Graph-Structured Representations for Visual Question Answering

Cheng Guan

May 21 , 2018

1 Introduction

The task of Visual Question Answering has received growing interest in the recent years(see [1, 3]for example). One of the more interesting aspects of the problem is that it combines computer vision, natural language processing, and artificial intelligence. In its open-ended form, a question is provided as text in natural language together with an image, and a correct answer must be predicted, typically in the form of a single word or a short phrase. In the multiple-choice variant, an answer is selected from a provided set of candidates, alleviating evaluation issues related to synonyms and paraphrasing.Some explanation of nouns in Table 1.

Multiple datasets for VQA have been introduced with either real [3] or synthetic images [2, 3]. Our experiments uses the latter, being based on clip art or cartoon images created by humans to depict realistic scenes (they are usually referred to as abstract scenes, despite this being a misnomer). Our experiments focus on this dataset of clip art scenes as they allow to focus on semantic reasoning and vision-language interactions, in isolation from the performance of visual recognition see examples in Fig. 1.They also allow the manipulation of the image data so as to better illuminate algorithm performance. A particularly attractive VQA dataset was introduced in [2] by selecting only the questions with binary answers (e.g. yes/no) and pairing each (synthetic) image with a minimally-different complementary version that elicits the opposite (no/yes) answer (see examples in Fig. 1, bottom rows). This strongly contrasts with other VQA datasets of real images, where a correct answer is often obvious without looking at the image, by relying on systematic regularities of frequent questions and answers [2, 3].

Table 1: Paraphrase

specialized word	paraphrase
VQA	Visual Question Answering
CNN	Convolutional Neural Network
LSTM	Long Short-Term Memory

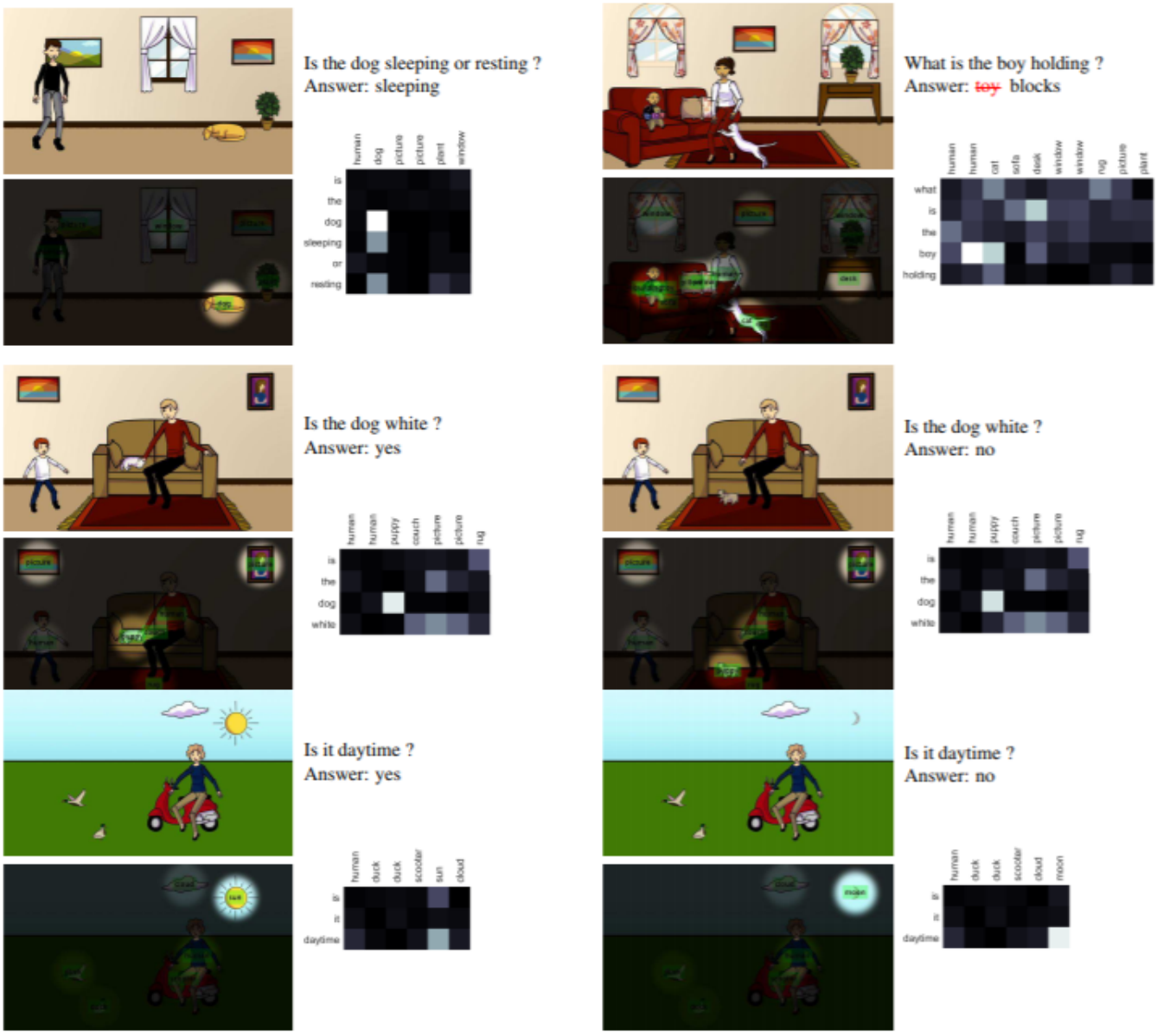

Figure 1: Qualitative results on the abstract scenes dataset (top row) and on balanced pairs (middle and bottom row). We show the input scene, the question, the predicted answer, and the correct answer when the prediction is erroneous. We also visualize the matrices of matching weights between question 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] M. Malinowski and M. Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. *In Proc. Advances in Neural Inf. Process. Syst*, pages 1682–1690, 2014.

[2] D. Summers-Stay D. Batra P. Zhang, Y. Goyal and D. Parikh. Yin and yang: Balancing and answering binary visual questions. *In Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2016.

[3] J. Lu-M. Mitchell D. Batra C. L. Zitnick S. Antol, A. Agrawal and D. Parikh. Vqa: Visual question answering. *In Proc. IEEE Int. Conf. Comp. Vis*, 2015.