Query and Attention Augmentation for Knowledge-Based Explainable Reasoning (Supplementary Materials)

Yifeng Zhang, Ming Jiang, Qi Zhao
University of Minnesota
{zhan6987, mjiang}@umn.edu, qzhao@cs.umn.edu

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_t^v and q_t^e into $q_t = [q_t^v, q_t^e]$, 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.,* $\alpha_t = [\alpha_t^v, \alpha_t^e]$). The augmented attention (*i.e.,* $\hat{\alpha}_t = [\hat{\alpha}_t^v, \hat{\alpha}_t^e]$) 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 α_{t+1}. First, following NSM [2], to determine how attention should be shifted, we compute the relevance scores of each graph node (denoted as γ_a) and edge (denoted as γ_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 α_{t+1}^{in} of each node as a combination of the input attention $\hat{\alpha}_t$ in the neighboring nodes, using the edge relevance scores γ_b as the combination weights, and (2) the remaining attention α_{t+1}^{remain} from the node relevance scores γ_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 α_{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 α_t, node features h_t or query q_t as the input and compute the output attention α_{t+1}. In our work, to jointly reason about the visual knowledge and external knowledge, we represent h_t as the concatenation of node features $h_t = [h_t^v, h_t^e]$ from both sources. Logic modules shift attention by conducting logical operations on the current attention α_t or two previously computed attention vectors α_t^1 and α_t^2. 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

Modules	Category	Operation
Attend	Attention	$\alpha_t = \text{softmax}(\text{MLP}(h_t, q_t))$
Relate	Attention	$\alpha_t, h_t, q_t \rightarrow \alpha_{t+1}$
Or	Logic	$\alpha_{t+1} = \min(\alpha_t^1, \alpha_t^2)$
And	Logic	$\alpha_{t+1} = \max(\alpha_t^1, \alpha_t^2)$
Not	Logic	$\alpha_{t+1} = 1 - \alpha_t$
Compare	Output	$h_{out} = \text{MLP}(h_t^1 - h_t^2)$
Exist	Output	$h_{out} = \text{MLP}(\text{sum}(\alpha_t))$
Describe	Output	$h_{out} = \text{softmax}(\text{MLP}(q_t))\boldsymbol{W}(\alpha_t \circ h_t)$

Table 1. An overview of XNM neural modules. $\text{MLP}(\cdot)$ indicates a multi-layer perceptron consisting of several fully-connected and ReLU layers, and $\boldsymbol{W}$ is a matrix of learnable weights. The parameters α_t, h_t, and q_t indicate attention, features, and query. α_t^1, α_t^2 are two previously computed attention vectors that will be processed with logical operations.

model, which is $\eta^v = 0.6$, $\eta^e = 0.8$, and $L_d = 3$. We report detailed discussions about these hyperparameters below.

Hyperparameters η^v and η^e. The hyperparameters η^v and η^e 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 η^v and η^e. Our method performs the best when $\eta^v = 0.6$ and $\eta^e = 0.8$. These weights are high enough for the two sets of augmented queries to complement each other but also remain relevant to the question.

η^v \ η^e	0	0.2	0.4	0.6	0.8	1.0
0	20.59	20.78	20.84	21.02	21.46	21.31
0.2	21.38	21.59	21.89	23.05	23.52	23.36
0.4	21.46	22.31	22.90	23.64	23.25	23.12
0.6	21.93	24.64	25.22	25.94	**26.52**	26.17
0.8	22.37	23.39	24.25	25.73	25.41	25.55
1.0	22.24	22.76	23.41	23.20	22.17	22.39

Table 2. Hyperparameter selection of η^v and η^e 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 overlarge 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.

η^v \ η^e	0	0.2	0.4	0.6	0.8	1.0
0	21.97	22.57	23.09	23.59	23.72	23.79
0.2	23.92	23.74	24.15	25.32	25.87	25.49
0.4	22.49	22.57	24.19	24.78	26.14	26.62
0.6	22.26	24.73	24.78	28.64	**29.24**	28.31
0.8	22.41	24.50	24.92	26.28	27.73	27.48
1.0	22.35	23.92	25.11	26.86	27.51	27.89

Table 3. Hyperparameter selection of η^v and η^e on the OK-VQA dataset, based on the NSM model. Best results are highlighted in bold.

L_d	1	2	3	4	5
Accuracy	24.89	25.64	**26.52**	26.27	26.21

Table 4. Hyperparameter selection of L_d on the OK-VQA dataset, based on the XNM model. Best results are highlighted in bold.

L_d	1	2	3	4	5
Accuracy	27.29	28.52	**29.24**	28.67	27.51

Table 5. Hyperparameter selection of L_d on the OK-VQA dataset, based on the NSM model. Best results are highlighted in bold.

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.

	(a)	(b)	(c)
Images			
Questions	Which object in this image can be created with dough, tomato sauce and mozarrella cheese?	Which object in this image is used for play with your pet?	What object in this image has shallow roots?
Answers	**NSM+Ours:** pizza **NSM+KI-Net:** cookie **NSM+AN:** salad **GT:** pizza	**NSM+Ours:** frisbee **NSM+KI-Net:** man **NSM+AN:** man **GT:** frisbee	**NSM+Ours:** grass **NSM+KI-Net:** tree **NSM+AN:** tree **GT:** grass
Queries	**B-Q:** 1.object, 2.dough, 3.tomato, 4.cheese **V-Q:** 1.cookie, 2.dough, 3.tomato, 4.cheese **E-Q:** 1.food, 2.bake, 3.vegetable, 4.topping	**B-Q:** 1.object, 2.play, 3.pet **V-Q:** 1.man, 2.play, 3.pet **E-Q:** 1.frisbee, 2.recreation, 3.dog	**B-Q:** 1.object,, 2.shallow, 2.root **V-Q:** 1.tree, 2.shallow, 3. root **E-Q:** 1.plant, 2.short, 3. branch
Visual	cookie-rightOf-beer	man-locationOf-frisbee	tree-rightOf-tower tree-leftOf-tower
External	pizza-typeOf-bake pizza-relationOf-cheese pizza-relationOf-tomoto pizza-relationOf-dough cheese-typeOf-topping pizza-relationOf-topping	dog-typeOf-pet frisbee-relationOf-dog firsbee-relationOf-recreation man-capableOf-play	grass-relationOf-root grass-typeOf-plant tree-typeOf-plant grass-relationOf-short

	(d)	(e)	(f)
Images			
Questions	Which thing in the image can be used for carrying goods?	What is this cake for?	What object in this image can work for days without water?
Answers	**NSM+Ours:** horse **NSM+KI-Net:** woman **NSM+AN:** woman **GT:** horse	**NSM+Ours:** weddings **NSM+KI-Net:** birthday **NSM+AN:** birthday **GT:** weddings	**NSM+Ours:** man **NSM+KI-Net:** man **NSM+AN:** man **GT:** camel
Queries	**B-Q:** 1.carry, 2.goods **V-Q:** 1.woman, 2.drive **E-Q:** 1.carriage, 2.transport	**B-Q:** 1.cake **V-Q:** 1.cake **E-Q:** 1.dress	**B-Q:** 1.object, 2.water **V-Q:** 1.man, 2.water **E-Q:** 1.camel, 2.grass
Visual	horse-leftOf-baby horse-leftOf-woman	woman-wear-dress cake-leftOf-woman cake-leftOf-man	man-topOf-grass camel-topOf-grass
External	horse-relationOf-carriage carriage-capableOf-carry goods-relationOf-transport horse-capableOf-transport	cake-relationOf-birthday cake-relationOf-wedding dress-relationOf-wedding	man-relationOf-water camel-relationOf-water water-relationOf-desert

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

[1] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. Mutan: Multimodal tucker fusion for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2612–2620, 2017.

[2] Drew A Hudson and Christopher D Manning. Learning by abstraction: The neural state machine. *arXiv preprint arXiv:1907.03950*, 2019.

[3] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3195–3204, 2019.

[4] Jiaxin Shi, Hanwang Zhang, and Juanzi Li. Explainable and explicit visual reasoning over scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8376–8384, 2019.

[5] Yifeng Zhang, Ming Jiang, and Qi Zhao. Explicit knowledge incorporation for visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2021.