objects in an image is still difficult
- Framework allows for integrating richer ontologies
Thank you for your attention!