Image Understanding via Continuous Questioning and Answering

Vatsal Mahajan (vmahaja1@asu.edu), and Saurabh Singh (ssing139@asu.edu), *ASU*

Abstract—In this project we take inspiration from "Neural Self Talk: Image Understanding via Continuous Questioning and Answering" [Yang *et al.*, 2015] to work on the problem of continuously discovering image contents by actively asking image based questions and subsequently answering the questions being asked. We will create two modules, Visual Question Generation(VQG) and Visual Question Answering(VQA) module for which Convolution Neural Network and Recurrent Neural Network will be used. Both VQG and VQA are trained simultaneously, VQG uses image as input and corresponding question as output whereas VQA module uses images and questions as input and corresponding answer as output.

1 Problem Description

One of the important research work going in deep neural network is Image captioning and Visual Question Answering. This task is to include the semantic analysis of the image and answering the question referring to the image. The project proposes an end to end system that can continuously discover novel questions on an image, and then provide legitimate answers. This "self talk" approach for image understanding goes beyond as just a visual task, but to solve an interdisciplinary AI problem in vision & language.

2 Related Work

We are witnessing a renewed interest in interdisciplinary AI research in vision & language. The most established work in the vision & language community is 'image captioning', where the task is to produce a literal description of the image. It has been shown [Devlin *et al.*, 2015; Fang *et al.*, 2014] that a reasonable language modeling paired with deep visual features trained on large enough datasets promise a good performance on image captioning, making it a less challenging task from language learning perspective.

Previous approaches of question generation from natural language sentences are mainly through template matching [Brown *et al.*, 2005]. In [Yang *et al.*, 2015] they propose a visual question generation module through a technique directly adapted from image captioning system [Karpathy and Li, 2014], which is data driven and the potential output questions space is significantly larger than previous template based approaches, and the trained module only takes in image as input.

In the filed of Visual Question Answering, very recently researchers spent a significant amount of efforts on both creating datasets and proposing new models [Antol *et al.*, 2015; Malinowski *et al.*, 2015; Gao *et al.*, 2015]. Interestingly both [Antol *et al.*, 2015] and [Gao *et al.*, 2015] adapted MS-COCO [Lin *et al.*, 2014] images and created an open domain dataset with human generated questions and answers. More recently, the work from [Ren *et al.*, 2015] reported state-of-the-art VQA performance using multiple benchmarks. The progress is mainly due to formulating the task as a classification problem and focusing on the domain of questions that can be answered with one word.

3 Approach

The paper [Yang *et al.*, 2015] descirbes an approach for Image Understanding via Continuous Questioning and Answering. (1) We can use the current state-of-the art image captioning systems to geneate questions. (2) And then by using the previous ouput, a system could be trained from human like self-talk. The approach is focused on building 2 modules- (1) Question Generation Module, and (2) Question Answering Module. Fig. 1 shows the architechure descrised in the paper [Yang *et al.*, 2015].

Our work is related to three lines of research of natural image understanding: 1) question generation, 2) image captioning and 3) visual question answering.

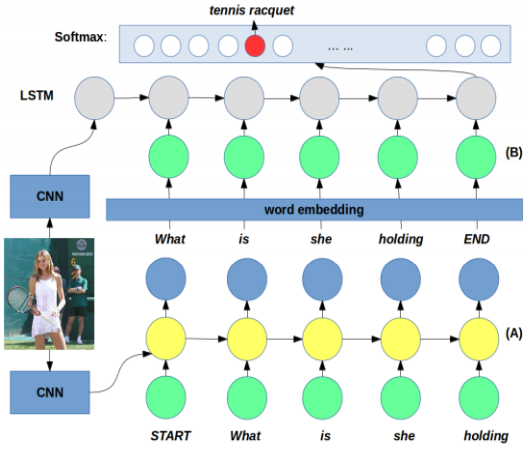

Fig. 1. The architecture of question generation module (part A) and question answering module (part B), and how they are connected as in 'Neural Self Talk'.

4 Timeline

The Table 1 describes the tasks along with their deliverables. Also, every task has the ownership assigned for each of the team members.

TABLE 1
TASKS ANS OWNERSHIP

Dates	Tasks	Extension goals
Task 1 (1 Mar - 20 Mar)	Question Generation Module: We will implement the method from [Karpathy and Li, 2014], where a simple but effective extension is introduced from previously developed Recurrent Neural Networks (RNNs) based language models to train image captioning model effectively.	Using the method from [Malinowski *et al.*, 2015] to improve readability of the conversation.
ownership	Vatsal	Saurabh
Task 2 (21 Mar - 12 Apr)	Question Answering Module: We implement the approach from [Ren *et al.*, 2015], which introduced a model builds directly on top of the long short-term memory (LSTM) [Hochreiter and Schmidhuber, 1997] sentence model and is called the VIS+LSTM model.	Adding common sense knowledge for answering questions as described in [Antol *et al.*, 2016].
ownership	Saurabh	Vatsal
Task 3 (13 Apr - 20 Apr)	Evaluation	

5 EXPECTED RESULTS

Given an image the systems iteratively for N times (typically N=5) generates a question and passes it through the VQA system along with the image to achieve the answer.

Fig. 2. Expected result of the system.

Fig. 2 shows the expected result. For this project, we will be using these two datasets, namely, DARQUAR [Malinowski and Fritz, 2014] and MSCOCO-VQA [Antol *et al.*, 2015].

6 EVALUATION

For the generated question answer pairs, since there are no ground truth annotations that could be used for automatic evaluation, we would be using human evaluation. For evaluation, we will use these three metrics: 1) Readability Score: measures how readable is the generated conversation (range from 1 to 5). 2) Correctness Score: measures how correctly the content of the generated QA pairs describes the image content (range from 1 to 5); 3) Human-likeness Score: measures how human-like does the robot perform (range from 1 to 5).

REFERENCES

[1] [Yang *et al.*, 2015] Yezhou Yang, Yi Li, Cornelia Fermuller, Yiannis Aloimonos. Neural Self Talk: Image Understanding via Continuous Questioning and Answering. arXiv: 1512.03460, 2015.

[2] [Devlin *et al.*, 2015] Jacob Devlin, Hao Cheng, Hao Fang, Saurabh Gupta, Li Deng, Xiaodong He, Geoffrey Zweig, and Margaret Mitchell. 2015. Language models for image captioning: The quirks and what works. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 100–105, Beijing, China, July. Association for Computational Linguistics.

[3] [Brown *et al.*, 2005] Jonathan C Brown, Gwen A Frishkoff, and Maxine Eskenazi. Automatic question generation for vocabulary assessment. In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, pages 819–826. Association for Computational Linguistics, 2005.

[4] [Karpathy and Li, 2014] Andrej Karpathy and Fei-Fei Li. Deep visual-semantic alignments for generating image descriptions. arXiv preprint arXiv:1412.2306, 2014.

[5] [Antol *et al.*, 2015; Antol *et al.*, 2016] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In International Conference on Computer Vision (ICCV), 2015 and arXiv:1505.00468, 2016.

[6] [Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C Lawrence Zitnick. Microsoft coco: Com- ´ mon objects in context. In Computer Vision–ECCV 2014, pages 740–755. Springer, 2014.

[7] [Ren *et al.*, 2015] Mengye Ren, Ryan Kiros, and Richard Zemel. Exploring models and data for image question answering. arXiv preprint arXiv:1505.02074, 2015.

[8] [Fang *et al.*, 2014] Hao Fang, Saurabh Gupta, Forrest N. Iandola, Rupesh Srivastava, Li Deng, Piotr Dollar, Jianfeng ´ Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. 2014. From captions to visual concepts and back. CoRR, abs/1411.4952.

[9] [Malinowski *et al.*, 2015] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. Ask your neurons: A neuralbased approach to answering questions about images. arXiv preprint arXiv:1505.01121, 2015.

[10] [Gao *et al.*, 2015] Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. Are you talking to a machine? dataset and methods for multilingual image question answering. arXiv preprint arXiv:1505.05612, 2015.

[11] [Malinowski and Fritz, 2014] Mateusz Malinowski and Mario Fritz. Towards a visual turing challenge. arXiv preprint arXiv:1410.8027, 2014.

[12] [Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jurgen Schmidhuber. Long short-term memory. Neural computation, 9(8):1735–1780, 1997.