What Value Do Explicit High Level Concepts Have in Vision to Language Problems? Supplementary Material

Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony Dick, Anton van den Hengel  
School of Computer Science, The University of Adelaide, Australia  
{qi.wu01,chunhua.shen,lingqiao.liu,anthony.dick,anton.vandenhengel}@adelaide.edu.au

1. Image Captioning Results on the Flickr

In this section, we report results on the Flickr8k [3] and Flickr30k [8]. These datasets contain 8,000 and 31,000 images respectively, and each image is annotated with 5 sentences. In our reported results, we use pre-defined splits for Flickr8k, 1000 for validation, 1000 for testing and the rest for training. For Flickr30k, we report results with the widely used publicly available splits in the work of [4], which use 1000 images for validation, 1000 for testing.

<table>
<thead>
<tr>
<th>Flickr8K</th>
<th>B-1</th>
<th>B-2</th>
<th>B-3</th>
<th>B-4</th>
<th>PPL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Karpathy &amp; Li (NeuralTalk) [4]</td>
<td>0.58</td>
<td>0.38</td>
<td>0.25</td>
<td>0.16</td>
<td>-</td>
</tr>
<tr>
<td>Chen &amp; Zintick (Mind's Eye) [1]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>15.10</td>
</tr>
<tr>
<td>GoogleNIC [6]</td>
<td>0.66</td>
<td>0.42</td>
<td>0.27</td>
<td>0.18</td>
<td>-</td>
</tr>
<tr>
<td>Mao et al. (m-Rnn-AlexNet) [5]</td>
<td>0.57</td>
<td>0.39</td>
<td>0.26</td>
<td>0.17</td>
<td>24.39</td>
</tr>
<tr>
<td>Xu et al. (Hard-Attention) [7]</td>
<td>0.67</td>
<td>0.46</td>
<td>0.31</td>
<td>0.21</td>
<td>-</td>
</tr>
</tbody>
</table>

Baseline - CNN(I)

VggNet-LSTM 0.56 0.37 0.24 0.16 15.71  
VggNet-PCA+LSTM 0.56 0.38 0.25 0.16 16.07  
GoogleLeNet+LSTM 0.56 0.38 0.24 0.16 15.71  
VggNet+LSTM 0.64 0.43 0.30 0.20 14.69

Ours - Valt

Attributes-GT+LSTM ‡ 0.76 0.57 0.41 0.29 12.52  
Attributes-SVM+LSTM 0.73 0.53 0.38 0.26 12.63  
Attributes-CNN+LSTM 0.74 0.54 0.38 0.27 12.60

Flickr30K

<table>
<thead>
<tr>
<th>Flickr30K</th>
<th>B-1</th>
<th>B-2</th>
<th>B-3</th>
<th>B-4</th>
<th>PPL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Karpathy &amp; Li (NeuralTalk) [4]</td>
<td>0.57</td>
<td>0.37</td>
<td>0.24</td>
<td>0.16</td>
<td>-</td>
</tr>
<tr>
<td>Chen &amp; Zintick (Mind's Eye) [1]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>19.10</td>
</tr>
<tr>
<td>GoogleNIC [6]</td>
<td>0.66</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Donahue et al. (LRCN) [2]</td>
<td>0.59</td>
<td>0.39</td>
<td>0.25</td>
<td>0.17</td>
<td>-</td>
</tr>
<tr>
<td>Mao et al. (m-Rnn-AlexNet) [5]</td>
<td>0.54</td>
<td>0.36</td>
<td>0.23</td>
<td>0.15</td>
<td>35.11</td>
</tr>
<tr>
<td>Mao et al. (m-Rnn-VggNet) [5]</td>
<td>0.60</td>
<td>0.41</td>
<td>0.28</td>
<td>0.19</td>
<td>20.72</td>
</tr>
<tr>
<td>Xu et al. (Hard-Attention) [7]</td>
<td>0.67</td>
<td>0.44</td>
<td>0.30</td>
<td>0.20</td>
<td>-</td>
</tr>
</tbody>
</table>

Baseline - CNN(I)

VggNet-LSTM 0.57 0.38 0.25 0.17 18.83  
VggNet-PCA+LSTM 0.59 0.40 0.26 0.17 18.92  
GoogleLeNet+LSTM 0.58 0.39 0.26 0.17 18.77  
VggNet+LSTM 0.67 0.47 0.31 0.21 16.62

Ours - Valt

Attributes-GT+LSTM ‡ 0.78 0.57 0.42 0.30 14.88  
Attributes-SVM+LSTM 0.68 0.49 0.33 0.23 16.01  
Attributes-CNN+LSTM 0.73 0.55 0.40 0.28 15.96

Table 1. BLEU-1,2,3,4 and PPL metrics compared to other state-of-the-art methods and our baseline on Flickr8k and Flickr30k dataset. ‡ indicates ground truth attribute labels are used, which (in gray) will not participate in rankings. Our PPLs are based on Flickr8k word dictionaries of size 2538 and Flickr30k word dictionaries of size 7414.

Results Table 1 reports image captioning results on Flickr8k, Flickr30k. The Attributes-GT+LSTM models perform best over all datasets and all evaluation metrics, because the ground truth attributes are used. Apart from using ground truth attributes, our Attributes-CNN+LSTM models generate the best results on both Flickr8k and Flickr30K over all evaluation metrics.

2. Some Example Results

This section shows some qualitative results demonstrating our attribute predictions, image captions and question answering.

Figure 1. Some qualitative results demonstrating our attribute predictions, image captions and question answering. Ground truth answers are in parentheses. Blue indicates we give the right answer, red means we are wrong.
Figure 2. Some qualitative results demonstrating our attribute predictions, image captions and question answering. Ground truth answers are in parentheses. Blue indicates we give the right answer, red means we are wrong.
Figure 3. Some qualitative results demonstrating our attribute predictions, image captions and question answering. Ground truth answers are in parentheses. Blue indicates we give the right answer, red means we are wrong.
Figure 4. Some qualitative results demonstrating our attribute predictions, image captions and question answering. Ground truth answers are in parentheses. Blue indicates we give the right answer, red means we are wrong.
References


