1. Introduction

The supplementary materials consist of implementation details of neural module methods, ablation studies, and supplementary qualitative results:

1. we present the implementation details of NSM [2] and XNM [4] in Section 2;
2. we present ablation studies about the selection of hyperparameters for NSM [2] and XNM [4] models in Section 3;
3. we demonstrate additional qualitative examples in Section 4.

2. Implementation details of NSM and XNM

One of the key differences between conventional XNM methods and ours lies in the attention mechanism. In the main paper, we have discussed the memory-augmented attention that jointly allocates attention to the visual knowledge and external knowledge with information sharing. In this section, we present additional details about how the attention mechanisms of XNM and NSM models are adapted to work with our method.

Specifically, we concatenate the pair of queries $q_v^t$ and $q_e^t$ into $q_t = [q_v^t, q_e^t]$, which is used as the input to the neural modules. Similarly, the intermediate attention output from the neural modules is also defined as the combination of two attention vectors (i.e., $\alpha_t = [\alpha_v^t, \alpha_e^t]$). The augmented attention (i.e., $\hat{\alpha}_t = [\hat{\alpha}_v^t, \hat{\alpha}_e^t]$) is then used as the input of the next neural modules in the sequence of execution. Under this design, each neural module of NSM and XNM can jointly process a pair of input queries and output a pair of attention vectors each corresponding to one knowledge source. Below we introduce how the two methods are adapted to process the augmented queries and attention.

Adaptation of NSM. With the input query $q_t$ and the augmented attention $\hat{\alpha}_t$, we adopt the attention update mechanism of NSM [2] to compute the attention of the next reasoning step $\alpha_{t+1}$. First, following NSM [2], to determine how attention should be shifted, we compute the relevance scores of each graph node (denoted as $\gamma_a$) and edge (denoted as $\gamma_b$) to the input query. Note that the difference between our implementation and the original NSM method is that we compute the relevance scores on both the visual scene graph and the external knowledge graph. We then compute (1) the influx attention $\alpha_{t+1}^{in}$ of each node as a combination of the input attention $\alpha_t$ in the neighboring nodes, using the edge relevance scores $\gamma_b$ as the combination weights, and (2) the remaining attention $\alpha_{t+1}^{remain}$ from the node relevance scores $\gamma_a$. Each of the attention vector is composed of two softmax-normalized components within the visual knowledge graph and the external knowledge graph independently. Finally, they are linearly combined to generate the output attention $\alpha_{t+1}$.

Adaptation of XNM. We implement the XNM method following its original design [4]. Specifically, Tab. 1 presents the list of XNM neural modules and their implementation details. These modules are grouped into three categories: attention, logic, and output. Attention modules take previous attention $\alpha_{t-1}$, node features $h_t$ or query $q_t$ as the input and compute the output attention $\alpha_{t+1}$. Logic modules shift attention by conducting logical operations on the current attention $\alpha_t$ or two previously computed attention vectors $\alpha_{t-1}^{in}$ and $\alpha_{t-1}^{remain}$. The output modules predict the answers by performing feature comparison, measuring the amount of attention, or applying the query to the attended features.

3. Ablation analysis of hyperparameters

We conduct ablation analyses to evaluate the effects of different hyperparameters. The model performance is evaluated on the OK-VQA [3]. Overall, the optimal hyperparameter settings for XNM are the same as those for the NSM...
model, which is $\eta^v = 0.6$, $\eta^c = 0.8$, and $L_d = 3$. We report detailed discussions about these hyperparameters below.

**Hyperparameters $\eta^v$ and $\eta^c$.** The hyperparameters $\eta^v$ and $\eta^c$ are the weights that determine how much the agents should be rewarded for selecting knowledge-related queries. Higher values enforce the generated queries of both agents being more different from the question but more similar to the visual or external sources, respectively. Tab. 2 and Tab. 3 present the results of our method with different combinations of $\eta^v$ and $\eta^c$. Our method performs the best when $\eta^v = 0.6$ and $\eta^c = 0.8$. These weights are high enough for the two sets of augmented queries to complement each other but also remain relevant to the question.

<table>
<thead>
<tr>
<th>$\eta^v$</th>
<th>$\eta^c$</th>
<th>0</th>
<th>0.2</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td>20.59</td>
<td>20.78</td>
<td>20.84</td>
<td>21.02</td>
<td>21.46</td>
<td>21.31</td>
</tr>
<tr>
<td>0.2</td>
<td></td>
<td>21.38</td>
<td>21.59</td>
<td>21.89</td>
<td>23.05</td>
<td>23.52</td>
<td>23.36</td>
</tr>
<tr>
<td>0.4</td>
<td></td>
<td>21.46</td>
<td>22.31</td>
<td>22.90</td>
<td>23.64</td>
<td>23.25</td>
<td>23.12</td>
</tr>
<tr>
<td>0.6</td>
<td></td>
<td>21.93</td>
<td>24.64</td>
<td>25.22</td>
<td>25.94</td>
<td>26.52</td>
<td>26.17</td>
</tr>
<tr>
<td>0.8</td>
<td></td>
<td>22.37</td>
<td>23.39</td>
<td>24.25</td>
<td>25.73</td>
<td>25.41</td>
<td>25.55</td>
</tr>
<tr>
<td>1.0</td>
<td></td>
<td>22.24</td>
<td>22.76</td>
<td>23.41</td>
<td>23.20</td>
<td>22.17</td>
<td>22.39</td>
</tr>
</tbody>
</table>

Table 2. Hyperparameter selection of $\eta^v$ and $\eta^c$ on the OK-VQA dataset, based on the XNM model. Best results are highlighted in bold.

**Hyperparameter $L_d$.** The maximum distance $L_d$ is a key hyperparameter to control the inclusion of visual and external concepts into the dictionaries. Though $L_d$ needs to be sufficiently large for the dictionaries to include the most relevant concepts as query candidates, a large $L_d$ may also result in the inclusion of less relevant concepts. Tab. 4 and Tab. 5 demonstrate the model performance with respect to different choices of $L_d$.

4. **Supplementary qualitative results**

We demonstrate supplementary qualitative examples in Fig. 1 to show the effectiveness and generalizability of our method. As shown in Fig. 1a–e, the knowledge-augmented queries guide attention to concepts that are more relevant to the answers (e.g., vegetable and topping in Fig. 1a, recreation and dog in Fig. 1b, plant and tiny in Fig. 1c, carriage and transport in Fig. 1d, dress in Fig. 1e). With these queries, the attention of the model can be better allocated to the corresponding answers. In contrast, the state-of-the-art AN [3] and KI-Net [5] are more vulnerable to external knowledge biases (e.g., salad is more probable than pizza to be present together with vegetable). These examples show that our method can better locate the relevant knowledge for answer prediction.

Fig. 1f demonstrates a failure case of our model. It suggests that when the reasoning considers logical operations (e.g., without), allocating attention solely based on semantic similarities can be problematic and may lead to incorrect answers (e.g., man instead of camel). Such issue is pervasive among many state-of-the-art VQA models [1, 2], as well as the compared AN [3] and KI-Net [5] methods.
Figure 1. Additional qualitative examples of our method. Each example shows the input image, question, ground-truth (GT) answer and model predictions, base queries (B-Q) and the queries augmented with visual knowledge (V-Q) and external knowledge (E-Q), followed by the attended visual and external knowledge. Highlighted knowledge (blue) indicates the FVQA supporting fact of the question.

Therefore, future extensions of our method may consider explicitly representing the logical operations in the neural modules and knowledge graphs to tackle the issue.
References


