Project Proposal: Visual Question Answering

Arushi Kumar (2014023), Avneet Kaur (2014027) , Purusharth Dwivedi(2014081)

Abstract— **With the growth in the field of camera technologies and the advent of social media applications, it has become a common activity to click and upload pictures of anything and everything around the world. As a result, there is a plethora of information on the internet available in the form of natural language as well as images. Hence, there is a need to develop a system for answering factoid as well as descriptive questions about or based on these images, in order to present knowledge in a readable way without human intervention. This has an application in developing search engines that provide to the point answers to image based, topic specific questions rather than presenting unnecessary information in the form of an entire document pertaining to a particular topic.**

I. INTRODUCTION

Language and vision problems like visual question answering(VQA)[1] have gained popularity in recent years towards solving multi-modal problems. A VQA system takes an image and a natural language question as an input and produces an answer to that question as an output. These questions could target different areas of the image like background context, object details, etc. The questions could be based on object detection (Eg. How many balls are there?), activity recognition (Eg. Is this man laughing?), fine-grained recognition (Eg. What is the color of her eyes?). This kind of an AI system will enable perception in computers in the form of visuals, processing questions as natural language and answering them, thus mimicking the human brain. Also, it has an application in knowledge representation and reasoning.

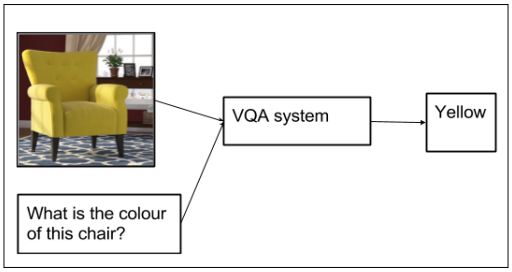

Fig. 1. System

II. WORKFLOW

A. Literature Survey

B. Dataset Curation and Analysis

Obtaining the already available datasets used in this problem like : COCO-QA, DAQUAR (Malinowski and Fritz, 2014) etc. Analyzing the nature of dataset in terms of types of questions (open ended questions, yes/no questions etc.), types of answers (yes/no, descriptive answers, object, color, counting, and location,quantity based etc.) , length of answers, analysis of image tags in terms of what descriptive features are available.

C. Dataset Preparation (Probable,if time permits)

Scraping social media websites like facebook,twitter for images about events, places, etc and their tags, and preparing question/answer pairs for the same.

D. Exploration of probabilistic based (and other non deep-learning) models

One of the models is based on Answer Type Prediction [2].The answer type is crucial to any natural language question so as to know the user intention. This part includes exploring a Bayesian framework in which the answer type is predicted for a question and then this is used to generate the natural language answer for a particular question related to an image.

E. Exploration of image captioning techniques

This would be done in order to provide contextual background of the scene / objects in an image. In case of an unseen dataset, for a new image, without any tags, it would be useful if the system could provide some contextual background in the form of captions.

F. Exploration of Deep Learning based models

Trying out various CNN ,LSTM, word-embedding based models applied to this problem [3].

G. Comparisons and Evaluations

This would be done for the various techniques applied to this task.

H. Developing a complete system

This would involve integrating the model developed so far with a user interface in order to complete the system. The input to the system will be an image chosen by the user and a question associated with it. The output of the same will be a answer (may be descriptive or factoid) along with a knowledge graph representation / description if required, (for factoid questions) in order to provide complete context to the user about a particular topic. The complete system outline is given in Figure 3.

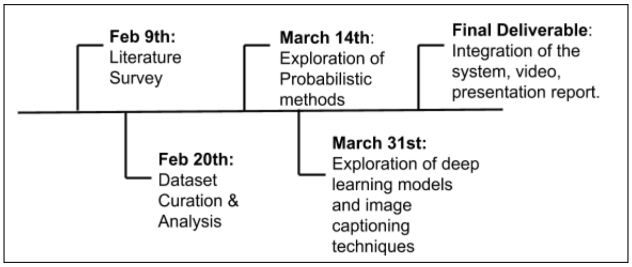
Feb 9th:
Literature Survey
March 14th:
Exploration of Probabilistic methods
Final Deliverable:
Integration of the system, video, presentation report.
Feb 20th:
Dataset Curation & Analysis
March 31st:
Exploration of deep learning models and image captioning techniques

Fig. 2. Timeline for the project

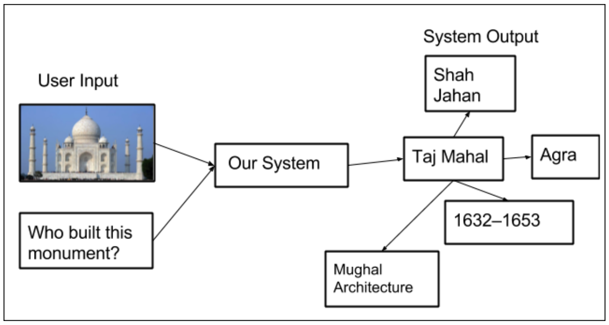

Fig. 3. System

III. Timeline and Deliverables

The timeline for the project is given in Figure 2.

Interim evaluation:

- Dataset curation and analysis.
- Exploration of probabilistic models and developing a baseline model.

Final Deliverable:

- Exploration of image captioning techniques.
- Exploration of Deep Learning based models.
- Integration of the developed model and the user interface to form a complete system.
- Code, report,presentation and video.

References

[1] http://www.visualqa.org/

[2] Kushal Kafle and Christopher Kanan. 2016. Answer- type prediction for visual question answering. In CVPR.

[3] Stanislaw Antol and Aishwarya Agrawal. 2017 Vqa: Visual question answering. International Journal of Computer Vision.