Contextualized Bilinear Attention Networks

Gi-Cheon Kang\textsuperscript{1}, Seonil Son\textsuperscript{1}, Byoung-Tak Zhang\textsuperscript{1,2}

\textsuperscript{1}Seoul National University, \textsuperscript{2}Surromind Robotics

Abstract. An agent that can see everyday scenes and fluently communicate with people is one of the ambitious goals of artificial intelligence. To achieve that, it is crucial to exploit visually-grounded information and capture subtle nuances from human conversation. To this end, Visual Dialog (VisDial) task has been introduced. In this paper, we propose a new model for visual dialog. Our model employs Bilinear Attention Network (BAN) and Embeddings from Language Models (ELMo) to exploit visually-grounded information and context of dialogs, respectively. Our proposed model outperforms previous state-of-the-art on VisDial v1.0 dataset by a significant margin (5.33% on recall @10)

Keywords: Visual Dialog, Visual QA, Attention mechanisms

1 Introduction

With the recent progress of Visual Question Answering (VQA) \cite{1, 2}, visual dialog \cite{3} task has been introduced as a general version of VQA. Different from VQA, it requires to answer multiple questions in a single image. Accordingly, the visual dialog task has two key challenges which are exploiting visually-grounded information and catching the context of the dialog. To deal with two challenges, we propose a Contextualized Bilinear Attention Networks (CBAN) for visual dialog task. CBAN can be viewed as an extended idea of BAN \cite{2} which was originally proposed in VQA task. Also, we employ newly proposed word embeddings, ELMo \cite{4} to utilize a contextualized word representation.

![Diagram of Contextualized Bilinear Attention Network](image)

Fig. 1. Contextualized Bilinear Attention Network
Table 1. Test-standard score on VisDial v1.0 dataset, measured by mean reciprocal rank (MRR), recall @k and mean rank [3]. ATT indicates a use of attention mechanism.

<table>
<thead>
<tr>
<th>Model</th>
<th>ATT</th>
<th>MRR</th>
<th>R@1</th>
<th>R@5</th>
<th>R@10</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Memory Network</td>
<td>✓</td>
<td>56.90</td>
<td>42.43</td>
<td>74.00</td>
<td>84.35</td>
<td>5.59</td>
</tr>
<tr>
<td>Late Fusion</td>
<td>✓</td>
<td>57.07</td>
<td>42.08</td>
<td>74.83</td>
<td>85.05</td>
<td>5.41</td>
</tr>
<tr>
<td>CBAN (ours)</td>
<td>✓</td>
<td>58.86</td>
<td>42.85</td>
<td>78.70</td>
<td>90.38</td>
<td>4.13</td>
</tr>
</tbody>
</table>

2 Contextualized Bilinear Attention Networks

Figure 1 shows the overview of our model. Our model has two sub-modules which are BAN [2] and contextualized history using ELMo [4]. BAN efficiently extracts visually-grounded representation using low-rank bilinear pooling [2], and history embedding with ELMo has a rich representation of the previous conversation. As a sequence of questions has an interdependent property, history representation plays a key role in catching the context (e.g. co-reference, temporal topic) of the dialog. Finally, the two representations are concatenated and passed to a decoder to respond valid answer. $\phi$ and $N$ denote the number of objects in the image and the number of questions in one dialog, respectively.

Table 1 shows the performance comparison with the other models on VisDial v1.0 test-standard. All scores are measured by the rank of ground-truth answer. Higher is better except for mean rank. [3] introduced Memory Network (MN) and Late Fusion (LF) based architectures. Our model significantly outperforms previous state-of-the-art except for R@1.

3 Conclusion

In this work, we introduce a Contextualized Bilinear Attention Networks (CBAN) and show experimental results. We believe that the CBAN approach can be utilized for practical application, including assistants for blind people.

References