Most existing downstream systems integrate visual features and textual concepts to form visual representations. These representations only contain information from one modality. We evaluate the proposed MIA on two multi-modal datasets: image captioning and visual question answering, is typically represented in two fundamental forms: visual features and textual concepts (see Figure 1).

Limitation & Challenge:

- Most existing downstream systems integrate visual features and the textual concepts in the decoding process, mostly ignoring the inherent alignment between the two modalities.
- The systems have to learn the alignment between each individual visual feature and textual concept.
- These representations only contain individual features, lacking the meaningful combinations and structural relationships among them.

Those problems hinder the system from understanding images efficiently.

Solution:

- We propose the Mutual Iterative Attention (MIA) module to align the visual features and textual concepts in the encoding process. Using textual concepts to query and integrate visual features with attention, we could get image representations centered upon each concept forming meaningful visual feature groups, and vice versa. The representations are refined by applying MIA iteratively.

Our approach based on the Multi-Head Attention (MHA) and Feed-Forward Network (FCN) from Transformer [1].

Mutual Attention

Given visual features I and textual concepts T, the mutual attention is conducted as:

\[ I' = FCN(MHA(I, T)), \quad T' = FCN(MHA(I', T)) \]

i.e., visual features are first integrated according to textual concepts, and then textual concepts are integrated according to integrated visual features.

Mutual Iterative Attention (MIA)

We perform mutual attention iteratively to refine both visual features and textual concepts:

\[ I_k = FCN(MHA(T_{k-1}, I_{k-1})), \quad T_k = FCN(MHA(I_k, T_{k-1})) \]

Semantic-Grounded Image Representations

Since the visual features and the textual concepts are already aligned, we can add them up to get the semantically-grounded image representations:

\[ MIA(I, T) = LayerNorm(N + T') \]

Experiments

- We evaluate the proposed MIA on two multi-modal tasks (image captioning and visual question answering (VQA)).
- We translate the baseline models (w/ MIA) on the MSCOCO image captioning benchmark dataset. B-n, M, R, C and S are short for BLEU-n, METEOR, ROUGE-L, CIDEr and SPICE, respectively.

```
<table>
<thead>
<tr>
<th>Methods</th>
<th>B-1</th>
<th>B-2</th>
<th>B-3</th>
<th>M</th>
<th>R</th>
<th>C</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCN</td>
<td>70.2</td>
<td>70.3</td>
<td>68.8</td>
<td>45.2</td>
<td>39.5</td>
<td>26.7</td>
<td>129.6</td>
</tr>
<tr>
<td>MHA</td>
<td>70.3</td>
<td>70.4</td>
<td>69.0</td>
<td>45.1</td>
<td>39.5</td>
<td>26.7</td>
<td>129.6</td>
</tr>
<tr>
<td>w/ MIA</td>
<td>69.3</td>
<td>69.4</td>
<td>67.9</td>
<td>44.9</td>
<td>39.1</td>
<td>26.5</td>
<td>128.5</td>
</tr>
<tr>
<td>BAN</td>
<td>70.2</td>
<td>70.3</td>
<td>68.8</td>
<td>45.2</td>
<td>39.5</td>
<td>26.7</td>
<td>129.6</td>
</tr>
</tbody>
</table>
```

Table 1. Results of representative systems on the MSCOCO image captioning benchmark dataset. B-n, M, R, C and S are short for BLEU-n, METEOR, ROUGE-L, CIDEr and SPICE, respectively.

- As we can see, the proposed MIA exhibits compelling effectiveness in boosting the baseline systems.

References


Contact Us

- fenglinliu@pku.edu.cn
- liuyuanxin@ie.ac.cn
- renxc@pku.edu.cn
- xiaoxun@pku.edu.cn
- xiaodong.he@jd.com