Graph-Structured Representations for Visual Question Answering

Wenjie Niu

June 3, 2018

Abstract

This paper proposes to improve visual question answering (VQA) with structured representations of both scene contents and questions. A key challenge in VQA is to require joint reasoning over the visual and text domains. The predominant CNN/LSTM-based approach to VQA is limited by monolithic vector representations that largely ignore structure in the scene and in the question. CNN feature vectors cannot effectively capture situations as simple as multiple object instances, and LSTMs process questions as series of words, which do not reflect the true complexity of language structure. We instead propose to build graphs over the scene objects and over the question words, and we describe a deep neural network that exploits the structure in these representations. We show that this approach achieves significant improvements over the state-of-the-art, increasing accuracy from 71.2% to 74.4% on the abstract scenes multiple-choice benchmark, and from 34.7% to 39.1% for the more challenging balanced scenes, i.e. image pairs with fine-grained differences and opposite yes/no answers to a same question.

1 Introduction

The task of Visual Question Answering has received growing interest in the recent years (see [3], [1], [6] 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.

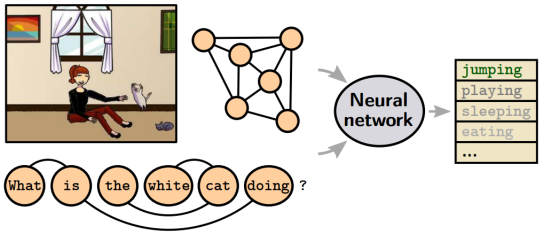

Figure 1: We encode the input scene as a graph representing the objects and their spatial arrangement, and the input question as a graph representing words and their syntactic dependencies. A neural network is trained to reason over these representations, and to produce a suitable answer as a prediction over an output vocabulary.

Multiple datasets for VQA have been introduced with either real [1], [2], [4], [5], [8] or synthetic images [1], [7]. Our experiments uses the latter, being based on clip art or cartoon images created by humans to depict realisticscenes (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. 2). They also allow the manipulation of the image data so as to better illuminate algorithm performance. A particularly attractive VQA dataset was introduced in [7] 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. 2, 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 [1], [7]. Performance improvements reported on such datasets are difficult to interpret as actual progress in scene understanding and reasoning as they might similarly be taken to represent a better modeling of the language prior of the dataset. This hampers, or at best obscures, progress toward the greater goal of general VQA. In our view, and despite obvious limitations of synthetic images, improvements on the aforementioned balanced dataset constitute an illuminating measure of progress in scene-understanding, because a language model alone cannot perform better than chance on this data.

The advantage of this approach with text- and scene-graphs, rather than more typical representations, is that the graphs can capture relationships between words and between objects which are of semantic significance. This contrasts with the typical approach of representing the image with CNN activations (which are sensitive to individual object locations but less so to relative position) and the processing words of the question serially with an RNN (despite the fact that grammatical structure is very non-linear). The graph representation ignores the order in which elements are processed, but instead represents the relationships between different elements using different edge types. Our network uses multiple layers that iterate over the features associated with every node, then ultimately identifies a soft matching between nodes from the two graphs. This matching reflects the correspondences between the words in the question and the objects in the image. The features of the matched nodes then feed into a classifier to infer the answer to the question (Fig. 1).

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. *ICCV*, 2015. 1, 2

[2] 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 Fei-Fei Li. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *CVPR*, 2016. 1

[3] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. *NIPS*, 2014. 1

[4] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. *CVPR*, 2014. 1

[5] Mengye Ren, Ryan Kiros, and Richard S. Zemel. Image question answering: A visual semantic embedding model and a new dataset. *CVPR*, 2015. 1

[6] Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony R. Dick, and Anton van den Hengel. Visual question answering: A survey of methods and datasets. *CVPR*, 2016. 1

[7] Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Yin and yang: Balancing and answering binary visual questions. *CVPR*, 2015. 1, 2

[8] Yuke Zhu, Oliver Groth, Michael S. Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. *CVPR*, 2015. 1

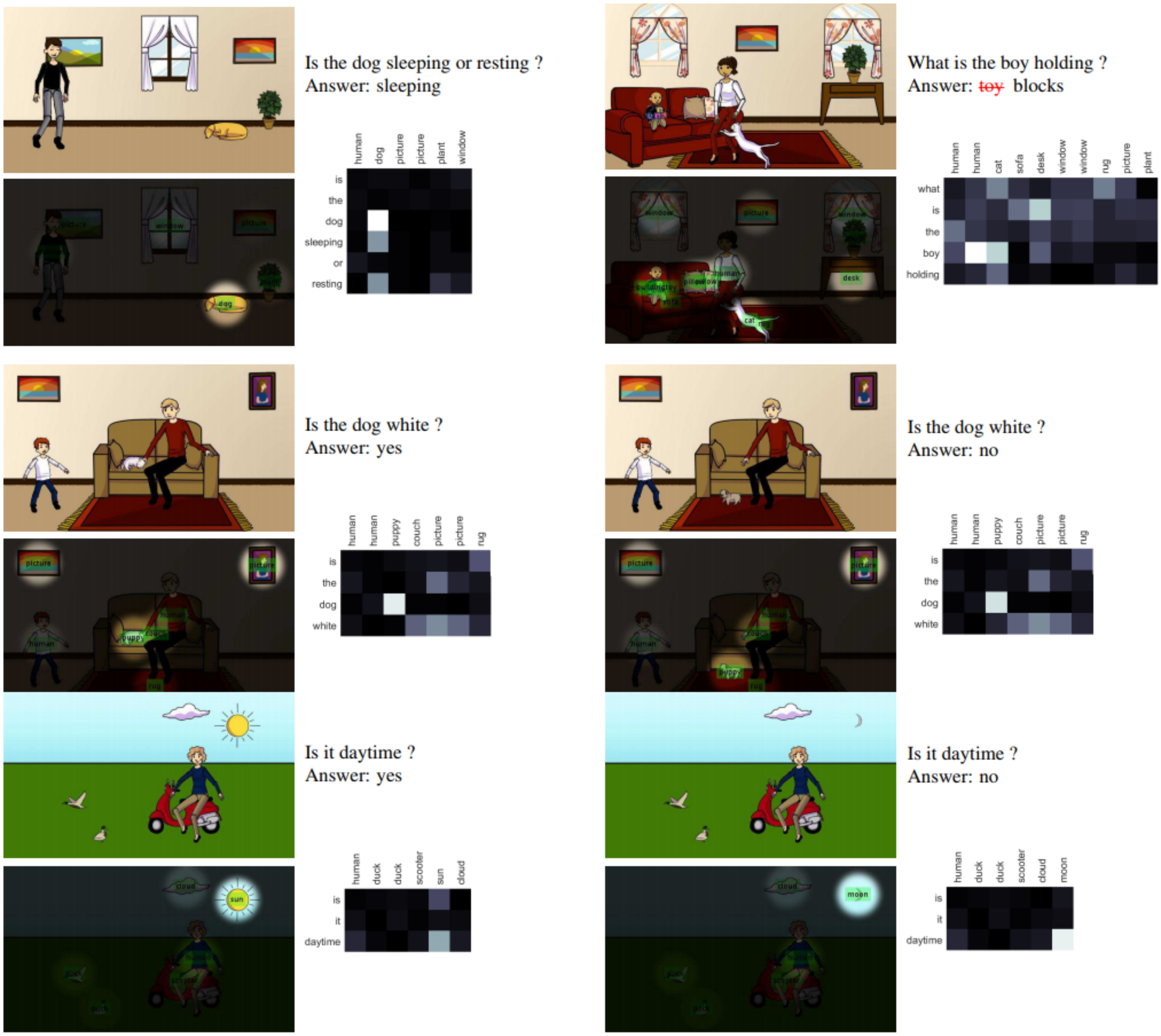
Is the dog sleeping or resting ?
Answer: sleeping
human dog picture picture plant window
is the dog sleeping or resting
picture window picture plant dog
What is the boy holding ?
Answer: ~~toy~~ blocks
human human cat sofa desk window window rug picture plant
what is the boy holding
window picture window desk sofa pillow human cat rug
Is the dog white ?
Answer: yes
human human puppy couch picture picture rug
is the dog white
picture picture human puppy couch rug
Is the dog white ?
Answer: no
human human puppy couch picture picture rug
is the dog white
picture picture human couch puppy rug
Is it daytime ?
Answer: yes
human duck duck scooter sun cloud
is it daytime
cloud sun
Is it daytime ?
Answer: no
human duck duck scooter cloud moon
is it daytime
cloud moon

Figure 2: Qualitative results on the abstract scenes dataset (top row) and on balanced pairs (middle and bottom row).