Knowledge Acquisition for Visual Question Answering via Iterative Querying

Yuke Zhu, Joseph J. Lim, Li Fei-Fei
Computer Science Department, Stanford University

Introduction

• Most of today’s VQA methods make their predictions based on a predefined set of information
• These VQA models have been shown to be “myopic” (tend to fail on novel instances) [1]
• Current state-of-the-art VQA models have indicated that they could benefit from better visually grounded evidence [4]
• We push one step further by enabling a VQA model to ask for and collect “clues” — in particular, visually grounded evidence.

Iterative Training Algorithm

1: procedure
2: Generate random query rollouts $R^{(i)}$
3: Train initial core network $C^{(0)}$ with rollout $R^{(1)}$
4: Generate training samples $S^{(i)}$ for query scoring network with $C^{(0)}$
5: Train initial query scoring network $Q^{(0)}$ with $S^{(i)}$
6: for $i = 1, \ldots, N$ do
7: Generate query rollouts $R^{(i)}$ with query scoring network $Q^{(i-1)}$
8: Finetune core network $C^{(i)}$ from $C^{(i-1)}$ with rollout $R^{(i)}$
9: Generate training samples $S^{(i)}$ for query scoring network from $C^{(i)}$
10: Finetune query scoring network $Q^{(i)}$ from $Q^{(i-1)}$ with $S^{(i)}$
11: end for
12: return $(Q^{(N)}, C^{(N)})$
13: end procedure

Experiment Results

New state-of-the-art on Visual7W
On par with VQA challenge winning model

Knowledge Sources

Visual Genome scene graphs [5]

<table>
<thead>
<tr>
<th>Method</th>
<th>VQA (dev)</th>
<th>VQA (standard)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two-layer LSTM [2]</td>
<td>0.627</td>
<td>0.631</td>
</tr>
<tr>
<td>Co-Attention [6]</td>
<td>0.658</td>
<td>0.661</td>
</tr>
<tr>
<td>MCB + Att. + GloVe [3]</td>
<td>0.691</td>
<td>–</td>
</tr>
<tr>
<td>MCB Ensemble + Genome [3]</td>
<td>0.702</td>
<td>0.701</td>
</tr>
<tr>
<td>MCB [4]</td>
<td>0.659</td>
<td>–</td>
</tr>
<tr>
<td>MCB + query generator</td>
<td>0.691</td>
<td>0.689</td>
</tr>
</tbody>
</table>

References