MuKEA: Multimodal Knowledge Extraction and Accumulation for Knowledge-based Visual Question Answering

Supplementary Material

Yang Ding[1,2], **Jing Yu**[1,2*], **Bang Liu**[3,4†], **Yue Hu**[1,2], **Mingxin Cui**[1,2], **Qi Wu**[5]
[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
[3]Université de Montréal, Canada [4]Mila - Quebec AI Institute, Canada
[5]University of Adelaide, Australia
{dingyang, yujing02, huyue, cuimingxin}@iie.ac.cn, bang.liu@umontreal.ca
qi.wu01@adelaide.edu.au

Abstract

We provide additional materials to supplement our main submissions. In Section A, we introduce explicit multimodal knowledge construction, knowledge graph characteristics, application scenarios in detail, and provide extracted multimodal knowledge embeddings as off-the-shelf knowledge features to serve knowledge-based downstream tasks. Based on the knowledge graph constructed above, in Section B and C, we respectively introduce how MuKEA performs multimodal knowledge accumulation and complex reasoning. Then we compare the model size of MuKEA with pre-trained models and analyse the influence of multimodal knowledge base size on inference time respectively in Section D and E, which proves that the inference time is not affected much when varying the knowledge size. In Section F, we study the effect of hyper-parameters in model ensemble corresponding to the knowledge complementary experiments. At last, in Section H, we introduce the implementation details about training.

A. Multimodal Knowledge Construction and Related Applications

In Figure 1, we show a 1-hop sub-graph of our accumulated knowledge triplets centered in top-3 frequent answers in the OK-VQA train set. To construct multimodal knowledge graph, firstly we store the extracted knowledge triplet (h, r, t) from training data, where h is the visual region in the image focused by the question, t is the ground-truth answer, and r is the embedding of implicit relation between h and t. We only display the image corresponding to the h and t to explicitly show the structure of the accumulated multimodal knowledge graph. Then we merge all the tail entities with the same answer and merge all the head entities with the same image, while preserving object regions in images as shown in the example. After the pre-training and fine-tuning stages, we accumulate 218,135 multimodal knowledge triplets for knowledge graph construction.

We summarize the characteristics of our proposed multimodal knowledge graph as follows:

- MuKEA extracts different instantiated knowledge for the same image based on different objects in the image. As shown in the left zoom-in example in Figure 1, the same image connects both 'cell phone' and 'travel' since different objects in the same image related to different knowledge. On the contrary, existing multimodal knowledge graph [22] is a complement of general knowledge graphs with entity-referred images.
- The same concept is associated with different visual knowledge in different scenes. As shown in the right zoom-in example in Figure 1, 'airport' can correspond to different scenarios, such as airport hall or suitcases in airport.
- Compared to existing knowledge graphs with predefined types of relations, relation in our proposed multimodal knowledge graph is extensible and supports retrieval as well.
- By correlating relevant knowledge, the knowledge graph is capable of supporting complex reasoning. In Section C we provide a detailed analysis.

Furthermore, We propose the following potential application scenarios for using our multimodal knowledge graph:

*Corresponding author.
†Canada CIFAR AI Chair.

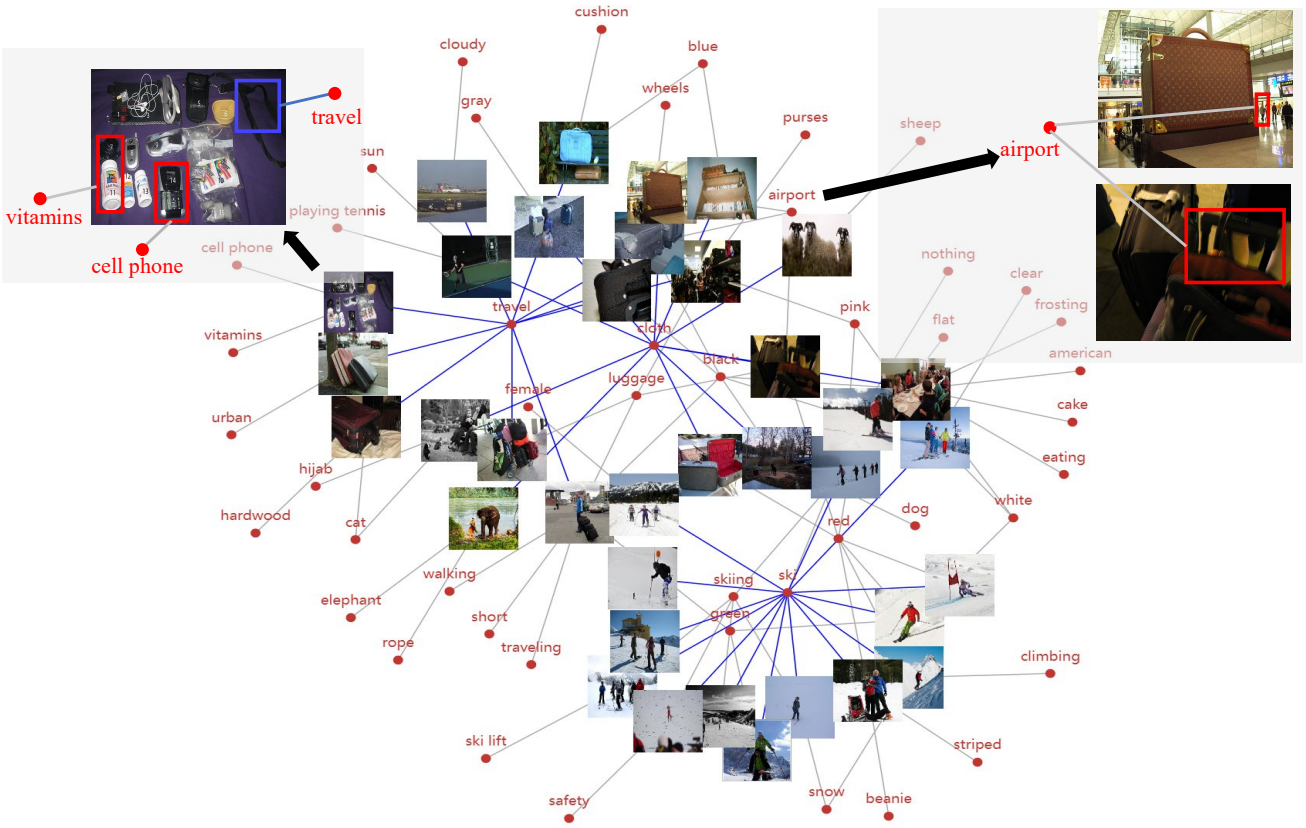

Figure 1. Visualization of the accumulated knowledge graph. Gray lines and blue lines means knowledge accumulated in the VQA 2.0 and OK-VQA respectively. We show extra zoom-in examples for demonstration.

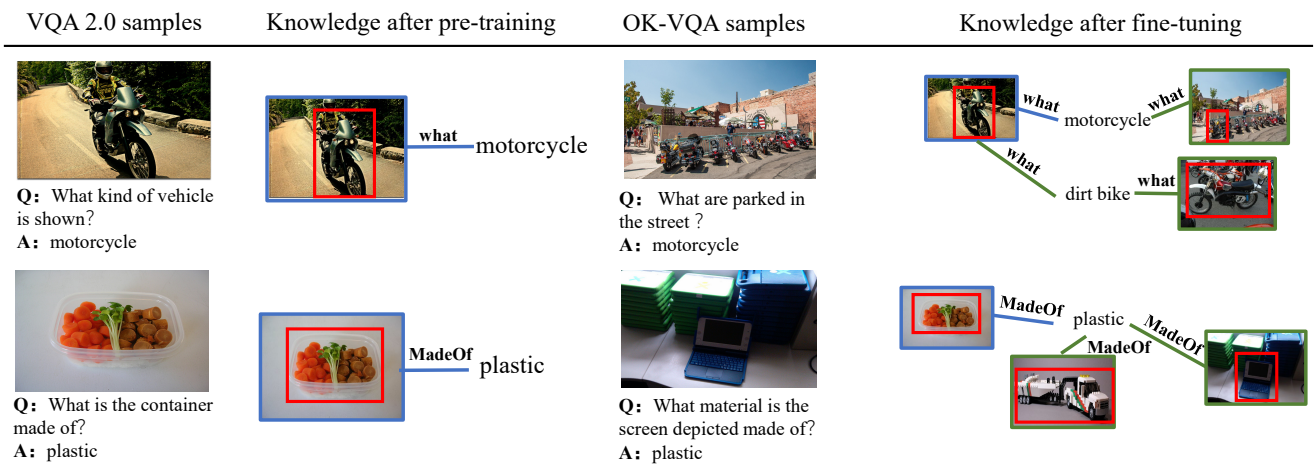

Figure 2. Knowledge accumulation in the pre-training stage and fine-tuning stages.

- Model-based knowledge search. MuKEA is capable of retrieving relevant knowledge for multimodal input.
- Knowledge-based vision-language tasks, such as image caption, referring expression comprehension, vision-language navigation etc.
- Explainable deep learning, especially in the legal, medical fields.

The checkpoint of MuKEA, extracted multimodal

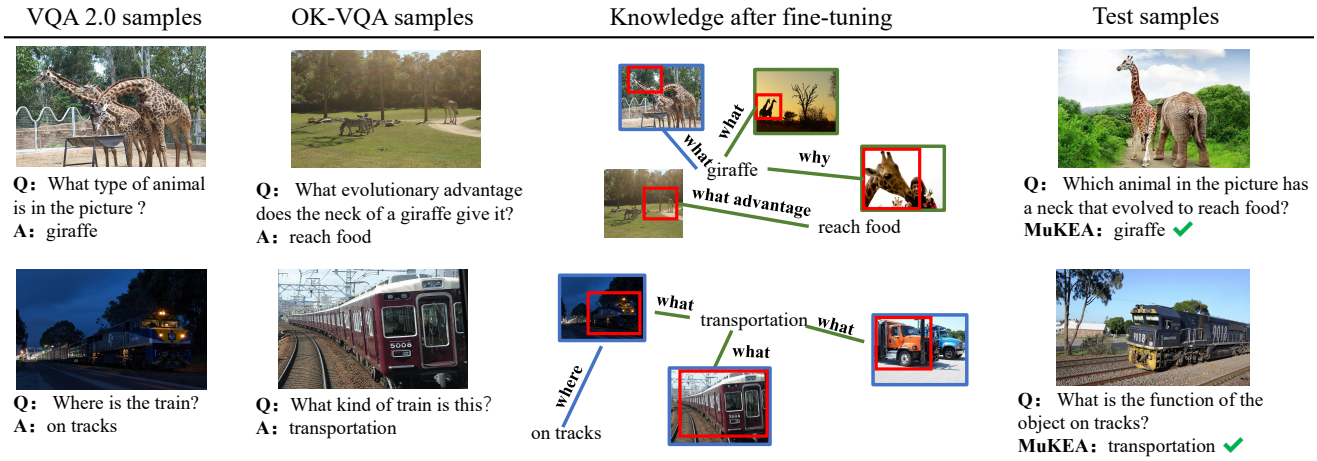

Figure 3. Testing samples based on manually constructed questions in zero-shot setting.

knowledge graph, and off-the-shelf knowledge embeddings are available at https://github.com/AndersonStra/MuKEA

B. Analysis of Progress Knowledge Accumulation

From case study in Figure 2, we illustrate how the basic visual knowledge in VQA 2.0 helps to learn more complex knowledge in OK-VQA: (1) In the first row, benefiting from the question about the appearance of motorcycle, MuKEA is capable to correlate the visual content of motorcycles with the answer in a multi-object scenario. (2) In the second row, benefiting from the prior knowledge of visual content with plastic materials, MuKEA has the advantage of focusing on the key region and obtaining more generalized representation for objects made of plastic.

C. Zero-shot Analysis of Accumulated Multimodal Knowledge

In Figure 3, we show that our model is capable to combine different accumulated knowledge to answer complex questions in the zero-shot setting. (1) In the test sample of the first row, we correlate 'giraffe' with 'evolution' through the manually constructed question. (2) In the test sample of the second row, we construct the question to correlate 'track' and 'transportation'. MuKEA performs correct prediction on both questions, which indicates that the accumulated multimodal knowledge can be applied to complex reasoning tasks in similar way as existing knowledge graphs.

D. Model Size Analysis

In Table 1, we compare the model size of MuKEA with pre-trained models [11, 12, 19, 20]. For MuKEA, we set the knowledge base size to the accumulated knowledge from VQA 2.0 [6] and OK-VQA [13]. The model size increases as the multimodal knowledge base size increases. Compared to ViLBERT, our model size only increases by 8.36% with the performance boost by 11.24%. The model size of KRISP is larger than ours by 86.78%, but its performance is inferior to ours by 3.69%. It indicates that our improvement is not from more parameters, but from the model structure.

Model	Parameter
VL-BERT [19]	138.4M
ViLBERT [11]	218.9M
LXMERT [20]	183.5M
KRISP [12]	443.04M
MuKEA	237.2M

Table 1. Comparison of model size.

E. Efficiency Analysis

To verify that MuKEA strikes a good balance between efficiency and effectiveness , we compare the inference time and ranking time separately based on different scale of multimodal knowledge base. We test on OK-VQA dataset [13], which contains 5,046 samples for testing. Knowledge scale means the number of accumulated multimodal knowledge triplets. Inference time means the time spent on predicting over the entire test set. Ranking time means the time it takes to calculate the similarity with all tail entities in the knowledge base and rank the candidate tail entity.

Table 2 shows that although the ranking time is positively correlated with the size of knowledge base. It is relatively faster in the total inference time (less than 0.01%, as shown in the column of 'Ranking/Inference'). Since the GPU uses

Knowledge Scale	Inference Time(s)	Ranking Time(s)	Ranking/Inference
1000	65.1431	0.0026	0.0040%
10000	65.1459	0.0054	0.0083%
100000	67.3542	0.0064	0.0095%

Table 2. Inference time and ranking time comparison based on different scale of knowledge base on OK-VQA.

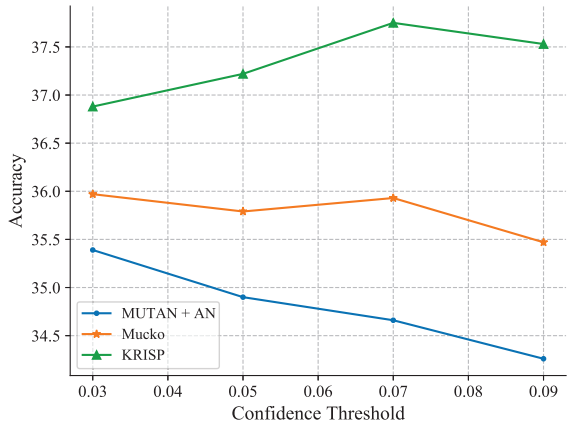

Figure 4. Quantitative study of the confidence threshold τ_t.

threads to process matrix multiplication in parallel, ranking time is not linearly related to the size of knowledge base. Our model still has good efficiency based on large-scale multimodal knowledge base.

F. Effect of Ensemble Hyper-parameters

To verify the robustness of our ensemble model, we report the results of different threshold m for model ensemble in Figure 4. Although we propose a simple method based on confidence to perform model ensemble, we can find that the performance of ensemble model remains stable as the hyper-parameters change in a reasonable range. The threshold m is set in the range of 0.03 and 0.09, and the performance of the ensemble models varies in the range of 34.24% to 35.39%, 35.49% to 35.97%, 36.88% to 37.79% respectively. How to effectively combine MuKEA with knowledge bases will be the future work.

G. Analysis on VQA 2.0

To prove the generalization ability of our method, we compare our model with state-of-the-art models on the VQA 2.0 dataset [6], which requires models to understand the visible content instead of incorporating external knowledge. All questions in VQA 2.0 are divided into three categories: *Yes/No*, *Number*, and *Other*. Since our model is pre-trained on *Other* type questions for accumulating basic multimodal knowledge, we only keep *Other* type questions for comparison.

Method	test-dev	test-std
	Other	Other
MLB [9]	56.34	-
MUTAN [3]	56.50	-
DCN [14]	57.44	56.83
DA-NTN [2]	57.92	-
Counting [25]	58.97	
BLOCK [4]	58.51	58.79
UpDn [1]	56.05	56.26
CGN [15]	-	56.22
CRA-Net [17]	59.08	59.42
MRA-Net [16]	59.46	59.86
SceneGCN [24]	57.77	57.89
TRN+UpDn [7]	57.44	-
MuRel [5]	57.85	-
VCTREE+HL [21]	59.11	59.34
LENA [8]	59.52	59.87
Ours	57.45	57.84

Table 3. Comparison on *Other* split of VQA 2.0 dataset.

Q: What colors are the dogs?
A: black, white and yellow

Q: What is this guy holding?
A: frisbee

Figure 5. samples in the VQA 2.0 dataset.

Table 3 shows that our model achieves comparable results compared with state-of-the-art models. This is mainly due to the following reasons: (1) VQA 2.0 mainly relies on visual appearance clues instead of external knowledge. As shown in Figure 5, the example on the left requires the model to sense colors in multiple regions, while the example on the right requires the model to accurately detect object in the target region. (2) Existing models takes the head answers as the candidate answers, we accumulate multimodal knowledge on the whole dataset to ensure the diver-

sity of answers, which is 10 times larger than the candidate answer set.

H. Implementation Details

For all experiments, we implement our model on top of LXMERT-based-uncased [23] with 2 NVIDIA V100 GPUs. We follow [26] to use Faster R-CNN model [18] pre-trained by the bottom-up model [1] on the Visual Genome dataset [10]. The dimension of inner feed-forward network layer before head entity and relation is set to 1024. The dimensions of entity and relation in multimodal knowledge triplet are set to 300. The parameters in the look-up table are initialized by uniform distribution.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018. 4, 5

[2] Yalong Bai, Jianlong Fu, Tiejun Zhao, and Tao Mei. Deep attention neural tensor network for visual question answering. In *Proceedings of the European Conference on Computer Vision*, pages 20–35, 2018. 4

[3] 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. 4

[4] Hedi Ben-Younes, Remi Cadene, Nicolas Thome, and Matthieu Cord. Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8102–8109, 2019. 4

[5] Remi Cadene, Hedi Ben-Younes, Matthieu Cord, and Nicolas Thome. Murel: Multimodal relational reasoning for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1989–1998, 2019. 4

[6] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2017. 3, 4

[7] Xinzhe Han, Shuhui Wang, Chi Su, Weigang Zhang, Qingming Huang, and Qi Tian. Interpretable visual reasoning via probabilistic formulation under natural supervision. In *European Conference on Computer Vision*, pages 553–570. Springer, 2020. 4

[8] Yudong Han, Yangyang Guo, Jianhua Yin, Meng Liu, Yupeng Hu, and Liqiang Nie. Focal and composed vision-semantic modeling for visual question answering. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4528–4536, 2021. 4

[9] Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling. *International Conference On Learning Representations*, 2016. 4

[10] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael S Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 5

[11] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 13–23, 2019. 3

[12] Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14111–14121, 2021. 3

[13] 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. 3

[14] Duy-Kien Nguyen and Takayuki Okatani. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6087–6096, 2018. 4

[15] Will Norcliffe-Brown, Stathis Vafeias, and Sarah Parisot. Learning conditioned graph structures for interpretable visual question answering. In *Conference and Workshop on Neural Information Processing Systems*, 2018. 4

[16] Liang Peng, Yang Yang, Zheng Wang, Zi Huang, and Heng Tao Shen. Mra-net: Improving vqa via multi-modal relation attention network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 4

[17] Liang Peng, Yang Yang, Zheng Wang, Xiao Wu, and Zi Huang. Cra-net: Composed relation attention network for visual question answering. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1202–1210, 2019. 4

[18] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28:91–99, 2015. 5

[19] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*, 2019. 3

[20] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 5100–5111, 2019. 3

[21] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6619–6628, 2019. 4

[22] Meng Wang, Haofen Wang, Guilin Qi, and Qiushuo Zheng. Richpedia: a large-scale, comprehensive multi-modal knowledge graph. *Big Data Research*, 22:100159, 2020. 1

[23] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, 2020. 5

[24] Zhuoqian Yang, Zengchang Qin, Jing Yu, and Tao Wan. Prior visual relationship reasoning for visual question answering. In *2020 IEEE International Conference on Image Processing*, pages 1411–1415, 2020. 4

[25] Yan Zhang, Jonathon Hare, and Adam Prügel-Bennett. Learning to count objects in natural images for visual question answering. In *International Conference on Learning Representations*, 2018. 4

[26] Zihao Zhu, Jing Yu, Yujing Wang, Yajing Sun, Yue Hu, and Qi Wu. Mucko: Multi-layer cross-modal knowledge reasoning for fact-based visual question answering. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 1097–1103, 2020. 5