**KB-VLP: Knowledge Based Vision and Language Pretraining**

**Motivation**
- Vision-Language Pretraining (VLP) has received increasing attention.
- Existing models ignore external knowledge.
- Models should consider both:
  - multiple modalities
  - rich structural information in knowledge.
- Knowledge embeddings in pretraining improve:
  - Standard VL tasks
  - Commonsense tasks

**Contributions**
- Knowledge-based self-supervised pretraining.
- Use Wikidata to get external knowledge.
- Experiments and analysis demonstrate the effectiveness of our approach.

**Knowledge Based Vision-Language Multi-Layer Transformer**

**Experiments**

<table>
<thead>
<tr>
<th>VQA</th>
<th>NLVR2</th>
<th>GQA</th>
<th>OK-VQA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>Dev</td>
<td>Test-std</td>
<td>Dev</td>
</tr>
<tr>
<td>NSM (Gao et al., 2019)</td>
<td>79.63</td>
<td>78.94</td>
<td>78.41</td>
</tr>
<tr>
<td>VILBERT (Li et al., 2019)</td>
<td>79.50</td>
<td>78.85</td>
<td>78.33</td>
</tr>
<tr>
<td>VisualBERT (Li et al., 2019)</td>
<td>79.80</td>
<td>79.00</td>
<td>78.67</td>
</tr>
<tr>
<td>LAMBERT (Li et al., 2019)</td>
<td>72.42</td>
<td>72.54</td>
<td>71.30</td>
</tr>
<tr>
<td>12-in-1 (Lyu et al., 2019)</td>
<td>72.27</td>
<td>72.46</td>
<td>71.14</td>
</tr>
<tr>
<td>UBIR-B (Chen et al., 2020)</td>
<td>73.16</td>
<td>73.44</td>
<td>72.07</td>
</tr>
<tr>
<td>KB-VLP (ours)</td>
<td>73.63</td>
<td>73.89</td>
<td>72.23</td>
</tr>
</tbody>
</table>

**Experiments (KB-VLP)**
- KB-VLP is finetuned on four tasks – VQA, NLVR2, GQA, and OK-VQA.
-KB-VLP outperforms baseline VLP models.
-On OK-VQA, KB-VLP has significant improvements than Oscar.
-The results show:
  - Using knowledge in pretraining improves standard VL tasks.
  - Using knowledge in pretraining enhances commonsense tasks.

**Knowledge Embedding**

- KB-VLP: Knowledge Based Vision and Language Pretraining
- Kezhen Chen*, Qiuyuan Huang+, Yonatan Bisk$, Daniel McDuff+, Jianfeng Gao+
- Northwestern University*, Microsoft Research+, Carnegie Mellon University$

**Experiments**

- Example 1: KB-VLP generates the correct answer, "Shibuya," while Oscar fails.
- Example 2: KB-VLP identifies the correct action, "Shibuya," whereas Oscar does not.
- Example 3: KB-VLP infers the correct answer, "Tokyo," unlike Oscar.

**Three examples from OK-VQA that KB-VLP model generates correct answer, but Oscar does not. Comparing the generated answers from KB-VLP and Oscar indicates that Oscar model is limited to visual detection and KB-VLP has stronger reasoning and understanding ability.**