# MCA Project Report

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

## MILESTONES ACHIEVED

## 1)Data Curation & analysis:

The dataset for VQA was curated from various sources like DAQAR, MS-COCO, FM-IQA, VQA-Real, Visual Genome, etc. Following is a table summarising the exploratory dataset analysis for each dataset.

| Dataset | Number Of Questions | Number Of images | Categories of Questions | Method for collecting question |
|---|---|---|---|---|
| DAQUAR | 12,468 | 1449 | 4 | Human |
| COCO-QA | 1,105,904 | 204,721 | 4 | Automatic |
| VISUAL GENOME | 1,445,332 | 108,000 | 7 | Human |
| VQA-real | 614,163 | 204,721 | 20+ | Human |
| VISUAL7W | 327,939 | 47,300 | 7 | Human |
| VISUAL MADLIBS | 360,001 | 10,738 | 12 | Human |

## 2)Exploration of Deep Learning Based Method for Visual Question Answering:

### Simple Baseline: CNN and InceptionV3 based model:

**Overview:**

In this method, 3 CNN models based framework for learning image features , question representations as well as a multi-modal (image + question features) framework for learning the interactions between the two in order to predict the answer is provided. The first model encodes the image content. The second model first finds a word embedding for each word in the question and then considers a convolution unit with a local receptive field and shared weights like in order to obtain important structures between a set of consecutive words. The max pooling layer is used to select interesting sequence of words, and also filters meaningless ones (eg: "red chair" would be selected and "in front of the" would be filtered out.). After several layers of convolution and max pooling are applied, a representation for the question is finally obtained. The third CNN model, in addition to taking the image representation as an individual semantic component, combines it along with the textual representations in order to learn the interactions between the multi-modal data and produces the desired output. The dataset for this purpose was MS-COCO dataset.

**Methodology and Implementation (Feature Extraction):**

To capture the semantic content of an image, we use a pre-trained InceptionV3 network trained on the ImageNet database. InceptionV3 by Google is one of the best performing models on the ImageNet competition. For our purposes however, we remove the softmax classification layer from the end, and use the values of the last dense layer as a feature vector. This feature vector captures the semantic representation of the image in a vector form.

**Sentence Embedding**

To capture the question semantics, we are exploring 3 different approaches for generating question embedding vectors. All these 3 different approaches however first require converting the individual words into embedding vectors. This can be done by using pre trained word-to-vec models or Glove embeddings.
Example question: How many people are present in this image ?
The 3 approaches are as follows:

→ CNN on word embedding matrix of the sentence.

→Using LSTMs for learning sentence embeddings.

→ Word vector averaging:
In this approach, we simply add the word vectors of all the words and divide by the total number of words to obtain the resultant vector. This vector is further normalised to obtain a unit vector. Some people have used a slightly modified approach, where normalised unit vectors are added to get the final sentence vector.

→ Word vector concatenation:
Consider the following example sentence, “The cat is sitting on a red table” and the sentence, “ The red cat is sitting on a table”. The embedding that will be generate for

these 2 sentences, using the first approach will be the same. Therefore, in this method, instead of averaging the word vectors, we concatenate the word vectors. This preserves the relationship between adjacent vectors. A potential drawback of this method however is as size of the sentence increases, the embedding size of the sentence increases. It is general practice to truncate the size of the resultant vector to some reasonable size.

**Generating Prediction**:
After the vectors for both the image obtained from the InceptionV3 CNN model and the vectors for the question generated using LSTM, we multiply the two in order to learn a combined representation  and given the superset of all possible answers, we predict the probability that an answer is for a particular question and assign that answer which has the higher score.

## MILESTONES REMAINING

### I. Models to be tried for comparison and improvement:

1. Exploration of Probabilistic Method (Answer Type Prediction)For VQA:
   Answer Type Prediction helps us to find user intent behind the question, a probability based model is proposed in this. This approach can be mainly divided into two stages : predicting answer types, and then predicting the probability of an answer.
2. Explainable VQA using captions and attributes:
   The approach hence proposed breaks the VQA pipeline into two steps: explaining (understanding image content) and reasoning (inferences related to the understanding regarding the answer).
3. Knowledge Based Reasoning using support from external knowledge bases:
   The idea is that some questions may lack background contextual information and are difficult to answer. We also need to know how we arrived at a particular answer, and the explanation for that answer, for which additional support can be taken from the knowledge bases available to us.

### II. Compiling a system for VQA:

In the end, we plan to integrate all of this into a complete system which we will provide as a service, in which the input is images, and question related to that image, and the output is an answer for the same question. We plan to represent it in the form of a knowledge graph along with some description for ease of understanding for the user.