CS671: NATURAL LANGUAGE PROCESSING
PROJECT PROPOSAL

Visual Question Answering

Avi Singh
October 2, 2015

1 INTRODUCTION AND MOTIVATION

In recent years, there has been a lot of progress in AI problems at the intersection of NLP and Computer Vision. One problem that has garnered a lot of attention recently is Image Captioning[1, 2, 3]. However, the task is not well suited to track the progress of AI since image captions are nonspecific, and its automatic evaluation is still an open problem[4]. Another such problem is Visual Question Answering[5, 6]. In this task, the input is an image and question based on the image, and an output is an one or more words that answer the question. Open-ended question answering requires one to solve several lower-level problems like fine-grained recognition, object detection, activity recognition, common-sense reasoning, and knowledge-based reasoning[6]. Due to the specificity of the task, it can also be evaluated automatically, making it easier to track progress. These characteristics make it an ideal "AI-complete" problem, suitable for a Turing Test[7].

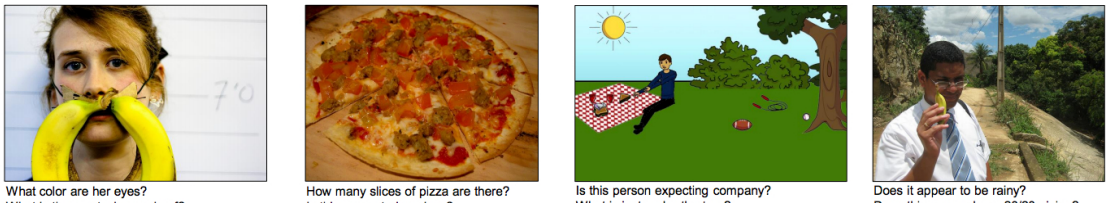

Figure 1.1: A teaser from the VQA dataset[6]

2 Related Work

Almost all of the work in Visual Question and answering has been done in the last two years. Malinowski et al.[5] released the first image Q&A dataset, DAQUAR. They initially used a multi-word symbolic approach to the problem, but their subsequent work was based on a combination of Long Short Term Memory Networks (LSTM) and Convolutional Neural Networks (CNN)[8]. A research group at Baidu has reported results on a dataset with Chinese questions[9], automatic translations of which are also available, and results are reported on both. The most recent result is that of Ren et al.[10], and it shows that even simple Bag of Words models are hard to beat using LSTMs. They reported result on a dataset of autmatically generated questions. Recently, the VQA[6] dataset was released, and it came with strong baselines based on CNN features combined with a bag-of-word model. Apart from the original paper that introduces the dataset, no results have been published on it.

3 Datasets

1. Visual Question Answering (VQA) dataset[6]: Based on images from the COCO dataset,it currently has 360K questions on 120K images. There are plans of releasing questions on the rest of the COCO images and an additional 50K abstract images. All the questions are human-generated, and were specifically designed to stump a "smart robot".

2. Visual Madlibs[11]: It contains fill-in-the-blank type questions along with standard question-answer pairs. It has 360K questions on 10K images from the COCO dataset. A lot of questions require high-level human cognition, such as describing what one feels on seeing an image.

3. Toronto COCO-QA Dataset[10]: Automatically generated questions from the captions of the MS COCO dataset. At 115K questions, it is smaller than the VQA dataset. Answers are all one word.

4. DAQUAR[5] - DAtaset for QUestion Answering on Real-world images: A much smaller dataset, with about 12K questions. This was one of the earliest datasets on image question and answering.

4 Proposed Method

This project will utilize the VQA dataset[6], since it is much larger and more challenging than previous datasets. The aim of the project is to beat the existing benchmarks set by the authors of VQA. We would be using different types of Neural Networks, including variants of Recurrent Neural Networks for modeling questions and Convolutional Neural Networks for extracting features from images. The project would begin with implementing existing Neural Network-based approaches[8, 10]. Subsequently, modern neural network architectures (especially those augmented with a memory) will be explored.

REFERENCES

[1] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[2] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[3] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[4] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[5] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. *Neural Information Processing Systems (NIPS)*, 2014.

[6] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. *International Conference on Computer Vision (ICCV)*, 2015.

[7] Mateusz Malinowski and Mario Fritz. Towards a visual turing challenge. *NIPS Workshop*, 2014.

[8] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. Ask your neurons: A neural-based approach to answering questions about images. *International Conference on Computer Vision (ICCV)*, 12/2015 2015.

[9] Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, and Alan Yuille. Are you talking to a machine? dataset and methods for multilingual image question answering. *Neural Information Processing Systems (NIPS)*, 2015.

[10] Mengye Ren, Ryan Kiros, and Richard S. Zemel. Exploring models and data for image question answering. *Neural Information Processing Systems (NIPS)*, 2015.

[11] Licheng Yu, Eunbyung Park, Alexander C. Berg, and Tamara L. Berg. Visual Madlibs: Fill in the blank Image Generation and Question Answering. *arXiv preprint arXiv:1506.00278*, 2015.