VTransE: Visual Translation Embedding Network for Visual Relation Detection

Hanwang Zhang\textsuperscript{1}, Zawlin Kyaw\textsuperscript{2}, Shih-Fu Chang\textsuperscript{1}, and Tat-Seng Chua\textsuperscript{2}
\textsuperscript{1}Columbia University \textsuperscript{2}National University of Singapore

**What is Visual Relation Detection?**

**Task** detect \(<\text{subject}, \text{predicate}, \text{object}>\) in an image

**Network Architecture**

Training Loss Function: Translation Embedding

Visual relations offer the direct understanding of object interactions, which provide further semantic information for applications such as captioning and QA

**Dataset**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Image</th>
<th>Object Class</th>
<th>Predicate Class</th>
<th>Unique Relation</th>
<th>Relation/Object</th>
</tr>
</thead>
<tbody>
<tr>
<td>VRD\textsuperscript{1}</td>
<td>5,000</td>
<td>100</td>
<td>70</td>
<td>6,672</td>
<td>24.25</td>
</tr>
<tr>
<td>VG\textsuperscript{2}</td>
<td>99,658</td>
<td>200</td>
<td>100</td>
<td>19,237</td>
<td>57</td>
</tr>
</tbody>
</table>

Visual relations can capture the meaning of "on", "under", etc, while JointBox confuses "below" and "on".

1. VTransE learns relations regardless of objects priors, reducing VRD complexity from $O(N^2R)$ to $O(N+R)$
2. VTransE improves object detection.
3. VTransE is a pure visual model that achieves state-of-the-art.

**Detection and Retrieval Results**

1. Top five predicate predictions. Top row: joint region predicate classifier result. Btm row: VTransE results

2. Faster R-CNN (%mAP) before and after VTransE

3. Interactive Demo

**Codes**

https://github.com/zawlin/cvpr17_vtranse

**https://cvpr.zl.io/**