Explicit Knowledge Incorporation for Visual Reasoning (Supplementary Materials)

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

1. Introduction

The supplementary material consists of ablation studies and attention analysis of the proposed method:

1. We report ablation studies of different hyperparameters used in the proposed method (Section 2).

2. We analyze the significance of external knowledge by comparing the attention distribution over different types of entity nodes in the enriched scene graph (Section 3).

2. Ablation Studies of Hyperparameters

In this section, we present ablation studies on the hyperparameter selection for the proposed method.

Hyperparameters for Knowledge Inference Network (KI-Net). The KI-net depends on two hyperparameters: K_p controls the number of the highest relevance scores in the relevance matrix $\boldsymbol{M}$ to consider in the semantic refinement of the scene graph, and ϵ_p is the threshold that determines whether to discard a candidate entity with low relevance scores. In particular, a smaller K_p or a larger ϵ_p indicate a more strict standard to keep the most relevant entities, thus there is a risk to exclude relevant entities; on the other hand, a larger K_p or a smaller ϵ_p indicate a higher tolerance to low-relevance entities, thus less relevant nodes may be included. We evaluate how different combinations of K_p and ϵ_p collaboratively decide the relevance between the incorporated external knowledge and the originally observed scene graph. With the other hyperparameters fixed to their optimal values, we optimize the combination of K_p and ϵ_p by conducting a grid search on the VQAv2 validation set. Tab. 1 presents the results of this ablation study. As can be seen, for both scene graph recall and answer accuracy, our KI-Net performs the best when $K_p = 3$ and $\epsilon_p = 0.8$, by maintaining a good trade-off between semantic richness and relevance.

K_p	ϵ_p	Accuracy	mR@50	mR@100	R@50	R@100
3	0.8	**67.32**	**6.2**	**7.3**	**25.7**	**30.6**
1	0.8	64.47	5.7	6.4	25.1	27.8
2	0.8	66.83	6.1	6.9	25.5	30.4
4	0.8	67.19	5.9	7.1	25.4	30.6
3	0.6	67.14	5.9	7.2	25.4	30.2
3	0.7	67.21	5.9	7.1	25.6	30.5
3	0.9	67.27	6.1	7.2	25.6	30.4

Table 1. Experimental results of our method with different settings of the hyperparameters used in the KI-Net.

T	d_h	L	Accuracy
2	300	3	62.54
3	300	3	64.93
4	300	3	**67.32**
5	300	3	66.77
6	300	3	66.39
4	100	3	64.14
4	200	3	66.80
4	300	3	**67.32**
4	500	3	66.95
4	300	1	64.57
4	300	2	66.41
4	300	3	**67.32**
4	300	4	65.94

Table 2. Experimental results of our method with different settings of the hyperparameters used in the neural modules.

Hyperparameters for Neural Modules. We evaluate how the number of reasoning steps T, the feature dimension d_h, and the max length of path L collaboratively impact the reasoning process. With the other hyperparameters fixed to their optimal values, the combination of T, $d_h$and L by conducting a grid search on the VQAv2 validation set. Tab. 2 shows that our model performs the best with the combination $T = 4$, $d_h = 300$, $L = 3$. Degenerated results with the

[*] These authors contributed equally.

Method	Question	Initial Entities		Incorporated Entities	
		Before	After	Before	After
G-Relate	Yes/No	0.29	0.28	0.34	0.37
	Number	0.28	0.28	0.31	0.32
	Other	0.29	0.28	0.31	0.33
	Overall	0.29	0.28	0.32	0.34
XNM [2]	Yes/No	0.30	0.29	0.32	0.33
	Number	0.29	0.29	0.31	0.31
	Other	0.29	0.29	0.31	0.32
	Overall	0.29	0.29	0.31	0.32

Table 3. Attention distribution over the enriched scene graph based on KI-Net and different Relate modules: G-Relate (high-order) *vs.* XNM (first-order). Initial Entities are from the original scene graph and Incorporated Entities are incorporated from the external knowledge graph using the KI-Net. Before and After suggest the attention weights computed before and after executing the Relate module, respectively.

other combinations indicate that while a smaller T is insufficient to reason over the rich semantics, a larger T may introduce extra complexities to the model, increasing the difficulty of parameter optimization. Similarly, a lower feature dimension d_h is insufficient to fully represent the necessary information while larger d_h may store redundant information, which is inefficient. Lastly, the number of paths L used in G-Relate cannot be too small or too large, since a small L neglects the high-order node-wise relation, and a large L is also unnecessary since the composite relevance along long paths becomes weaker.

3. Analysis of Attention Distribution

Where a visual reasoning method attends and how it shifts its attention during the execution of neural modules can explain the correctness of its reasoning process. To demonstrate the roles of the KI-Net and the G-Relate in directing attention for visual reasoning, we analyze the attention distribution over the enriched scene graph. In particular, we analyze and compare the following three aspects among different questions: (1) the average attention weights that suggest the need for focused attention in answering the questions, (2) the difference in attention weights between initial entities and incorporated entities that show the usefulness of the incorporated knowledge, and (3) the difference in attention weights before and after executing the G-Relate that indicate the significance of high-order inference. We conduct comparative studies on a selected subset of questions from the VQAv2 validation dataset [1]. The questions are selected following two criteria: first, the question contains an object that is undetected from the image, and second, the generated program contains a G-Relate module. This subset allows us to focus our analysis on the significance of the proposed KI-Net and G-Relate.

In Tab. 3, we compare the average attention weights between the initial entities and incorporated entities, both before and after executing the high-order G-Relate module or the first-order Relate module (XNM), and across different question types.

Attention Analysis for KI-Net. First, we compare the attention weights between the two sets of entity nodes: all initial entities detected from the scene (*i.e.,* in the original scene graph), and all new entities from the external knowledge graph. As shown in Tab. 3, the attention weights of the incorporated entities are higher than those of the initial entities. Although neither the initial scene graph generation nor the knowledge incorporation is specifically based on the question, we can still observe significant differences between their average attention weights. The higher weights of the incorporated entities demonstrate the usefulness of these entities, thus the usefulness of the external knowledge to complement the visual information from the initial scene graph.

Attention Analysis for G-Relate. Next, to analyze how the proposed G-Relate module transfers attention through high-order relations, we compare the attention weights before and after executing the G-Relate module. As shown in Tab. 3 (top panel), the attention weights of the initial entities are decreased after the G-Relate operation, but those of the incorporated entities are increased. This observation suggests that the G-Relate module can transfer attention from the initially attended entities to the incorporated entities based on the external knowledge graph. Differently, when we replace the proposed G-Relate with a first-order Relate module (implemented following the XNM model [2]), the shift of attention weights becomes less significant. As shown in Tab. 3 (bottom panel), after executing the first-order Relate module, the overall attention weights on the initial entities remain the same, while the attention weights on the incorporated entities only have a mild improvement, not as significant as we observe during the experiment with G-Relate. This comparison confirms the finding that transferring attention along high-order relations is important for the neural modules to access external knowledge and perform effective reasoning.

Attention Analysis for Different Question Types. To evaluate the generalizability of our model, we look into the attention distribution across different question types (Yes/No, Number, and Other). As shown in Tab. 3, the aforementioned observations apply to all three question types. We further observe that among the various question types, Yes/No questions have both the highest average attention weights of the incorporated entities, and the biggest attention difference before and after executing the G-Relate. The more significant role of KI-Net and G-Relate in Yes/No questions is likely because these questions depend heavily on the focused attention to the correct entities, which may be originally missing in the initial scene graph, and are only made accessible with KI-Net and G-Module. Differently, for Number and Other questions, most of them require broader attention to be directed to multiple entities (*e.g.,* for counting or comparing objects), thus the attention weights on the incorporated entities are lower, and the attention differences between initial entities and incorporated entities are smaller. However, we still observe that the high-order G-Relate transfers more attention to the incorporated entities than the first-order Relate modules.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *Int. Conf. Comput. Vis.*, 2015.

[2] Jiaxin Shi, Hanwang Zhang, and Juanzi Li. Explainable and explicit visual reasoning over scene graphs. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8376–8384, 2019.