# 

**Ain Shams University**  
**Faculty of Computer & Information Sciences**

**Computer Science Department**

Visual question answering system

**July 2018**

**Ain Shams University**  
**Faculty of Computer & Information Sciences**

**Computer Science Department**



Visual question answering system

**By:**

Rabah Jamal Mohammed Ali [cs]

Abdallah Shaaban Elsayed [cs]

Abdullah Mahmoud Abdullah [cs]

Abdullah Abdelkader Roshdy [cs]

**Under Supervision of:**

[Dr. Sally Saad]

Computer science Department,

Faculty of Computer and Information Sciences,

Ain Shams University.

[Dr. Ahmed Salah]

Computer science Department,

Faculty of Computer and Information Sciences,

Ain Shams University.

# Acknowledgement

First of all we would like to thank ALLAH for granting us the power, faith and blessings throughout this project.

We would like to express our gratitude to our supervisor, Dr. Sally Saad for her guidance, motivation, help, and support at all times, her expertise and understanding, added considerably to our graduate experience.

We also want to thank our dear teaching assistant Dr. Ahmed Salah who has helped us since the idea’s searching phase till the very last day. Special thanks for his help, support and devotion.

We would also like to thank our families and friends for the support they provided us through our entire lives, without their love, encouragement and editing assistance, we would not have reached this stage of our life and we would not have succeeded. And finally we are grateful for everyone who helped us in this project.

# Abstract

We introduce a Visual Question Answering system based on the power of both computer vision and natural language processing. Given an image and a natural language question about the image, the task is to provide an accurate natural language answer, mirroring real-world scenarios, such as helping the visually impaired.

Our system is trained on a huge data-set "COCO-QA dataset" so it has the ability to predict an answer for these types of questions: objects, color, numbers, and locations.

The main idea of the system is divided into 4 main modules. The first module is generating features from an image using convolution neural network, the second one is generating the features from the questions using question hierarchy, the third one is to co-attend between both generated features and the last one is predicting an answer.

We achieved 52.06% accuracy on the testing dataset that covers 20,278 samples of the total 38,948 testing samples.

We enhanced our model to achieve better accuracy but we would not able to retrain the model because of the resources unavailability.

**Table of Contents**

[i](#_Toc516871126)

[Acknowledgement iii](#_Toc516871128)

[Abstract iv](#_Toc516871129)

[List of Figures vii](#_Toc516871130)

[List of Abbreviations viii](#_Toc516871131)

[1- Introduction 1](#_Toc516871132)

[1.1 Motivation 1](#_Toc516871133)

[1.2 Problem Definition 1](#_Toc516871134)

[1.3 Objective 1](#_Toc516871135)

[1.4 Time Plan 2](#_Toc516871136)

[1.5 Document Organization 2](#_Toc516871137)

[2- Background 3](#_Toc516871138)

[2.1 Problem background 3](#_Toc516871139)

[2.2 Scientific background 5](#_Toc516871140)

[2.2.1 Word embedding 5](#_Toc516871141)

[2.2.2 Recurrent Neural Network 7](#_Toc516871142)

[2.2.3 Convolution Neural Network 9](#_Toc516871143)

[2.3 Related works 12](#_Toc516871144)

[3- Analysis and Design 14](#_Toc516871145)

[3.1 System Overview 14](#_Toc516871146)

[3.1.1 System Architecture 14](#_Toc516871147)

[3.1.2 Modules Description 14](#_Toc516871148)

[3.1.3 System Users 15](#_Toc516871149)

[3.2 System Analysis & Design 15](#_Toc516871150)

[3.2.1 Use Case Diagram 15](#_Toc516871151)

[4- Implementation and Testing 16](#_Toc516871152)

[4.1 Implementation 16](#_Toc516871153)

[4.1.1 Preprocessing 16](#_Toc516871154)

[4.1.2 Convolution neural networks 16](#_Toc516871155)

[4.1.3 Question hierarchy 18](#_Toc516871156)

[4.1.4 Attention 21](#_Toc516871157)

[4.1.5 Encoding for Predicting Answers 26](#_Toc516871158)

[4.2 Testing 27](#_Toc516871159)

[5-User Manual 30](#_Toc516871160)

[6-Conclusion and Future Work 32](#_Toc516871161)

[6.1 Conclusion 32](#_Toc516871162)

[6.2 Future Work 33](#_Toc516871163)

[References 33](#_Toc516871164)

# List of Figures

[**Figure 1-Time plan** 2](#_Toc516871184)

[**Figure 2- project overview** 4](#_Toc516871185)

[**Figure 3- word embedding** 5](#_Toc516871186)

[**Figure 4- Question hierarchy** 5](#_Toc516871187)

[**Figure 5- skip gram model** 6](#_Toc516871188)

[**Figure 6- CBOW model** 7](#_Toc516871189)

[**Figure 7- recurrent neural network** 8](#_Toc516871190)

[**Figure 8- LSTM** 8](#_Toc516871191)

[**Figure 9 - Convolution Neural Networks general architecture** 9](#_Toc516871192)

[**Figure 10– Convolution Operation** 10](#_Toc516871193)

[**Figure 11-Relu operation** 10](#_Toc516871194)

[**Figure 12- Max pooling** 11](#_Toc516871195)

[**Figure 13- fully connected layer and output layer** 12](#_Toc516871196)

[**Figure 14-System Architecture** 14](#_Toc516871197)

[**Figure 15-Use case Diagram** 15](#_Toc516871198)

[**Figure 16-VGG19** 17](#_Toc516871199)

[**Figure 17-output of the last pooling layer** 17](#_Toc516871200)

[**Figure 18-Question hierarchy** 18](#_Toc516871201)

[**Figure 19– (a) unigram, (b) bi gram, (c) tri gram** 19](#_Toc516871202)

[**Figure 20- phrase level** 20](#_Toc516871203)

[**Figure 21-Describes the visual attention.** 21](file:///C:\Users\Lenovo-PC\Desktop\Visual-Question-Answering\Seminars%20+%20Proposal%20+%20Documentation\Documentation\Gp.docx2141277323%20(2).docx#_Toc516871204)

[**Figure 22-visual attention** 22](#_Toc516871205)

[**Figure 23 - Question attention** 22](#_Toc516871206)

[**Figure 24-Visualizes the parallel co-attention** 24](#_Toc516871207)

[**Figure 25-Visualizes the alternating co-attention** 25](#_Toc516871208)

[**Figure 26-Visualizes the process of predicting answers** 26](#_Toc516871209)

[**Figure 27-Distribution of questions by their first four words** 27](#_Toc516871210)

[**Figure 28-Distribution of answers per question type** 28](#_Toc516871211)

[**Figure 29- Extract the network IP address** 31](#_Toc516871212)

[**Figure 30 - Run the localhost** 31](#_Toc516871213)

[**Figure 31- Android app sample input** 32](#_Toc516871214)

# List of Abbreviations

* **Cbow Continuous bag of word**
* **COCO-MS Common Objects in Context – Microsoft**
* **COCO-QA Common Objects in Context – Question Answering**
* **Cnn Convolution neural network**
* **Relu Rectified linear unit**
* **Rnn Recurrent neural network**
* **Lstm Long short term memory**
* **Nlp Natural language processing**
* **VQA Visual Question Answering**

# Introduction

## 1.1 Motivation

Visual question answering is a task that was proposed to connect computer vision and natural language processing (NLP), to stimulate research, and push the boundaries of both fields. On the one hand, computer vision studies methods for acquiring, processing, and understanding images. In short, its aim is to teach machines how to see. On the other hand, NLP is the field concerned with enabling interactions between computers and humans in natural language, i.e. teaching machines how to read, among other tasks.

Both computer vision and NLP belong to the domain of artificial intelligence and they share similar methods rooted in machine learning. However, they have historically developed separately. Both fields have seen significant advances towards their respective goals in the past few decades, and the combined explosive growth of visual and textual data is pushing towards a marriage of efforts from both fields. For example, visual question answering.

## 1.2 Problem Definition

Building a system that can answer natural language questions about any image has been considered a very ambitious goal and one of the semantic scene understanding, in the previous years, both natural language and computer vision have been shown great results in solving many complicated tasks but now how the two models can be merged for generating an answer for a given question.

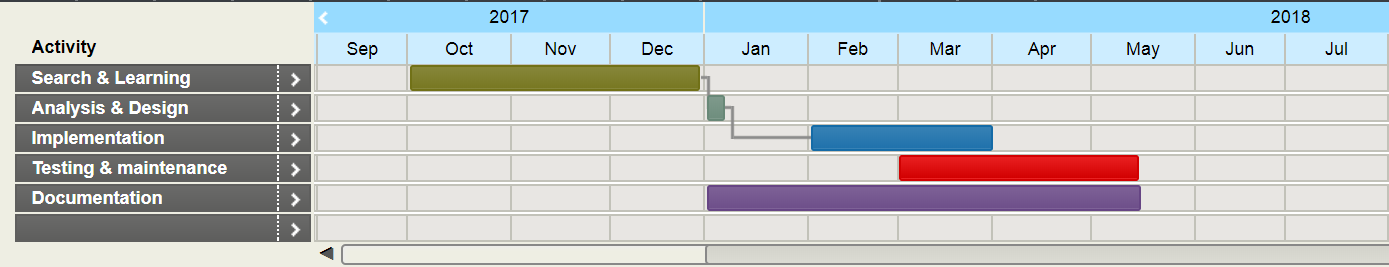
## 1.3 Objective

Our Graduation project aims to predict an answer for a given question about given image using co-attention technique. The system is able to answer different types of questions includes objects, numbers, colors and locations.

Including modules of:

* Extract image features.
* Extract question features
* Apply co-attention technique on the extracted image and question features.
* Predict the answer to the question.

## 1.4 Time Plan



**Figure 1-Time plan**

## 1.5 Document Organization

* **Chapter 2: Background and literature review**

This chapter includes introduction to Visual Question Answering, description of the field of the project and the scientific background related to the project.

* **Chapter 3: Analysis and Design**

This chapter includes diagrams that represent an overview of the whole system along with system functionalities and architecture.

* **Chapter 4: Implementation and Testing**

This chapter talks about implementation methodologies, the used dataset and the results achieved in the testing phase.

* **Chapter 5: User Manual**

This chapter shows how user can use application and how to use application features.

* **Chapter 6: Conclusion and Future work**

This chapter contains challenges faced during design and implementation, system limitations and future work that would be done to improve application performance and features.

# Background

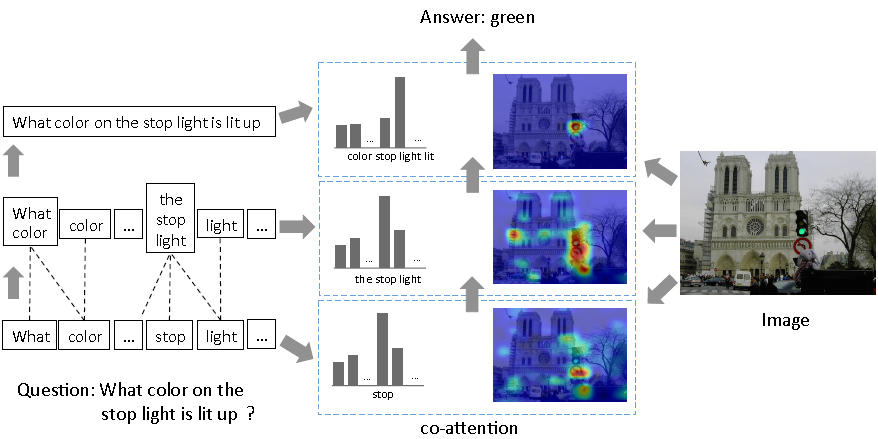
## 2.1 Problem background

Visual Question Answering (VQA) [2, 7, 16, 17, 29] has emerged as a prominent multi-discipline research problem in both academia and industry. To correctly answer visual questions about an image, the machine needs to understand both the image and question. Recently, visual attention based models [20, 23–25] have been explored for VQA, where the attention mechanism typically produces a spatial map highlighting image regions relevant to answering the question.

So far, all attention models for VQA in literature have focused on the problem of identifying “where to look” or visual attention. In this work, we argue that the problem of identifying “which words to listen to” or question attention is equally important. Consider the questions “how many horses are in this image?” and “how many horses can you see in this image?". They have the same meaning, essentially captured by the first three words. A machine that attends to the first three words would arguably be more robust to linguistic variations irrelevant to the meaning and answer of the question. Motivated by this observation, in addition to reasoning about visual attention, we also address the problem of question attention. Specifically, we use a novel multi-modal attention model for VQA with the following two unique features:

**Co-Attention:** We use a novel mechanism that jointly reasons about visual attention and question attention, which we refer to as co-attention. Unlike previous works, which only focus on visual attention, our model has a natural symmetry between the image and question, in the sense that the image representation is used to guide the question attention and the question representation(s) are used to guide image attention.

**Question Hierarchy**: We build a hierarchical architecture that co-attends to the image and question at three levels: (a) word level, (b) phrase level and (c) question level. At the word level, we embed the words to a vector space through an embedding matrix. At the phrase level, 1-dimensional convolution neural networks are used to capture the information contained in unigrams, bigrams and trigrams.



**Figure 2- project overview**

Specifically, we convolve word representations with temporal filters of varying support, and then combine the various n-gram responses by pooling them into a single phrase level representation. At the question level, we use recurrent neural networks to encode the entire question. For each level of the question representation in this hierarchy, we construct joint question and image co-attention maps, which are then combined recursively to ultimately predict a distribution over the answers.

Overall, the main contributions of our work are:

* We use a novel co-attention mechanism for VQA that jointly performs question-guided visual attention and image-guided question attention
* We use a hierarchical architecture to represent the question, and consequently construct image-question co-attention maps at 3 different levels: word level, phrase level and question level. These co-attended features are then recursively combined from word level to question level for the final answer prediction
* At the phrase level, we propose a novel convolution-pooling strategy to adaptively select the phrase sizes whose representations are passed to the question level representation
* Finally, we evaluate our proposed model on two large datasets, VQA [2] and COCO-QA [17].

## 2.2 Scientific background

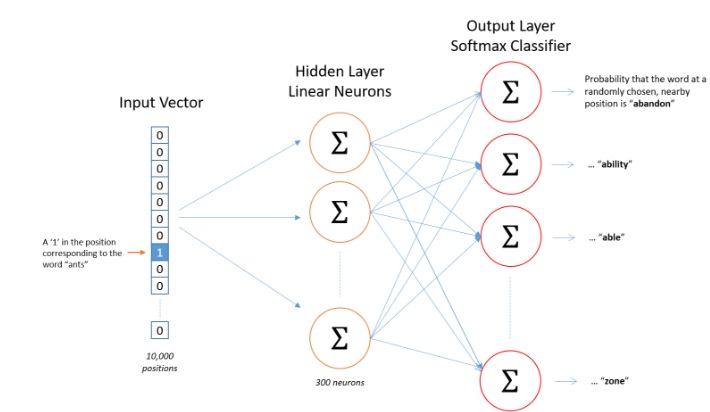
### 2.2.1 Word embedding

Word embedding is a learned representation of text where words that have the same meaning have a similar representation, trained over corpus "large amount of text" to find the similarity between words using conditional probability P(w1|w2).



**Figure 3- word embedding**

There are many word embedding training algorithms like “Skip gram model and CBOW” each of them use a shallow neural network with one hidden layer for training.



**Figure 4- Question hierarchy**

* **Skip gram**

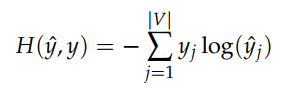
Skip gram is a word-embedding model, which pick a word from a sentence to be an input word then pick a random word from neighboring fixed size window to be target word, and then neural network starts to train to begin to find the similarity of each two words.



**Figure 5- skip gram model**

The number of neurons in the hidden layer is the number of features that we want to generate for a given data-set, so we train the neural network for generating the optimized weight for the hidden layer.

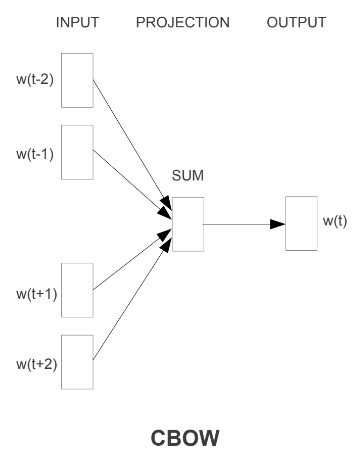
At the output layer we apply softmax function which assigns a probability for each word in the corpus given an input word using cross entropy loss function



**Eq. 1– cross entropy**

* **Continuous Bag of Words (CBOW)**

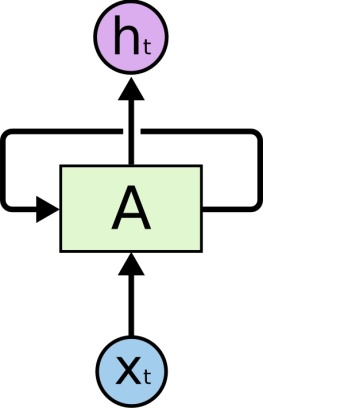
On the other side, the CBOW model aim is to fill in the missing word given its neighboring context, for example, "The cat \_\_ on the mat", the neural network will try to find the highest probability for words to fill the missing word.



**Figure 6- CBOW model**

### 2.2.2 Recurrent Neural Network

Recurrent neural network are popular model that have shown great promise in many NLP tasks, the idea behind RNN is to make use of sequential information. In a traditional neural network, we assume that all inputs are independent of each other, for the given question word's many words depend on the previous word. If you want to predict the next word in a sentence you better know which words came before it. RNN are called recurrent because they perform the same task for every element of a sequence, with the output being depended on the previous computations.



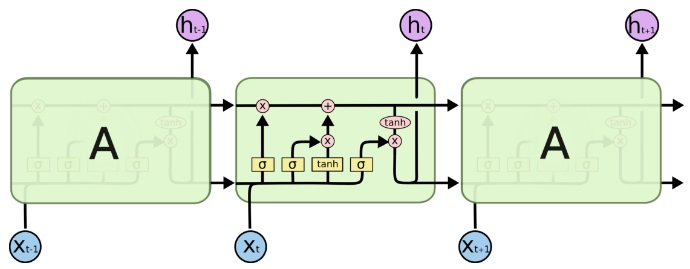
**Figure 7- recurrent neural network**

Vanishing gradient problem is one of the disadvantages of the recurrent neural network, for a given long sentence during the training process if the weight parameter bigger than zero the gradient will explode on the other side if it smaller than zero it will vanish, so the solution for this problem is using LSTM recurrent neural network.

* **Long Short Term Memory (LSTM)**

Long Short Term Memory networks are a special kind of RNN, capable of learning long-term dependencies, remembering information for long periods of time is practically their default behavior.

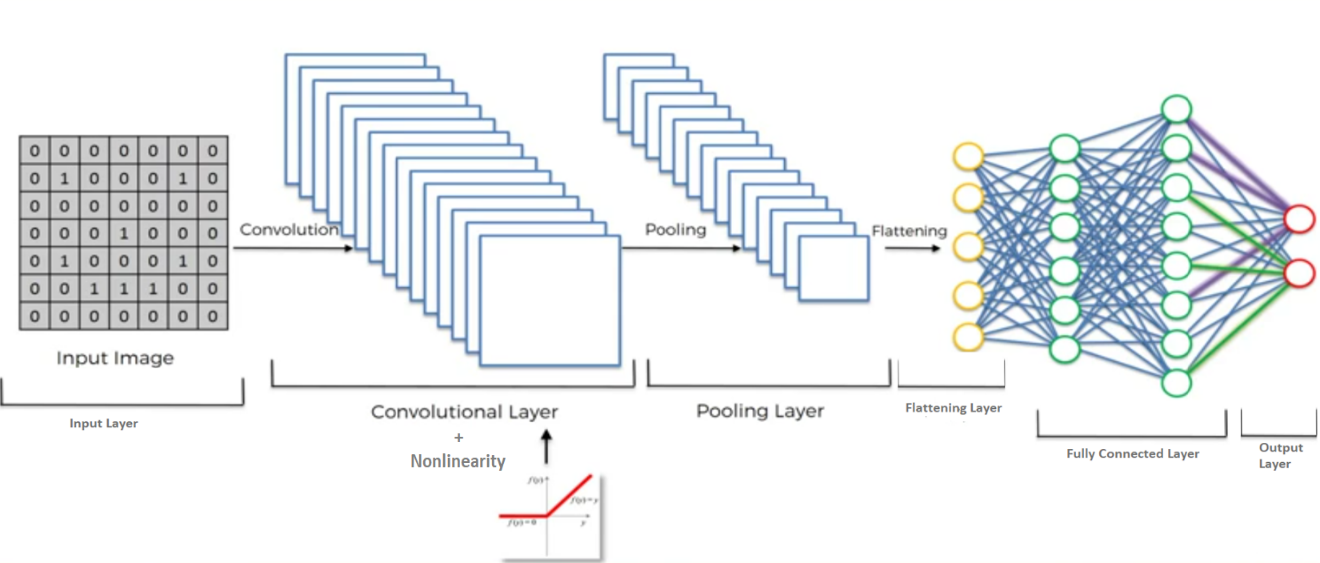
The reason of why LSTM network does not vanish is the activation function used is an identity function, which has a derivative of one, so it will never vanish or explode, also it has gates functions controlling memory cell uses for saving the output of the hidden state at each time step.



**Figure 8- LSTM**

### 2.2.3 Convolution Neural Network

In this subsection, we will talk about the Convolution Neural Network, we will talk about its components, architectures, and features extraction. The convolution neural network is used in our model for extracting the features of an input image.

****

**Figure 9 - Convolution Neural Networks general architecture**

* **Input layer**

The input layer of a neural network is composed of artificial input neurons, and brings the initial data into the system for further processing by subsequent layers of artificial neurons. The input layer is the very beginning of the workflow for the artificial neural network.

* **Convolution Layer**

The primary purpose of convolution is to extract features from the input image.

Convolution preserves the spatial relationship between pixels by learning image features using small squares (filter) of input data.

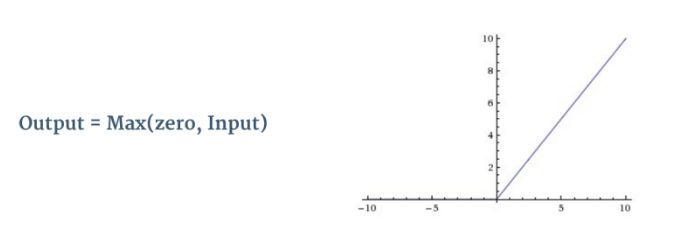


**Figure 10– Convolution Operation**

As the filter is sliding, or convolving, around the input image, it is multiplying the values in the filter with the original pixel values of the image (aka computing element-wise multiplications). These multiplications are all summed up. So, now you have a single number. Next step would be moving the filter to the right by 1 unit, then right again by 1, and so on. Every unique location on the input volume produces a number. After sliding the filter over all the locations, you will find out that you’re left with a smaller array of numbers, which we call an activation map or feature map.

* **Activation Function (RELU)**

An additional operation is to add an activation function, one of the most common used is “Relu” has been used after every Convolution operation in **Figure 10** above. Its output is given by:

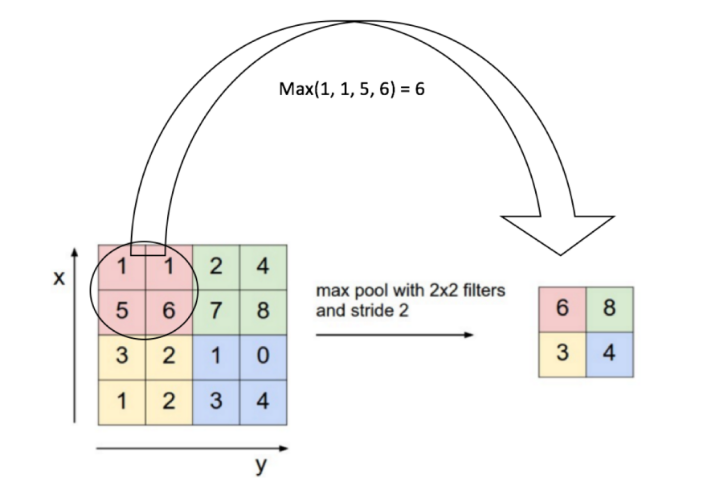


**Figure 11-Relu operation**

Relu is an element wise operation (applied per pixel) and replaces all negative pixel values in the feature map by zero. The purpose of Relu is to introduce non-linearity, since most of the real-world data would be non-linear (Convolution is a linear operation – element wise matrix multiplication and addition, so we account for non-linearity by introducing a non-linear function like Relu).

* **Pooling layer**

Spatial Pooling (also called subsampling or downsampling) reduces the dimensionality of each feature map but retains the most important information. Spatial Pooling can be of different types: Max, Average, Sum etc.



**Figure 12- Max pooling**

Combines the outputs of neuron clusters at one layer into a single neuron in the next layer.

* **Flattening Layer**

Convert the output of the convolution part of the CNN into a 1D feature vector, to be used by the ANN part of it. This operation is called flattening. It gets the output of the convolution layers, flattens all its structure to create a single long feature vector to be used by the dense layer for the final classification.

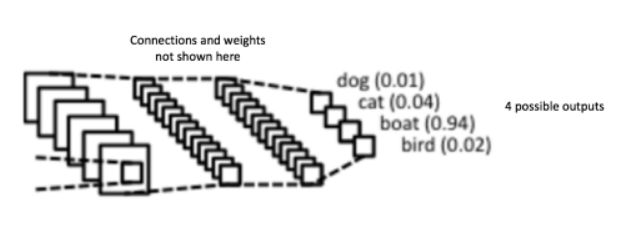
* **Fully connected layer**

The Fully Connected layer is a traditional Multi-Layer Perceptron that uses a softmax activation function in the output layer (other classifiers like SVM can also be used). The term “Fully Connected” implies that every neuron in the previous layer is connected to every neuron on the next layer

The purpose of the fully connected layer is to use output features of the flattening layer for classifying the input image into various classes based on the training dataset.

* **Output layer**

The last layer of the CNN that has m artificial neurons form classes, at the end of classification each neuron will have a probability that the input image belongs to its class.



**Figure 13- fully connected layer and output layer**

The sum of output probabilities from the output Layer is 1. This is ensured by using the [Softmax](http://cs231n.github.io/linear-classify/" \l "softmax" \t "_blank) as the activation function in the output layer

## 2.3 Related works

Many recent works [2, 7, 13, 16, 17, 27, 12, 6] have proposed models for VQA. We compare and relate our proposed co-attention mechanism to other vision and language attention mechanisms in literature.

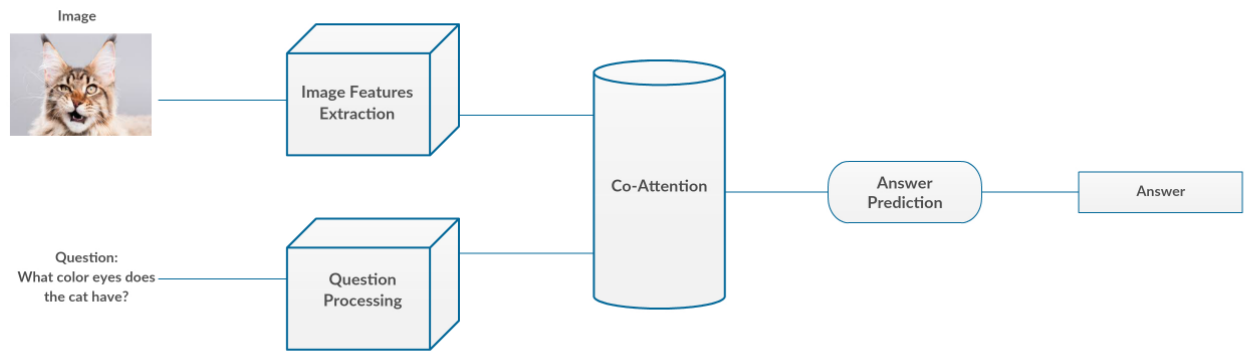
**Image Attention**: Instead of directly using the holistic entire-image embedding from the fully-connected layer of a deep CNN (as in [2, 15–17]), a number of recent works have explored image attention models for VQA. Zhu et al. [28] add spatial attention to the standard LSTM model for pointing and grounded QA. Andreas et al. [1] propose a compositional scheme that consists of a language parser and a number of neural modules networks. The language parser predicts which neural module network should be instantiated to answer the question. Some other works perform image attention multiple times in a stacked manner. In [25], the authors propose a stacked attention network, which runs multiple hops to infer the answer progressively. To capture fine-grained information from the question, Xu et al. [24] propose a multi-hop image attention scheme. It aligns words to image patches in the first hop, and then refers to the entire question for obtaining image attention maps in the second hop. In [20], the authors generate image regions with object proposals and then select the regions relevant to the question and answer choice. Xiong et al. [23] augments dynamic memory network with a new input fusion module and retrieves an answer from an attention based GRU. In concurrent work, [5] collected ‘human attention maps’ that are used to evaluate the attention maps generated by attention models for VQA. Note that all of these approaches model visual attention alone, and do not model question attention. Moreover, [24, 25] model attention sequentially, i.e., later attention is based on earlier attention, which is prone to error propagation. In contrast, we conduct co-attention at three levels independently.

**Language Attention:** Though no prior work has explored question attention in VQA, there are some related works in natural language processing (NLP) in general that have modeled language attention. In order to overcome difficulty in translation of long sentences, Bahdanau et al. [3] propose RNN Search to learn an alignment over the input sentences. In [9], the authors propose attention model to circumvent the bottleneck caused by fixed width hidden vector in text reading and comprehension. A more fine-grained attention mechanism is proposed in [18]. The authors employ a word-by-word neural attention mechanism to reason about the entailment in two sentences. Also focused on modeling sentence pairs, the authors in [26] propose an attention-based bigram CNN for jointly performing attention between two CNN hierarchies. In their work, three attention schemes are proposed and evaluated. In [19], the authors propose a two-way attention mechanism to project the paired inputs into a common representation space.

# Analysis and Design

## 3.1 System Overview

### 3.1.1 System Architecture



**Figure 14-System Architecture**

There are four main modules in the system, Image Features Extraction, Question Processing, Co-Attention and Answer Prediction. The system has two inputs the image and the question, and one output the answer.

### 3.1.2 Modules Description

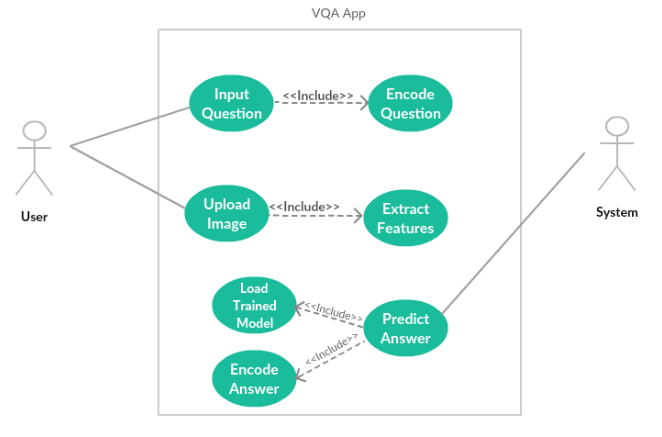
* **Image Features Extraction:** The aim of this module is to extract the features of the input image using VVG19 Convolutional Neural Network.
* **Question Processing:** This module is responsible to clean the question, and then extract its features using Question Hierarchy technique (will be illustrated later).
* **Co-Attention:** This module is responsible to combine the features extracted from both image and question.
* **Answer Prediction:** The aim of this module is to predict the answer based on the co-attended image and question features.

### 3.1.3 System Users

1. *System Characteristics:*
   1. The user can ask about most topics (objects, colors, numbers, locations, etc).
2. *User Characteristics*
   1. Familiar with computer Applications.
   2. Read the guidelines well.
   3. Follow the instruction and the same like picture.

## 3.2 System Analysis & Design

### 3.2.1 Use Case Diagram

******

**Figure 15-Use case Diagram**

# 4- Implementation and Testing

## 4.1 Implementation

In this chapter, we talk about the implementation phases for each module, you will see a detailed explanation about how the entire deep learning model work for each phase and the intuition behind each phase with some mathematical explanation.

We also mentioned a lot of references that we have used during the implementation phases, so, you can check them if you want to find more details about something not clear in the illustration.

Our model trying merging between natural language processing represented in question hierarchy and convolution neural network using co-attention, we will begin with an explanation of question hierarchy then convolution neural network for an image then co-attention between them.

### 4.1.1 Preprocessing

At the preprocessing, we start by removing the unnecessary characters that have no value during the training process like backslash or comma, also we return some of the abbreviations to its original for example " I'm " to " I am", the reason for doing this is to make the training faster and to consider just to the words that make difference in the training.

In the second part, we start counting each word in the data-set and assign a unique number for the top thousand words and the other be assigned to one unique number which relevant to "UNK".

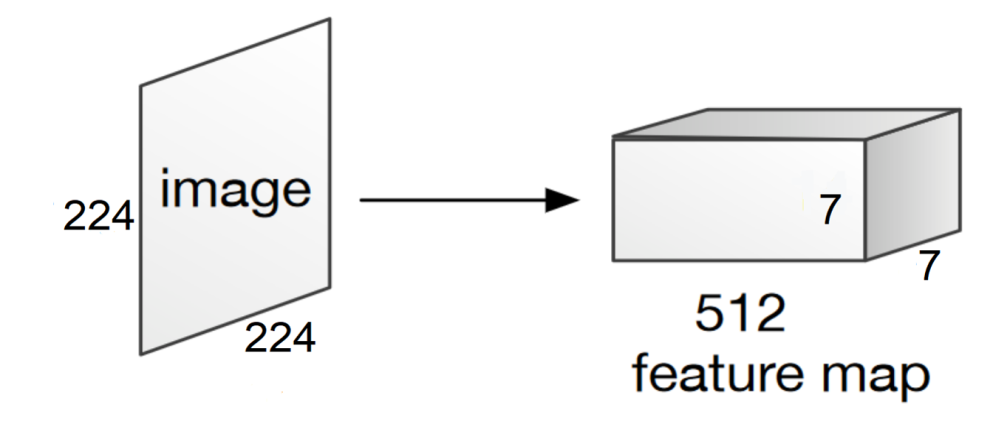
For every word in a given question we replace it with its unique id that has been assigned to and finally we pad the sequence to zeros to make every question has the same number of a maximum number of words so the training be more flexible. For every answer to a question in the data-set, we replace it with one hot encoding through all category types.

### 4.1.2 Convolution neural networks

* + There is several CNN architectures: AlexNet, VGG, GoogLeNet, ResNet etc.
  + In our model we will use **VGG19[1]**

****

**Figure 16-VGG19**

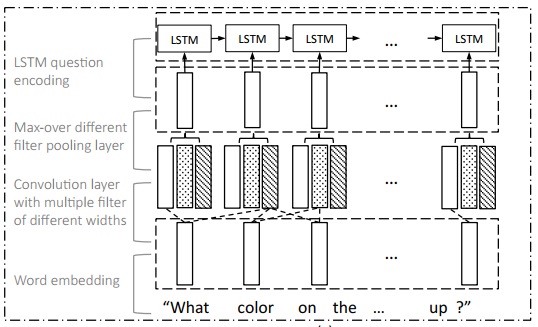
* We resize the input image into **224x224**.
* We extract the features of the input image from the **last pooling layer**.
* The last pooling layer has dimensions of **7x7x512**.

**Figure 17-output of the last pooling layer**

* We consider that the image has **49 regions** as 7x7 dimension according to the first and the second dimension of the output 7x7x512.
* Each region has **feature vector of length 512** as the third dimension of the output 7x7x512.
* So, we have **49 regions** each one has feature vector of **length 512**.

### 4.1.3 Question hierarchy

Question hierarchy is a model architecture that co-attends to the image and question at three levels: word level, phrase level and question level.



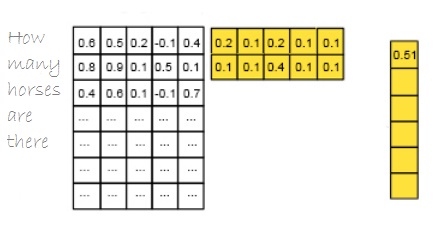
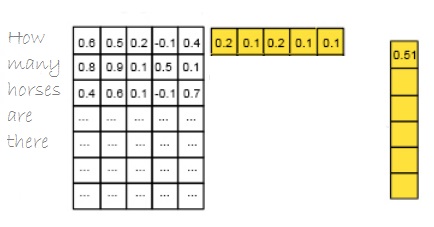
**Figure 18-Question hierarchy**

1. **Word level**

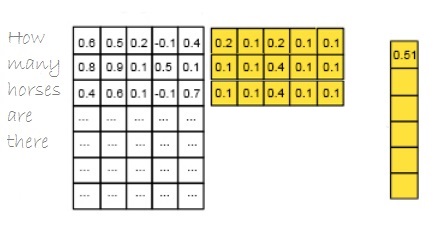
At the word level, we apply one hot encoding for the most frequent words in a given data-set which is a traditional way of representing words, a vector with only one target element being 1 and the others being 0, then generate word embedding vector for each word.

1. **Phrase level**

At the phrase features, we apply 1-D convolution on the word embedding vectors. Concretely, at each word location, we compute the inner product of the word vectors with filters of three window sizes: unigram, bigram and trigram, then combine the various n-gram responses by pooling them into a single phrase level representation.



1. (b)



(c)

**Figure 19– (a) unigram, (b) bi gram, (c) tri gram**

For the *t*-th word, the convolution output with window size *s* is given by

ph.jpg

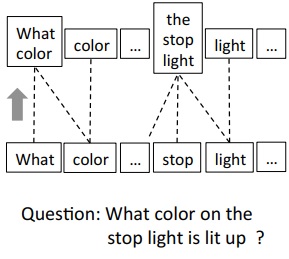
**Eq. 1**

The word-level features are appropriately 0-padded before feeding into bigram and trigram convolutions to maintain the length of the sequence after convolution, given the convolution result, we then apply max-pooling across different n-grams at each word location to obtain phrase-level features.

**ph.jpg**

**Eq. 2**

The reason for applying phrase level is to pick the most important phrase features from the given question, for each word we obtain the most important feature according to its neighboring word, for example the best representation for the word “what” is “what color" and the word “stop" is "the stop light" .

****

**Figure 20- phrase level**

The output of this level will be the input for the last level in question hierarchy "LSTM recurrent neural network".

1. **Question level**

At the question level, we use recurrent neural networks to encode the entire question. For each level of the question representation in this hierarchy, we construct joint question and image co-attention maps, which are then combined recursively to ultimately predict a distribution over the answers.

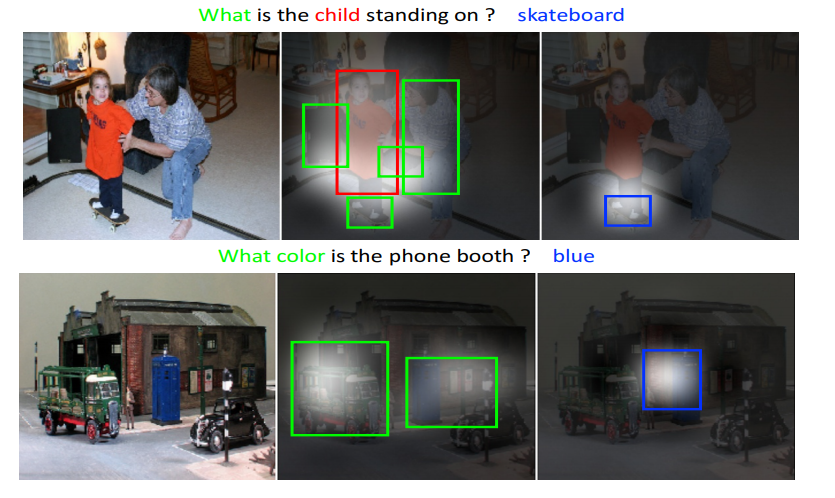
### 4.1.4 Attention

So now we have question features extracted from question hierarchy phase, and image features extracted from VGG19 Convolution Neural Network. But how to combine information from different domains (text and image)? Here comes the role of the Attention phase. Attention has been widely used in image or video captioning, and now it comes to VQA. With it, we can now focus on specific regions of an image base on the question which correspond to how a human understand an image and vice versa.

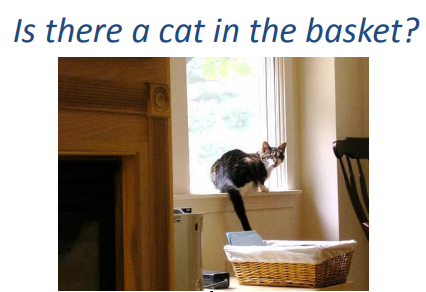
In this section you will see a detailed explanation for attention on the image which called visual attention, attention on the question and the co-attention which consider as a combination of both (visual and question attentions) and this is what we use in our project.

1. **Visual Attention**

This attention mechanism typically produces a spatial map highlighting image regions relevant to answering the question, where answering visual questions often involves looking at different spatial regions and comparing their contents and/or locations. For example, to answer the questions in Figure 21, we need to look at a portion of the image, such as the child or the phone booth. Similarly, to answer the question “Is there a cat in the basket?” in Figure 22, we can first find the basket and the cat objects, and then compare their locations.

******

**Figure 21-Describes the visual attention.**



**Figure 22-visual attention**

1. **Question Attention**

So far, most of the previous attention models for VQA have focused on the problem of identifying “where to look” or visual attention. In this model, the problem of identifying “which words to listen to” or question attention is equally important. Consider the questions about the image in figure 23 “how many horses are in this image?” and “how many horses can you see in this image?". They have the same meaning, essentially captured by the first three words. A machine that attends to the first three words would arguably be more robust to linguistic variations irrelevant to the meaning and answer of the question.



**Figure 23 - Question attention**

1. **Co-Attention**

The co-attention mechanism is a combination of visual attention and question attention. The idea of the co-attention is that the model has a natural symmetry between the image and question, in the sense that the image representation is used to guide the question attention and the question representations are used to guide image attention.

There are two co-attention mechanisms that differ in the order in which image and question attention maps are generated. The first mechanism, which called parallel co-attention, generates image and question attention simultaneously. The second mechanism, which called alternating co-attention, sequentially alternates between generating image and question attentions. These co-attention mechanisms are executed at all three levels of the question hierarchy.

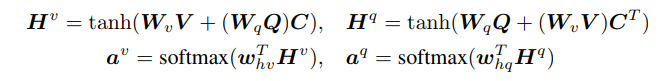
1. **Parallel Co-Attention**

Parallel co-attention attends to the image and question simultaneously. We connect the image and question by calculating the similarity between image and question features at all pairs of image-locations and question-locations as shown in figure 24. Specifically, given an image feature map ∈, and the question representation , we calculate the affinity matrix by



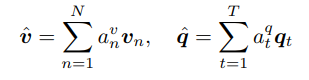
***Eq. 3***

where contains the weights, is the image representation dimension, is the number of the spatial locations on the image and is the number of words in the question. After computing this affinity matrix, one possible way of computing the image (or question) attention is to simply maximize out the affinity over the locations of other modality, i.e. and . Instead of choosing the max activation, the best way to improve the performance is to consider this affinity matrix as a feature and learn to predict image and question attention maps via the following:



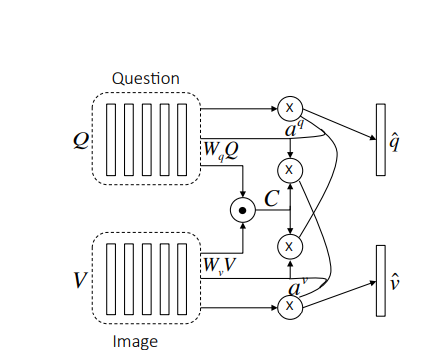
***Eq. 4***

where , are the weight parameters and k is a constant number. and are the attention probabilities of each image region and word respectively. The affinity matrix C transforms question attention space to image attention space (vice versa for ). Based on the above attention weights, the image and question attention vectors are calculated as the weighted sum of the image features and question features.



***Eq. 5***

The parallel co-attention is done at each level in the hierarchy, leading to and where r ∈ {word level, phrase level, question level}.



**Figure 24-Visualizes the parallel co-attention**

1. **Alternating Co-Attention**

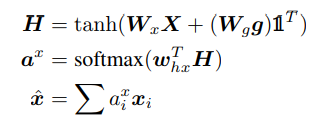
In this attention mechanism, the generating of image and question attention is sequentially alternated. Briefly, this consists of three steps:

1) Summarize the question into a single vector q.

2) Attend to the image based on the question summary q.

3) Attend to the question based on the attended image feature.

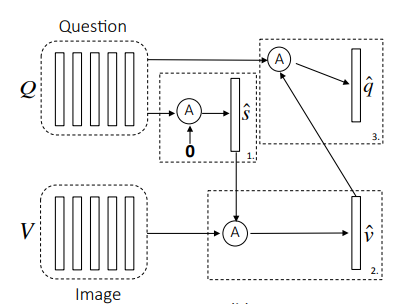
Concretely, the attention operation is defined as , which takes the image (or question) features and attention guidance derived from question (or image) as inputs, and outputs the attended image (or question) vector. The operation can be expressed in the following steps:



***Eq. 6***

where **1** is a vector with all elements to be 1. and are parameters. is the attention weight of feature .

The alternating co-attention process is illustrated in Figure 25. At the first step of alternating co-attention, and is zero. At the second step, where is the image features, and the guidance is intermediate attended question feature from the first step. Finally, the attended image feature is used as the guidance to attend the question again, i.e., and . Similar to the parallel co-attention, the alternating co-attention is also done at each level of the hierarchy.

****

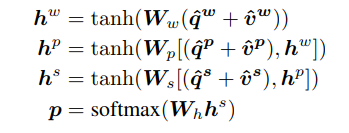
**Figure 25-Visualizes the alternating co-attention**

### 4.1.5 Encoding for Predicting Answers

So far, we have finished the features extraction phase and the co-attention phase. In this section, we will talk about the last phase of the VQA system to get the answer to the question out from the system.

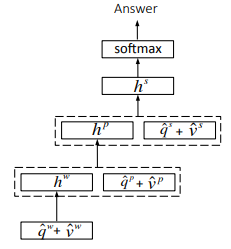
The VQA problem is treated as a classification problem where each answer should be considered as a class. The COCO-QA dataset contains 450 different answers for all questions, so we have 450 unique classes we need to classify each question to one of them.

The answer is predicted based on the co-attended image and question features from all three levels. Then a multi-layer perceptron (MLP) is used to recursively encode the attention features as shown in figure 26.



***Eq. 7***

Where and are the weight parameters. is the concatenation operation on two vectors. p is the probability of the final answer.



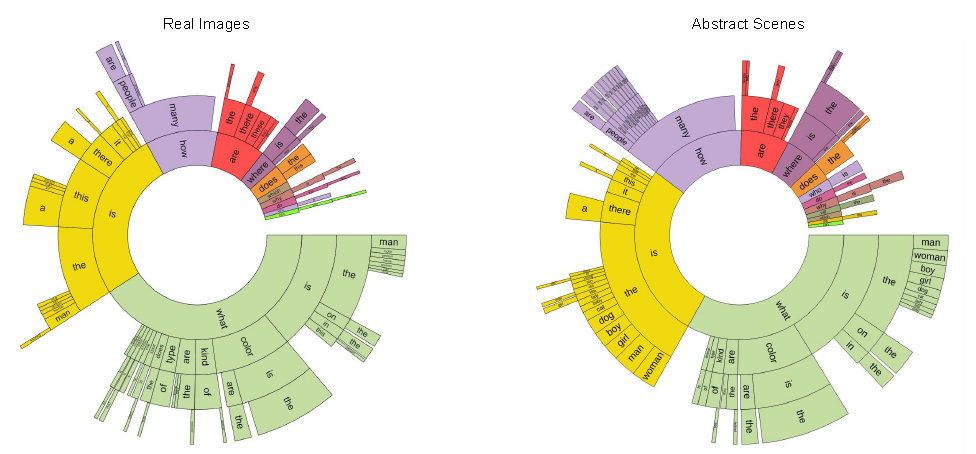
**Figure 26-Visualizes the process of predicting answers**

## 4.2 Testing

4.2.1 Dataset

In this section, we are going to talk about the most common datasets used in VQA and giving clear examples and comparisons.

1. **VQA dataset**
   * **VQA** dataset [2] is the largest dataset for this problem, containing human annotated questions and answers on Microsoft COCO dataset [14]. The dataset contains 248,349 training questions, 121,512 validation questions, 244,302 testing questions, and a total of 6,141,630 question-answers pairs. There are three sub-categories according to answer-types including yes/no, number, and other. Each question has 10 free-response answers.

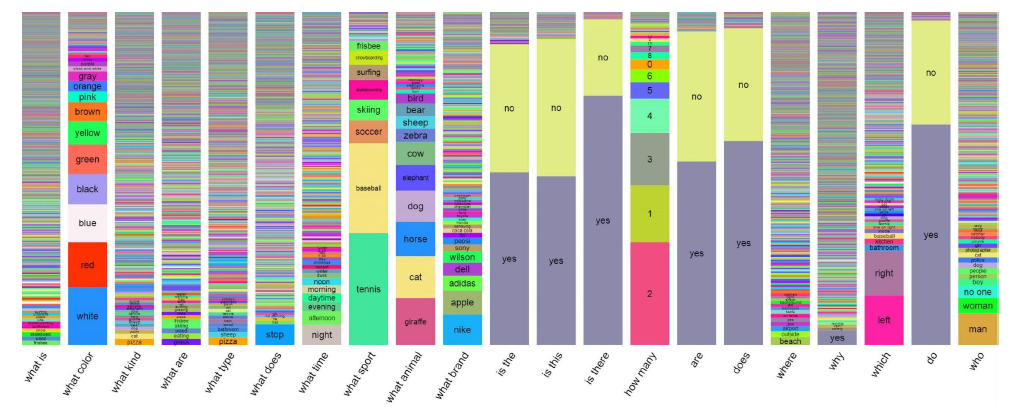


**Figure 27-Distribution of questions by their first four words**

|  |  |
| --- | --- |
| **Question**  **type** | **Percentage** |
| What is | 13.84 |
| What color | 8.98 |
| What kind | 2.49 |
| What are | 2.32 |
| What type | 1.78 |
| Is the | 10.16 |
| Is this | 8.26 |
| How many | 10.28 |
| Are | 7.57 |
| Does | 2.75 |
| Where | 2.90 |
| Is there | 3.60 |
| Why | 1.20 |
| Which | 1.21 |
| Do | 1.15 |
| What does | 1.12 |
| What time | 0.67 |
| Who | 0.77 |
| What sport | 0.81 |
| What animal | 0.53 |
| What brand | 0.36 |

***Table 1 - The percentage of questions for real images***

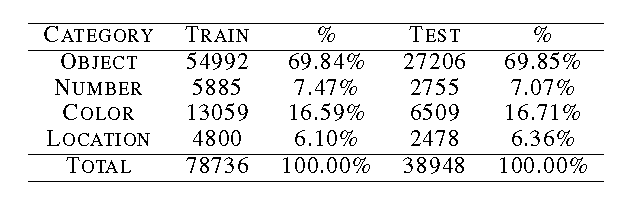
* + **Types of Question.** Given the structure of questions generated in the English language, we can cluster questions into different types based on the words that start the question. **Figure 27** shows the distribution of questions based on the first four words of the questions for both the real images (left) and abstract scenes (right). Interestingly, the distribution of questions is quite similar for both real images and abstract scenes. This helps demonstrate that the type of questions elicited by the abstract scenes is similar to those elicited by the real images. There exists a surprising variety of question types, including “What is…”, “Is there…”, “How many…”, and “Does the…”.Quantitatively, the percentage of questions for different types is shown in **Table 1**. A particularly interesting type of question is the “What is…” questions, since they have a diverse set of possible answers.



**Figure 28-Distribution of answers per question type**

* + **Typical Answers.** Figure **28** shows the distribution of answers for several question types. We can see that a number of question types, such as “Is the. . . ”, “Are. . . ”, and “Does. . . ”are typically answered using “yes” and “no” as answers. Other questions such as “What is. . . ” and “What type. . . ” have a rich diversity of responses. Other question types such as “What color. . . ” or “Which. . . ” have more specialized responses, such as colors, or “left” and “right”.
  + **Lengths**. Most answers consist of a single word, with the distribution of answers containing one, two, or three words, respectively being 89:32%, 6:91%, and 2:74% for real images and 90:51%, 5:89%, and 2:49% for abstract scenes.

1. **COCO-QA dataset**
   * **COCO-QA** dataset [17] is automatically generated from captions in the Microsoft COCO dataset [14]. There are 78,736 train questions and 38,948 test questions in the dataset. These questions are based on 46,236 and 24,648 images respectively. There are four types of questions including object, number, color, and location. Each type takes 70%, 7%, 17%, and 6% of the whole dataset, respectively. All answers in this data set are single word.

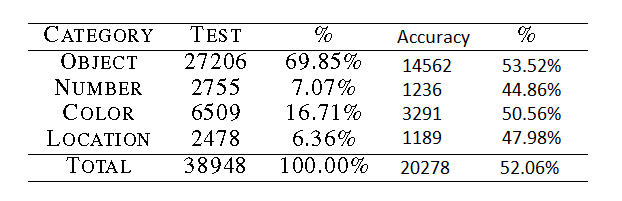


**Table 2 - The statistics of COCO-QA.**

* + The maximum question length is 55, and average is 9.65. The most common answers are “two” (3116, 2.65%), “white” (2851, 2.42%), and “red” (2443, 2.08%). The least common are “eagle” (25, 0.02%) “tram” (25, 0.02%), and “sofa”(25, 0.02%). The median answer is “bed” (867, 0.737%). Across the entire test set (38,948 QAs), 9072 (23.29%) overlap in training questions, and 7284 (18.70%) overlap in training question-answer pairs.
* In our project we used **COCO-QA** dataset[2]

4.2.2 Results

We achieved 52.06% accuracy on the testing dataset that covers 20,278 samples of the total 38,948 testing samples.



**Table 3 - Testing statistics**

We enhanced our model to achieve better accuracy but we would not able to retrain the model because of the resources unavailability.

# 5-User Manual

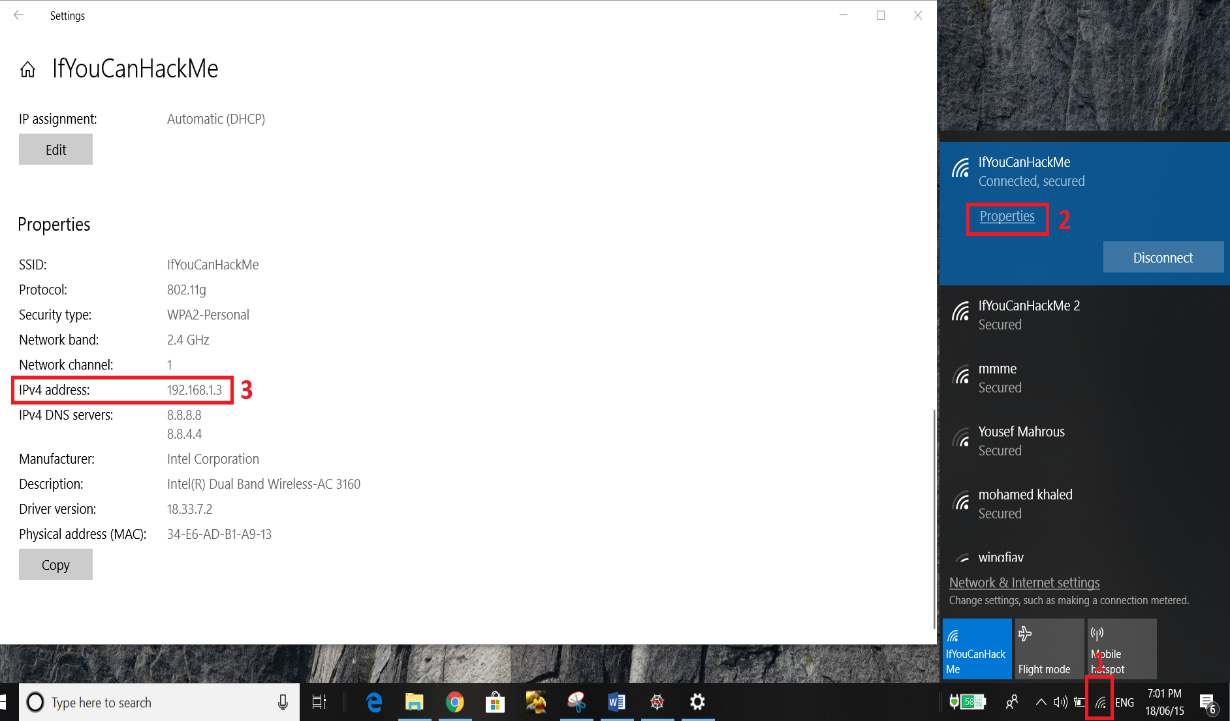
In this chapter, we provide a detailed description of how to be able to use our VQA android application.

We used a local host technique to host our python backend on the laptop, and then we make a communication between the Android interface and the python backend using Flask web framework.

1. **Create a network**

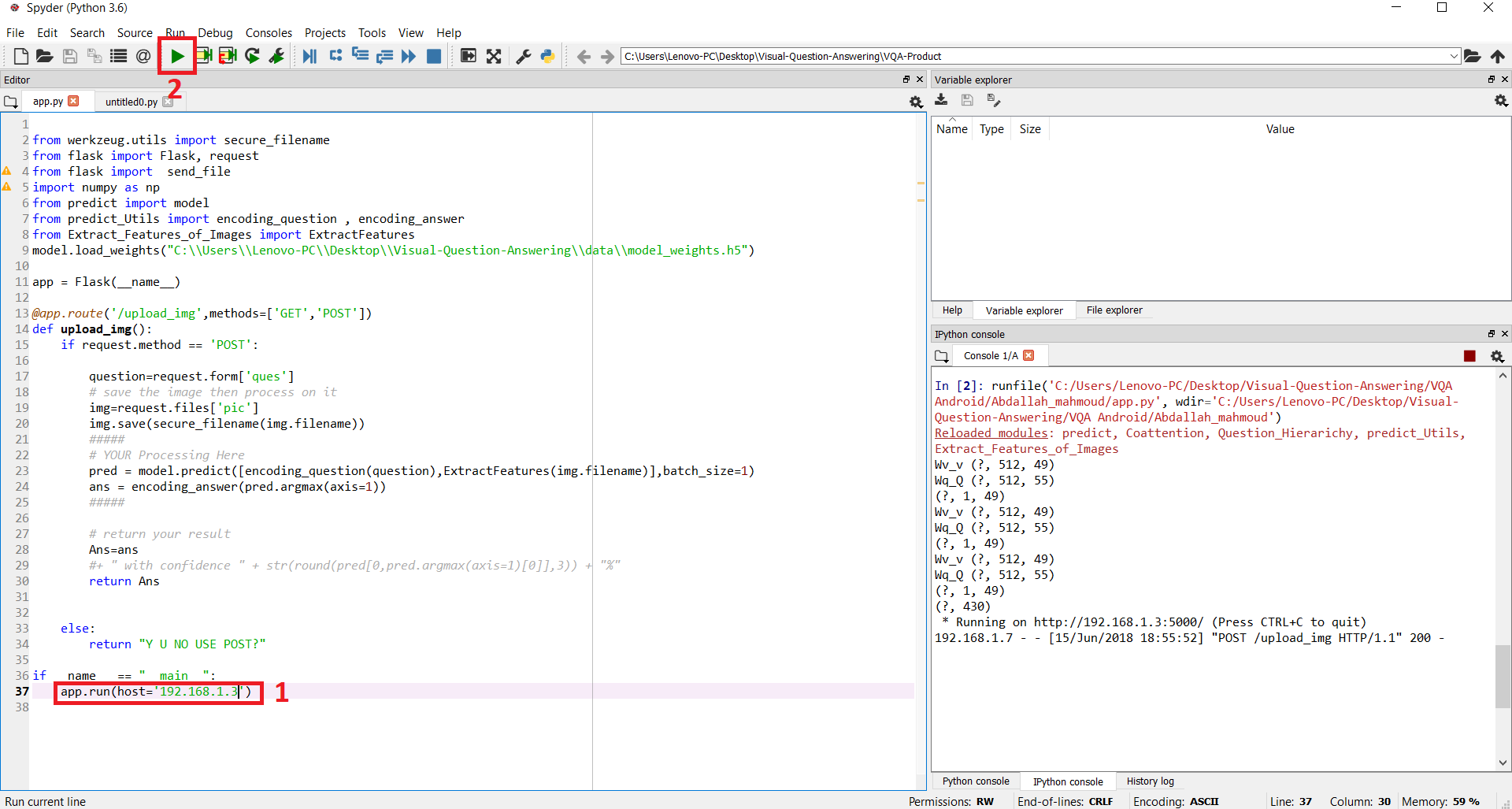
* make a network between your mobile and the laptop using possible two ways:
  1. **Wi-Fi connection**
     + Make your mobile and laptop connect on the same router.
  2. **mobile hotspot**
     + Open your mobile hotspot and connect on it using your laptop.

1. **Extract the network IP address**

****

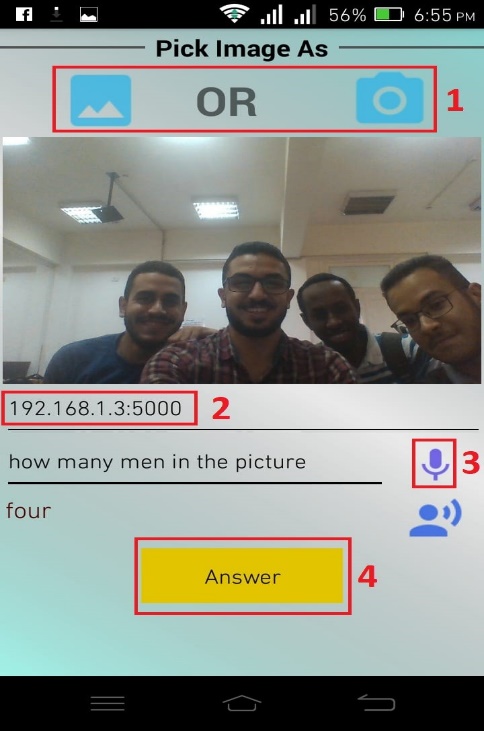
**Figure 29- Extract the network IP address**

1. Click on the Wi-Fi icon.
2. Open Wi-Fi properties.
3. Get Your IP address.
4. **Run the localhost**

****

**Figure 30 - Run the localhost**

1. Put your network IP in this line of the code (**app.py**).
2. Run the **app.py** file.
3. **Run the android application**

****

**Figure 31- Android app sample input**

1. Pick an image from your gallery or using your camera.
2. Put the network IP followed by “**:5000**” as shown in **Figure 31.**
3. Click on the microphone icon to pronounce your question or manually write it in the input textbox.
4. Click on “**Answer**” button to achieve the answer.

# 6-Conclusion and Future Work

## 6.1 Conclusion

In our system, we tried to make an effective application to answer an image related to a given question by merging between natural language processing and convolution neural network. We got an overall accuracy 52.06% accuracy on the testing dataset that covers 20,278 sample of the total 38,948 testing samples.

Making such a system helps us to learn many new concepts and understanding too many new basics in deep learning and natural language processing.

## 6.2 Future Work

Our future work includes:

* Deal with more different types of questions.
* Deal with low quality of images.
* Generate complete answers using encoder decoder model.
* Reduce the overall time complexity of the whole application.
* Increase the classifier training data.
* Try different methods of attentions.
* Try using Dynamic memory network for visual and textual answering.
* Visualizing the attention between the Image and the question.
* Publishing the application on Google store and Apple store.

# References

**[1]** Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein.Deep compositional question answering with neural module networks.In CVPR, 2016.

**[2]** Stanislaw Antol, AishwaryaAgrawal, Jiasen Lu, Margaret Mitchell, DhruvBatra, C Lawrence Zitnick, and Devi Parikh.Vqa: Visual question answering. In ICCV, 2015.

**[3]**DzmitryBahdanau, Kyunghyun Cho, and YoshuaBengio.Neural machine translation by jointly learning to align and translate.In ICLR, 2015.

**[4]** R. Collobert, K. Kavukcuoglu, and C. Farabet. Torch7: A matlab-like environment for machine learning. In BigLearn, NIPS Workshop, 2011.

**[5]**Abhishek Das, Harsh Agrawal, C Lawrence Zitnick, Devi Parikh, and DhruvBatra. Human attention in visual question answering: Do humans and deep networks look at the same regions? arXiv preprint arXiv:1606.03556, 2016.

**[6]** Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach.Multimodal compact bilinear pooling for visual question answering and visual grounding.arXiv preprint arXiv:1606.01847, 2016.

**[7]**HaoyuanGao, 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. In NIPS, 2015.

**[8]**Kaiming He, Xiangyu Zhang, ShaoqingRen, and Jian Sun. Deep residual learning for image recognition.In CVPR, 2016.

**[9]** Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, LasseEspeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom.Teaching machines to read and comprehend.In NIPS, 2015.

**[10]**Baotian Hu, Zhengdong Lu, Hang Li, and Qingcai Chen. Convolutional neural network architectures for matching natural language sentences.In NIPS, 2014.

**[11]**IlijaIlievski, Shuicheng Yan, and JiashiFeng.A focused dynamic attention model for visual question answering.arXiv:1604.01485, 2016.

**[12]**Jin-Hwa Kim, Sang-Woo Lee, Dong-Hyun Kwak, Min-Oh Heo,Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Multimodal residual learning for visual qa.arXiv preprint arXiv:1606.01455, 2016.

**[13]**Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, YannisKalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. arXiv preprint arXiv:1602.07332, 2016.

**[14]**Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, PietroPerona, Deva Ramanan, PiotrDollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In ECCV, 2014.

**[15]** Lin Ma, Zhengdong Lu, and Hang Li. Learning to answer questions from image using convolutional neural network.In AAAI, 2016.

**[16]** Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. Ask your neurons: A neural-based approach to answering questions about images. In ICCV, 2015.

**[17]**MengyeRen, Ryan Kiros, and Richard Zemel. Exploring models and data for image question answering. In NIPS, 2015.

**[18]** Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, TomášKoˇcisk`y, and Phil Blunsom.Reasoning about entailment with neural attention.In ICLR, 2016.

**[19]** Cicero dos Santos, Ming Tan, Bing Xiang, and Bowen Zhou. Attentive pooling networks. arXiv preprint arXiv:1602.03609, 2016.

**[20]** Kevin J Shih, Saurabh Singh, and Derek Hoiem. Where to look: Focus regions for visual question answering. In CVPR, 2016.

**[21]** Karen Simonyan and Andrew Zisserman.Very deep convolutional networks for large-scale image recognition.CoRR, abs/1409.1556, 2014.

**[22]** Christian Szegedy, Wei Liu, YangqingJia, Pierre Sermanet, Scott Reed, DragomirAnguelov, DumitruErhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions.In CVPR, 2015.

**[23]**CaimingXiong, Stephen Merity, and Richard Socher. Dynamic memory networks for visual and textual question answering. In ICML, 2016.

**[24]**HuijuanXu and Kate Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. arXiv preprint arXiv:1511.05234, 2015.

**[25]**Zichao Yang, Xiaodong He, JianfengGao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In CVPR, 2016.

**[26]**Wenpeng Yin, HinrichSchütze, Bing Xiang, and Bowen Zhou.Abcnn: Attention-based convolutionalneural network for modeling sentence pairs. In ACL, 2016.

**[27]**Peng Zhang, YashGoyal, Douglas Summers-Stay, DhruvBatra, and Devi Parikh. Yin and yang: Balancing and answering binary visual questions. arXiv preprint arXiv:1511.05099, 2015.

**[28]**Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In CVPR, 2016.

**[29]** C Lawrence Zitnick, AishwaryaAgrawal, Stanislaw Antol, Margaret Mitchell, DhruvBatra, and Devi Parikh. Measuring machine intelligence through visual question answering. AI Magazine, 37(1), 2016.