Graph representation of scenes and questions

Chaonan Song

May 21, 2018

1 Evaluation

1.1 Evaluation on the "abstract scenes" dataset

They report our results on the original "abstract scenes" dataset in Table 1. The evaluation is performed on an automated server that does not allow for an extensive ablative analysis. Anecdotally, performance on the validation set corroborates all findings presented above, in particular the strong benefit of pre-parsing, pretrained word embeddings, and graph processing with a GRU. At the time of their submission, our method occupies the top place on the leader board in both the open-ended and multiple choice settings. The advantage over existing method is most pronounced on the binary and the counting questions. Refer to Fig. 1 and to the supplementary for visualizations of the results.

2 Conclusions

They presented a deep neural network for visual question answering that processes graph-structured representations of scenes and questions. This can make use of existing natural language processing tools, in particular pre-trained word embedding and syntactic analysis. The latter shows a significant advantage over the traditional sequential processing of problems *e.g.* with LSTMs. In their opinion, VQA systems are unlikely to learn everything from question/answer examples alone. They believe that any significant improvement in performance will require additional sources of information and supervision. Their explicit processing of the language part is a small step in that direction. It has clearly shown that generalization can be improved without relying entirely on VQA-specific annotations. So far, they have applied their method to the dataset of the clip scene. By replacing the nodes in the input scene graph with suggestions made by pre-trained object detectors, it will solve the direct expansion of the actual image in future work.

Method	Overall	Yes/no	Other	Number	Overall	Yes/no	Other	Number
LSTM blind [2]	61.41	76.90	49.19	49.65	57.19	76.88	38.79	49.55
LSTM with global image features [2]	69.21	77.46	66.65	52.90	65.02	77.45	56.41	52.54
Zhang *et al.* [5] (yes/no only)	35.25	79.14			35.25	79.14		
Multimodal residual learning [3]	67.99	79.08	61.99	52.57	62.56	79.10	48.90	51.60
U. Tokyo MIL (ensemble) [1, 4]	71.18	79.59	67.93	56.19	69.73	80.70	**62.08**	58.82
Graph VQA (full model)	74.37	79.74	68.31	74.97	70.42	81.26	56.28	76.47

Table 1: Results on the test set of the "abstract scenes" dataset (average scores in percents).

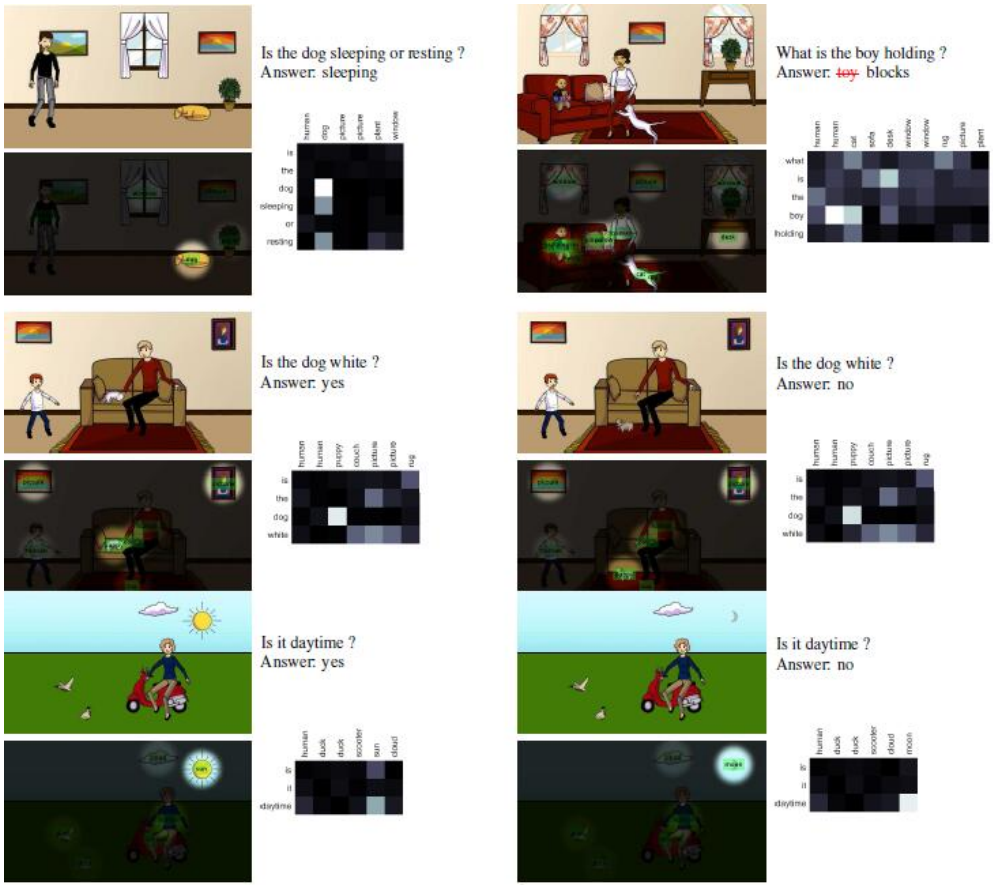

Figure 1: Qualitative results on the abstract scenes dataset (top row) and on balanced pairs (middle and bottom row). We show the input scene, the question, the predicted answer, and the correct answer when the prediction is erroneous. We also visualize the matrices of matching weights (Eq. 6, brighter correspond to higher values) between question words (vertically) and scene objects (horizontally). The matching weights are also visualized over objects in the scene, after summation over words, giving an indication of their estimated relevance. The ground truth object labels are for reference only, and not used for training or inference.

References

[1] Vqa challenge leaderboard. http://visualqa.org/challenge.html.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. *In Proc. IEEE Int. Conf. Comp. Vis*, 2015.

[3] Jin Hwa Kim, Sang Woo Lee, Dong Hyun Kwak, Min Oh Heo, Jeonghee Kim, Jung Woo Ha, and Byoung Tak Zhang. Multimodal residual learning for visual qa. 2016.

[4] Kuniaki Saito, Andrew Shin, Yoshitaka Ushiku, and Tatsuya Harada. Dualnet: Domain-invariant network for visual question answering. In *IEEE International Conference on Multimedia and Expo*, pages 829–834, 2017.

[5] Peng Zhang, Yash Goyal, Douglas Summersstay, Dhruv Batra, and Devi Parikh. Yin and yang: Balancing and answering binary visual questions. In *Computer Vision and Pattern Recognition*, pages 5014–5022, 2016.