Toward Multi-Granularity Decision-Making: Explicit Visual Reasoning with Hierarchical Knowledge (Supplementary Materials)

Yifeng Zhang, Shi Chen, Qi Zhao
University of Minnesota
{zhan6987, chen4595}@umn.edu, qzhao@cs.umn.edu

The supplementary materials provide additional experimental results and implementation details of our proposed work:

1. Section 1 provides ablation analyses of the hyperparameters used in the proposed knowledge representation and reasoning model.
2. Section 2 presents details of the incorporation of knowledge graphs (*i.e.*, SKG, UKG, HCG) for non-NMN models (*i.e.*, MCAN [5], UnifER [3]).

1. Ablation Study on Hyperparameters

To enable a comprehensive understanding of our method, we perform ablation experiments to evaluate the effects of different hyperparameters for multi-layered hierarchical structure (HCG) and the hierarchical concept neural module networks (HCNMN). The model performance is evaluated on the GQA [4] dataset.

r_v \ r_c	0.2	0.4	0.6	0.8	1.0
0.2	63.83	64.04	64.27	64.58	64.27
0.4	64.02	64.26	64.59	64.72	64.76
0.6	64.07	**64.97**	64.34	64.86	64.79
0.8	64.17	64.23	64.59	64.65	64.74
1.0	64.28	64.31	64.44	64.42	64.30

Table 1. Hyperparameter selections of r_v and r_c on the GQA validation set, based on the HCNMN model. The best result is highlighted in bold.

Hyperparameters for Concept Properties. The hyperparameters r_v and r_c are the weights that determine how concept properties are selected. Higher values enforce the extracted concepts being more unique (*i.e.,* r_c) or relevant (*i.e.,* r_v) to the dataset, respectively. Table 1 presents the results of our method with different combinations of r_v and r_c. Our method performs the best when $r_v = 0.6$ and $r_c = 0.4$. The results indicate that both uniqueness and relevance are important factors for extracting useful knowledge, and it is essential to maintain a reasonable balance between them.

t	0.1	0.3	0.5	0.7	0.9
GQA Val	64.26	**64.97**	64.87	64.25	63.98

Table 2. Hyperparameter selections of t on the GQA validation set, based on the HCNMN model. The best result is highlighted in bold.

Hyperparameters for Inter-Layer Information Decay Rate. The hyperparameter t controls the significance of the information that is propagated downwards across different granularity layers. A higher value enables models to emphasize more on the knowledge of more general concepts throughout the inter-layer propagation. Table 2 presents the results of our method with different t. Our method performs the best when $t = 0.3$, indicating the significance of taking into account the granularity of knowledge for decision-making. Additionally, it is also important to maintain a reasonable balance among knowledge of different granularity for visual reasoning.

k	1	2	3	4	5
GQA Val	62.89	64.14	**64.97**	64.39	63.78

Table 3. Hyperparameter selections of k on the GQA validation set, based on the HCNMN model. The best result is highlighted in bold.

Hyperparameters for Graph Layers. The hyperparameter k is the variable that controls the number of layers in our proposed hierarchical concept graph (HCG), and is relevant to the complexity of the knowledge representation. Table 3 presents the results of our method with different k. Our method performs the best result when $k = 3$, and thus applies this setting throughout our experiments in the main paper.

2. Implementation details of incorporating knowledge graphs for MCAN and UnifER

Several methods compared in the main paper, *i.e.*, MCAN [5] and UnifER [3], leverage attention mechanisms to combine implicit embeddings from knowledge sources and visual-linguistic inputs. On the other hand, our method emphasizes multi-granularity knowledge, and performs structured reasoning over a graph representation. Due to the discrepancies in knowledge representations, it is difficult to directly incorporate multi-granularity knowledge encoded in the multi-layered graph structure for implicit methods (*i.e.*, MCAN, UnifER) and fairly evaluate the roles of multi-granularity knowledge. For this, we make adjustments for the feature retrieval/encoding step of these methods in our study (see the main paper) by preprocessing graph-based knowledge representations (*i.e.*, SKG, UKG, HCG) into structure-aware embeddings.

Specifically, we leverage node2vec [2], a random walk-based approach, to perform the transformation from graph to embeddings through random walk generation and embedding creation. First, we randomly traverse the graph to create multiple "walks" (*i.e.*, paths consisting of multiple connected nodes) that encode the topology of the graph. In order to take into account the heterogeneity of HCG (*i.e.*, different types of edges), we generate two sets of walks that traverse through inter-layer and intra-layer edges, respectively. They implicitly encode the intra-layer and inter-layer relationships among concepts in different granularity. Next, those sequential "walks" are fed into an encoder-decoder framework [1] to create structure-aware knowledge embeddings for MCAN and UnifER.

References

[1] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014. 2

[2] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864, 2016. 2

[3] Yangyang Guo, Liqiang Nie, Yongkang Wong, Yibing Liu, Zhiyong Cheng, and Mohan Kankanhalli. A unified end-to-end retriever-reader framework for knowledge-based vqa. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2061–2069, 2022. 1, 2

[4] Drew A Hudson and Christopher D Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6700–6709, 2019. 1

[5] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6281–6290, 2019. 1, 2