Graph-Structured Representations for Visual Question Answering

Chaonan Song

May 9, 2018

1 Abstract

This article proposes to improve the visual question answer (VQA) by using a structured representation of scene content and problems. One of the difficulties with VQA is the requirement for joint reasoning in the visual and text areas. The dominant VQA approach based on CNN / LSTM is largely constrained by the overall vector representation and largely ignores the structure in the scenario and problem. The CNN feature vector cannot effectively capture the context, while the LSTM treats the problem as a series of words, which cannot truly reflect the structure of the language. They propose to create graphs on scene objects and problem words and describe a deep neural network with structures in these representations. They found that this method has made significant progress over the existing technology.

2 Introduction

The task of Visual Question Answering has received increasing attention in recent years see, for example [1,3]. One aspect of the problem is the combination of computer vision, natural language processing and artificial intelligence. The text as a natural language together with the image provides the problem, and the correct answer must be predicted. In multiple choice variants, one answer is selected from a set of candidate answers provided to alleviate the problems associated with synonyms and paraphrases. Some data sets of VQA have introduced real [1–4, 6] or composite images [1,5] Their experiments use "cartoon" images based on clip art or human creation to describe realistic scenes. Their experiments focused on editing this dataset of the art scene because they allowed to focus on semantic reasoning and visual language interaction (see the example in Figure 1). They also allow the processing of image data to better illustrate algorithm performance. In [5], the opposite answer is elicited by selecting only questions with binary answers, and each (synthetic) image with a minimally complementary version. This is very different from the VQA data set of other real images, in that the correct answer is usually obvious and does not require viewing the image by relying on the systematic laws of common questions and answers [1,5]. In their view, despite the obvious limitations of composite images, improvements to the above-mentioned "balanced" data set constitute heuristic measures of the progress of the scene understanding, because the language model itself does not perform as well as these data.

Challenges The problem of clip data sets varies greatly in complexity. Some can answered directly by observing visual elements. Other question need to involve multiple many infants or understand complex behavior. The other challenge impact all data sets of VQA is the Sparse of training data. Even a large number of training problem can not cover the diversity of possible objects and concepts.

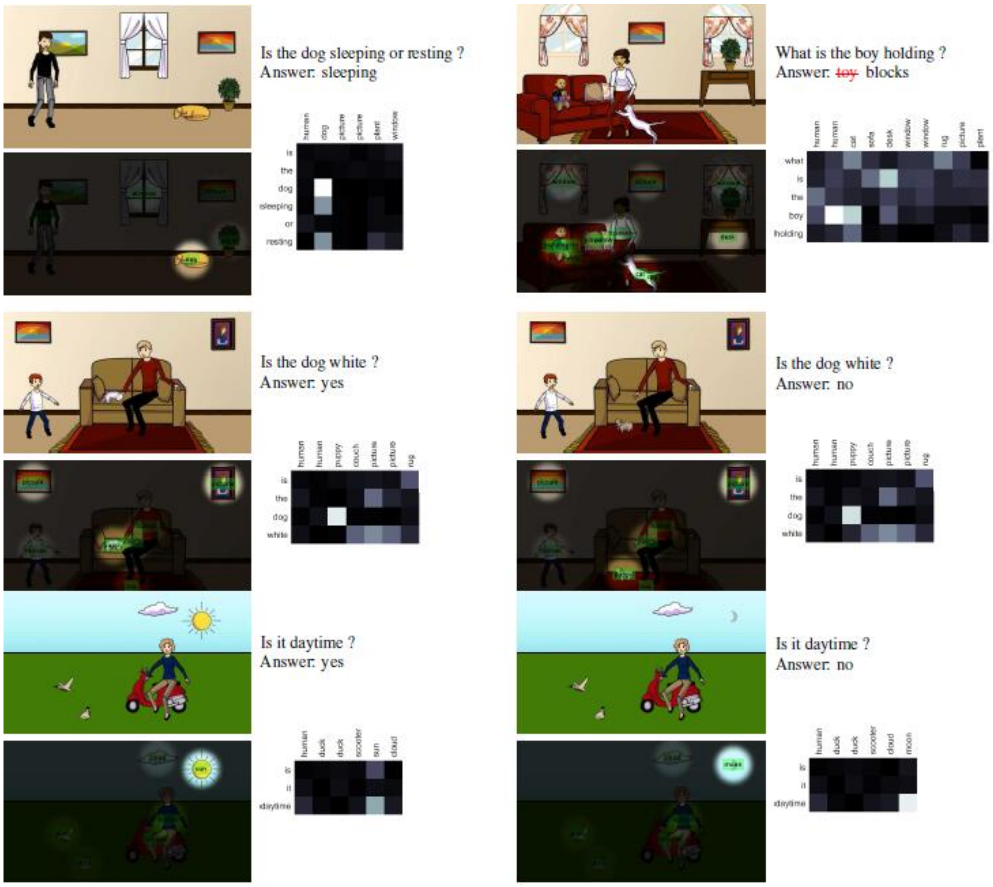

Figure 1: Qualitative results on the abstract scenes dataset (top row) and on balanced pairs (middle and bottom row). We show the input scene, the question, the predicted answer, and the correct answer when the prediction is erroneous. We also visualize the matrices of matching weights (Eq. 6, brighter 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] 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.

[2] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Li Jia Li, Li Jia Li, and David A. Shamma. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2016.

[3] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *International Conference on Neural Information Processing Systems*, pages 1682–1690, 2014.

[4] Mengye Ren, Ryan Kiros, and Richard Zemel. Image question answering: A visual semantic embedding model and a new dataset. *Litoral Revista De La Poesa Y El Pensamiento*, pages pgs. 8–31, 2015.

[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.

[6] Yuke Zhu, Oliver Groth, Michael Bernstein, and Fei Fei Li. Visual7w: Grounded question answering in images. In *Computer Vision and Pattern Recognition*, pages 4995–5004, 2016.