What Does BERT with Vision Look At?

Liunian Harold Li  
UCLA

Mark Yatskar  
AI2

Da Yin  
PKU

Cho-Jui Hsieh  
UCLA

Kai-Wei Chang  
UCLA

BERT with Vision: Pre-trained Vision-and-language (V&L) Models

Several people walking on a sidewalk in the rain with umbrellas.

Several people [MASK] on a [MASK] in the [MASK] with [MASK].

Pre-train on image captions and transfer to visual question answering
BERT with Vision: Pre-trained Vision-and-language (V&L) Models

<table>
<thead>
<tr>
<th>Task</th>
<th>Baseline</th>
<th>VisualBERT</th>
</tr>
</thead>
<tbody>
<tr>
<td>VQA</td>
<td>68.71</td>
<td>70.80</td>
</tr>
<tr>
<td>VCR</td>
<td>44.0</td>
<td>52.4</td>
</tr>
<tr>
<td>NLVR²</td>
<td>53.5</td>
<td>67.3</td>
</tr>
<tr>
<td>Flickr30K</td>
<td>69.69</td>
<td>71.33</td>
</tr>
</tbody>
</table>

Performance of VisualBERT compared to strong baselines

- Mask and predict on image captions
- Transformer over image regions and texts
- Significant improvement over baselines
- ViLBERT, B2T2, LXMERT, VisualBERT, Unicoder-VL, VL-BERT, UNITER, …
What does BERT with Vision learn during pre-training?

Entity grounding
Map entities to regions
Probing attention maps of VisualBERT: Entity Grounding

Certain heads can perform entity grounding

Accuracy peaks in higher layers
What does BERT with Vision learn during pre-training?

Syntactic grounding

Map $w_1$ to regions of $w_2$, if $w_1 \xrightarrow{r} w_2$
### Probing attention maps of VisualBERT: Syntactic Grounding

<table>
<thead>
<tr>
<th>Type</th>
<th>Baseline</th>
<th>Acc</th>
<th>Head</th>
</tr>
</thead>
<tbody>
<tr>
<td>det</td>
<td>19.59</td>
<td>54.01</td>
<td>10-1</td>
</tr>
<tr>
<td>pobj</td>
<td>17.34</td>
<td>32.82</td>
<td>11-11</td>
</tr>
<tr>
<td>amod</td>
<td>18.67</td>
<td>45.96</td>
<td>10-9</td>
</tr>
<tr>
<td>nsubj</td>
<td>23.19</td>
<td>44.64</td>
<td>5-1</td>
</tr>
<tr>
<td>prep</td>
<td>20.61</td>
<td>49.27</td>
<td>9-11</td>
</tr>
<tr>
<td>dobj</td>
<td>9.82</td>
<td>30.24</td>
<td>11-11</td>
</tr>
<tr>
<td>punct</td>
<td>23.32</td>
<td>48.80</td>
<td>3-6</td>
</tr>
<tr>
<td>partmod</td>
<td>21.41</td>
<td>38.15</td>
<td>4-9</td>
</tr>
<tr>
<td>nn</td>
<td>16.33</td>
<td>34.06</td>
<td>10-9</td>
</tr>
<tr>
<td>num</td>
<td>23.15</td>
<td>67.44</td>
<td>9-11</td>
</tr>
</tbody>
</table>

For each dependency relationship, there exists at least one accurate syntax grounding head
Probing attention maps of VisualBERT: Syntactic Grounding

Syntactic grounding accuracy peaks in higher layers
Probing attention maps of VisualBERT: Qualitative Example

Accurate entity and syntax grounding

Refined understanding over the layers
Discussion

Previous work

Pre-trained language models learn the classical NLP pipeline (Peters et al., 2018; Liu et al., 2019; Tenney et al., 2019)

Qualitatively, V&L models learn some entity grounding (Yang et al., 2016; Anderson et al., 2018; Kim et al., 2018)

Grounding can be learned using dedicated methods (Xiao et al., 2017; Datta et al., 2019)

Our paper

BERT with Vision learns grounding through pre-training

We quantitively verify both entity and syntactic grounding

https://github.com/uclanlp/visualbert