DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

A Dissertation Report on

Visual Question Answering

Submitted by

Anurag Srivastava	1MS14CS020
Chaitra H	1MS14CS030
Brinda V Eshwar	1MS14CS028
Abhijeet Shankar	1MS14CS005

in partial fulfillment for the award of the degree of

Bachelor of Engineering in Computer Science & Engineering

Under the guidance of

Dr. Shilpa Chaudhari
Associate Professor

M.S.RAMAIAH INSTITUTE OF TECHNOLOGY
(Autonomous Institute, Affiliated to VTU)
BANGALORE-560054
2017-2018, www.msrit.edu,

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

A Dissertation Report on

Visual Question Answering

Submitted by

Anurag Srivastava	1MS14CS020
Chaitra H	1MS14CS030
Brinda V Eshwar	1MS14CS028
Abhijeet Shankar	1MS14CS005

in partial fulfillment for the award of the degree of

Bachelor of Engineering in Computer Science & Engineering

Under the guidance of

Dr. Shilpa Chaudhari
Associate Professor

M.S.RAMAIAH INSTITUTE OF TECHNOLOGY
(Autonomous Institute, Affiliated to VTU)
BANGALORE-560054
2017-2018, www.msrit.edu,

Ramaiah Institute of Technology

(Autonomous Institute, Affiliated to VTU)
BANGALORE-560054

Department of Computer Science & Engineering

CERTIFICATE

This is to certify that the project work titled **Visual Question Answering** is a bonafide work carried out by **1MS14CS020 – Anurag Srivastava, 1MS14CS030 – Chaitra H, 1MS14CS028 – Brinda V Eshwar** and **1MS14CS005 – Abhijeet Shankar** in partial fulfillment for the award of degree of Bachelor of Engineering in Computer Science and Engineering during the year 2018. The Project report has been approved as it satisfies the academic requirements with respect to the project work prescribed for Bachelor of Engineering Degree. To the best of our understanding the work submitted in this report has not been submitted, in part or full, for the award of said degree.

Signature of the Guide	**Signature of the HOD**
Dr. Shilpa Chaudhari	Dr. Anita Kanavalli

External Examiners

Name of the Examiners: Signature

1.

2.

DECLARATION

We Students of Eighth semester BE, Dept. of Computer Science and Engineering, Ramaiah Institute of Technology, Bangalore, hereby declare that the project entitled "**Visual Question Answering",** report is completed and written by us under the guidance of **Dr. Shilpa Chaudhari,** Dept. of CSE, Bangalore for the partial fulfillment of the requirements for the award of the degree of Bachelor of Engineering.

Place: Ramaiah Institute of Technology
Date: 05 May 2018
(1MS14CS020 Anurag Srivastava)
(1MS14CS030 Chaitra H)
(1MS14CS028 Brinda V Eshwar)
(1MS14CS005 Abhijeet Shankar)

ACKNOWLEDGEMENT

First and foremost, our utmost gratitude to Dr. Shilpa Chaudhari, Dept. of CSE, MSRIT whose sincerity and encouragement we will never forget. She has been our inspiration as we overcame all the obstacles in the completion of this project work.

Dr. Anita Kanavalli, Head of the Department of Computer Science and Engineering, had kind concern and consideration regarding project work and we would like to thank her for continues support.

We would like to thank our beloved principal Dr. N.V.R Naidu for his support and encouragement.

This work would not have been possible without the guidance and help of several individuals who in one way or another contributed their valuable assistance in preparation and completion of this study.

We would like to express sincere thanks to all the teaching and non-teaching faculty of CSE Department and our dear friends who helped in all the ways while preparing the Report.

ABSTRACT

We propose a contemporary attention based deep learning framework for visual question answering task (VQA). Given an image and a question pertaining to the image, VQA seeks to apply deep learning tools and neural networks to enable the computer returning a natural language answer to the asked question. Since different questions inquire about the attributes of distinct image regions, generating answers with higher accuracy will require the model to have question guided attention, i.e., the attention on the regions analogous to the input question's intent. We propose an attention-based adaptive convolutional neural network to recognize the question-guided attention based on input queries. The CNN first deduces the attention regions by detecting the corresponding visual features in the visual feature maps with a configurable convolution operation. With the support of the question-guided attention, this CNN can attain both higher VQA accuracy and superior understanding of the visual question answering procedure. We assess the CNN architecture on multiple benchmark VQA datasets: COCO-QA, VQA datasets and its subsets. This refined model is poised to achieve significant improvements over the antecedent methods.

Contents

1. INTRODUCTION

1.1 General Introduction

VQA (Visual Question Answering) demands an image and an associated question (text or verbal), to which the system must attempt to determine the closest correct answer or an accurate response. This procedure comprises the disciplines of computer vision and natural language processing, as it vital to have the apprehension of the question as well as the parsing of the perceptible components of the image in consideration. VQA is a pragmatic experiment to assess deep learning and visual understanding, and contemplate the objectives of the domain of computer vision. Deep visual understanding can also be interpreted as the potential of algorithms and techniques to obtain information of significance from images and to perform some sort of reasoning or calculations based on the obtained information.

1.2 Problem Statement

This project is an attempt to apply deep learning tools and neural networks to enable a computer to answer questions by looking at images pertaining to the question. The answers can be in various forms such as a phrase, a word, a yes/no answer, selection from multiple choice or fill in the blank. Adding a voice interface facilitates the usability and the accessibility of the project and helps it connect better with the user group.

In our project, we think that it is not enough to only consider image attention, but it is essential to consider image attention based on question. Specifically, not only are we focusing on which region in the image we should look at, but which word and which POS tag feature in the sentence is of interest. So in this project, we are thinking of first implementing a baseline MLP model and then improving the model by changing the loss function.

1.3 Objectives of the Project

- To survey on the potential and significance of the ability of a computer to apprehend human language, respond to queries and examine images that can pave the way for computer sentience.
- To identify the miscellaneous parameters, variables and tools required for accurate language processing and precise image examination based on the surveyed information. Also identifying techniques to apprehend questions and arrive on correct answers.
- To analyze the various requirements for the efficient development of a software model to solve the proposed problem using the identified parameters.
- To design techniques and algorithms and develop the Visual Question Answering system according to the analyzed requirements.
- To test the functionality and efficiency of the developed system.

1.4 Project Deliverables

- A developed and trained system which is able to solve the proposed problem with satisfactory efficiency.
- The project will be delivered in the form of a software model.

1.5 Current Scope

Due to the rapid development of deep learning methods for visual recognition, natural language processing, computers could perform various complex and difficult tasks. One of the most important tasks is to have computer combining various tools for high-level scene interpretation, such as image captioning and visual question answering. With the emergence of large image dataset, text, questions, visual question answering by computers has been made possible. In general, the visual question answer system are required to answer all kinds of questions people might ask relate or not related with the image. Building a system that could properly answer questions would be important to the development of artificial intelligence.

1.6 Future Scope

Existing VQA benchmarks are not enough to evaluate whether an algorithm has 'solved' VQA. In this section, we discuss future developments in VQA datasets that will make them better benchmarks for the problem.
Future datasets need to be larger. While VQA datasets have been increasing in size and diversity, algorithms do not have enough data for training and evaluation. A small experiment was done where we trained a simple MLP baseline model for VQA employing ResNet-152 image features and skip-thought features for the questions, and we evaluated performance as a function of the amount of training data available on COCO-VQA.

2. LITERATURE SURVEY

2.1 Introduction

Recent advances in computer vision and image processing have guided us to the position where conventional object-recognition benchmarks are now considered to be outdated. However, this advent of technology has also triggered the question of how can one advance from object recognition to visual understanding; to elaborate, how can present recognition systems that yield “words” describing an image or a region in an image shift to systems that are able to yield a more thorough semantic representation of the contents of the image. Since benchmarks have conventionally been a cue for advancements in computer vision, various recent studies have suggested techniques to measure our capabilities to develop such renditions.

2.2 Related Works

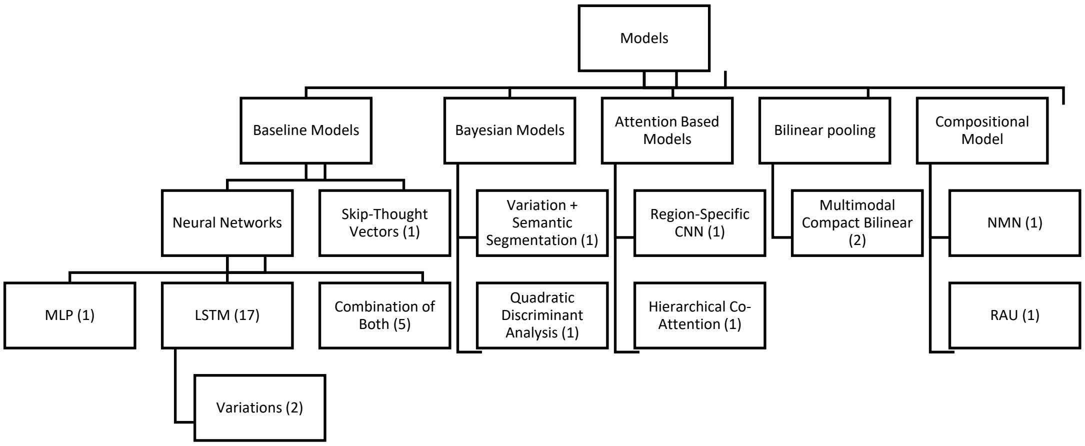
2.2.1 Previous Work:

The recent surge of studies on visual question answering has been fueled by the release of several visual question-answering datasets, most prominently, the VQA dataset, the DAQUAR dataset, the Visual Madlibs Q&A dataset, the Toronto COCO-QA dataset, and the Visual7W dataset. Most of these datasets were developed by annotating subsets of the COCO dataset. Geman proposed a visual Turing test in which the questions are automatically generated and require no natural language processing. Current approaches to visual question answering can be subdivided into “generation” and “classification” models.

a. **Generation models -** Bolaños trained a LSTM model to generate the answer after receiving the image features (obtained from a convolutional network) and the question as input. Wu et al. [30] extends an LSTM generation model to use external knowledge that is obtained from DBpedia. Whilst generation models are appealing because they can generate arbitrary answers (also answers that were not observed during training), in practice, it is very difficult to jointly learn the encoding and decoding models from the question answering datasets of limited size. In addition, the evaluation of the quality of the generated text is complicated in practice.

b. **Classification models -** Zhou studied an architecture in which image features are produced by a convolutional network, question features are produced by averaging word embedding over all words in the question, and a multi-class logistic regressor is trained on the concatenated features; the top unique answers are treated as outputs of the classification model. Similar approaches are also studied by Antol et al. [28] and Wu et al. [39], though they use a LSTM to encode the question text instead of an average over word embedding. Zhu et al. [15] present a similar method but extend the LSTM encoder to include an attention mechanism for jointly encoding the question with information from the image.

2.2.2 Classification of Models:

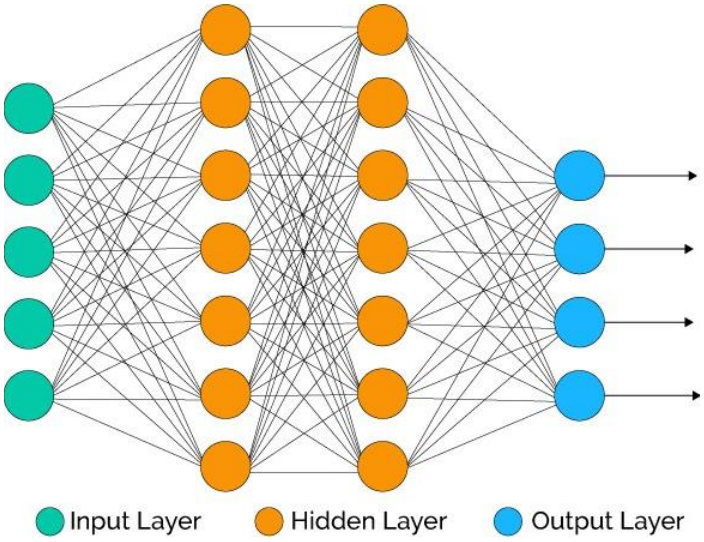


2.3 Baseline Models

Baseline method is one that uses heuristics, simple summary statistics, randomness, or machine learning to create predictions for a dataset. A machine learning algorithm tries to learn and retain a function that prototypes the relationship linking the input data and the target variable. During testing, performance is measured in one way or the other. For example, the algorithm may be 75% accurate, but what does this mean? You can come to a conclusion by comparing with a baseline's performance.

2.4 Neural Networks

An Artificial Neural Network (ANN) is an information processing paradigm which is composed of large number of highly interconnected processing elements working together to solve various specific problems, they learn by examples and is inspired by the way biological nervous systems, such as the brain, process information. Through a continuous learning process an ANN is trained for a specific application, such as pattern recognition or data classification.



Neural networks, with their exceptional ability to extract the meaning from complicated or imprecise data, can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques.. Other advantages may include:

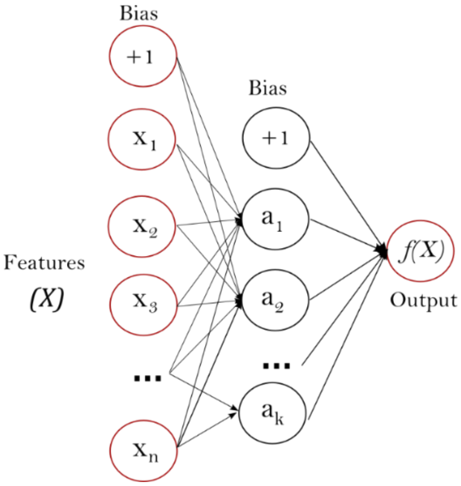
a. Adaptive learning: An ability to learn how to perform tasks based on the given training data.
b. Self-Organization: An ability wherein an ANN can create its own organization or representation of the information it receives during the time of learning.
c. Real Time Operation: ANN computations may be carried out in parallel, and special hardware devices are being designed and manufactured for this purpose.

2.4.1 MLP: Multilayer Perceptron

An MLP is a network of simple neurons called perceptron. In 1958, Rosenblatt introduced the basic concept of a single perceptron. Given multiple real-valued inputs, the perceptron computes a single output by forming a linear combination according to the input weights and then possibly putting the output through some nonlinear activation function. Mathematically this can be written as:

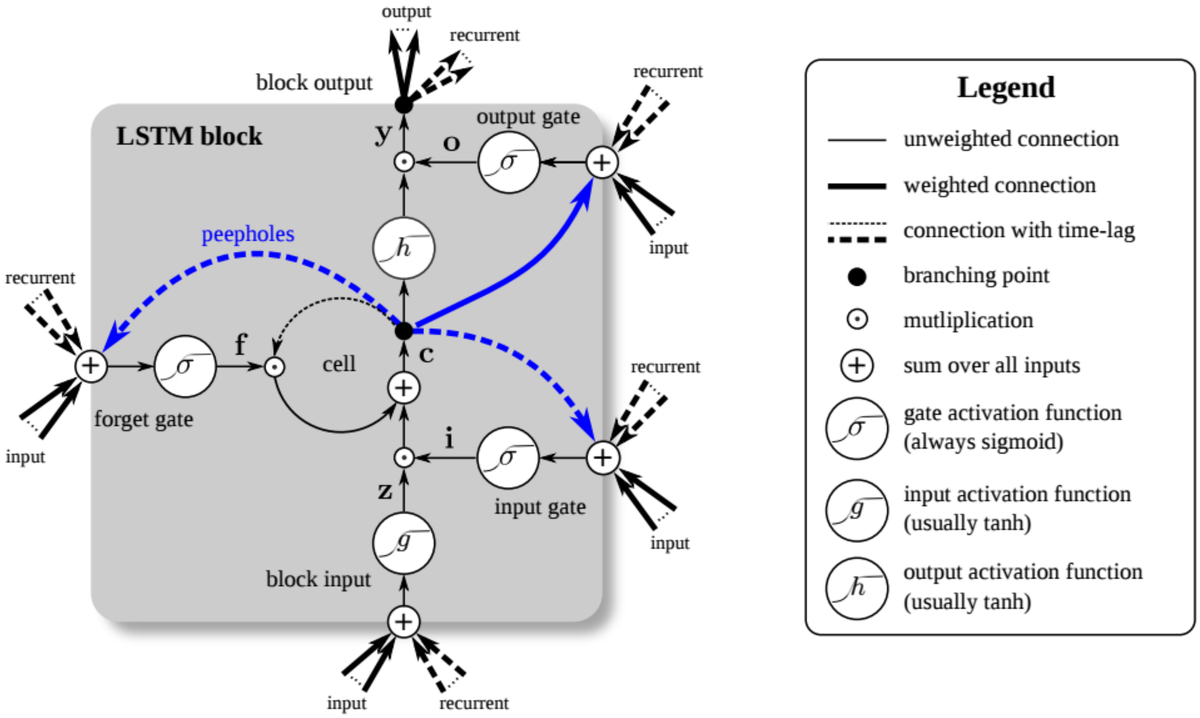
$$y = \varphi\left(\sum_{i=1}^{n} w_i x_i + b\right) = \varphi(w^t x + b)$$

Where w denotes the vector of weights, x is the vector of inputs, b is the bias and φ is the activation function.



An MLP can also be regarded as a logistic regression classifier where the input is first transformed using a learnt non-linear transformationΦ. This transformation projects the input data into a space where it becomes linearly separable. This intermediate layer is referred to as a hidden layer. A single hidden layer is sufficient to make MLPs a universal approximate**.** However, we will later see that there are considerable benefits to using many such hidden layers, i.e. the very premise of deep learning**.**

2.4.2 LSTM



LSTMs help maintain the error which can be back propagated through time and layers. By maintaining a more constant error, they let recurrent nets to continue to memorize over many time steps (over 1000), thereby opening a channel to link causes and effects remotely. LSTMs have information stored outside the normal flow of the recurrent network in a gated cell. Information can be stored in, written to, or read from a cell, more or less like data in a computer's memory. The cell has the authority to decide regarding what to store, and when to allow reads, writes and erasures, via gates that open and close. Unlike the digital storage on computers, however, these gates are analog, lie in the range 0-1and implemented with element-wise multiplication by sigmoid. Analog has the benefit over digital of being differentiable, and therefore suitable for backpropagation. Those gates act on the signals they receive, and identical to the neural network's nodes, they block or pass on information based on its strength and import, which they filter with their own sets of weights. Those weights, like the weights that modulate input and hidden states, are adjusted via the recurrent networks learning process; i.e, the cells learn when to allow data to enter, leave or be deleted through the iterative process of making guesses, back propagating error, and adjusting weights via gradient descent. Different sets of weights filter the input for inputting, outputting and forgetting.

2.4.3 Variations

One popular LSTM variant is adding "peephole connections" which was introduced by Gers & Schmidhuber (2000). It infers that we let the gate levels look at the cell state.
Another variation is to use coupled forget and input gates. Instead of separately deciding what to forget and what we should add new information to, those decisions are made together. We only forget when we're going to input something in its place. We only input new values to the state when we forget something older.

Cho, et al. (2014) introduced a slightly more dramatic variation on the LSTM is the Gated Recurrent Unit (GRU). In this variation, the forget and input gate are combined into a single "update gate." It also merges the cell state and hidden state. The resulting model is much simpler than the standard LSTM models, and this has been growing increasingly popular. These are only few significant LSTM variants. There are many others, like Depth Gated RNNs by Yao, et al. (2015). There is a complete different approach which tackles long-term dependencies, like Clockwork RNNs by Koutnik, et al. (2014).

2.5 Skip-thought Vector

A distributed representations learned by the "skip-thought" architecture is called skip-thought vector. So in simple words a skip-thought is the representation learned by the encoder of the model, which will map a sentence (or sequence of tokens) to a fixed vector. They propose a model for learning high-quality sentence vectors without a particular supervised task in mind. Using word vector learning as inspiration, an objective function was proposed that abstracts the skip-gram model of to the sentence level. They encode a sentence to predict the sentences around it, Instead of using a word to predict its surrounding context. Thus, any composition operator can be replaced by a sentence encoder which leaves only the objective function to be modified.

2.6 Bilinear Pooling

In Visual Question Answering and Visual Grounding, the visual (image) and textual information are represented in separate vectors which are modeled using CNN for image vector and RNN for text vector.

Now these two vectors are combined or pooled using different techniques in order to map the image and text to predict the answers to VQA or predict location in visual grounding. Bilinear pooling enables all the elements of the two vectors to interact with each other in multiplicative ways and is efficient for fine-grained classification for vision only tasks but due to its high dimensionality (n^2), bilinear pooling is not widely used.

2.7 Multimodal Compact Bilinear Pooling

In multimodal compact bilinear model, the problem of high dimensionality is resolved to efficiently and expressively combine the multimodal features (image and text). Due to high dimensionality, the mapping of the two models are not very expressive. Multimodal pooling are implemented using element-wise product or sum or even by concatenating the visual and texting representations. MCB can be used to predict most accurate answers for the VQA task and locations for the Visual Grounding task.

Paper Number	Method	Dataset used	Image Network	Accuracy
3	MCB	VQA real-image dataset +Visual Genome+Visual7W	ResNet-152	MCB(d=16k)with attention:62.50 Multiple choice QA-62.2 Based on dimension(16000):59.83

23	Multi-modal Factorized Bilinear Pooling (MFB) with Co-Attention Learning	VQA	152 layer ResNet	MFB+CoAtt+GloVe+VG : 71.3
36	iBOWIMG	COCO VQA dataset	VGGNet, GoogleNet	61.68
43	Multi-Modal Compact Bilinear Pooling (MCB)	Task Driven Image Understanding Challenge (TDIUC) COCO-VQA	ResNet	84.26

2.8 Compositional Model

Compositional modeling is a new modeling technique which generalizes some of the modeling ideas in Qualitative Process theory. This method provides explicit representations for modeling assumptions to support automatic formulation of models. Compositional modeling is a modern methodology for modeling that generalizes few of the modeling ideas in Qualitative Process theory. Compositional modeling provides explicit representations for modeling assumptions to support automatic formulation of models.

2.8.1 NMN

The neural module network is an interesting compositional approach to VQA task. The main idea being to compose a series of discrete modules (sub networks) that can be executed collectively to answer a given question. Diverse models are used to achieve this, for example the *find(x)* module outputs a heat map for detecting x. To organize the modules, the question is first parsed into a concise expression (called an S-expression), e.g., ‘What is to the right of the car?’ is parsed into *(what car); (what right); (what (*and *car right))*. Using these expressions, modules are composed into a sequence to answer the query.

2.8.2 RAU

The multi-step recurrent answering units (RAU) model for Visual Question Answering task is another state-of-the-art (modern) methodology. Each interface step in RAU consists of a complete answering block that takes an image, a question and the output from the previous LSTM step as inputs. And all of these are a part of larger LSTM network which continuously reasons about a question in a progressive manner.

2.9 Comparing Previous Work based on Model Used

From the time Malinowski et al [23] made an early attempt at solving the VQA task. Since then, VQA task has perceived increasing attention from the computer vision and natural language processing communities. VQA approaches can be classified into the following methodological Categories: the coarse joint-embedding models the fine-grained joint-embedding models with attention and the external knowledge based models.

The most straightforward VQA solutions are coarse joint-embedding models. Image and question are first represented as global features and then integrated to predict the answer. Zhou *et al.* put forth a baseline approach for the VQA task by making use of the concatenation of the image CNN features and the question BoW (bag-of-words) features. Few approaches propose more entangled deep models, e.g., LSTM networks or residual networks, to solve the VQA task in an end-to-end fashion.

2.10 Neural Networks:

2.10.1 Convolutional Neural Network

Extracting information from et al. [2] visual content is one component to answer questions about images. Since the proposal of AlexNet (Krizhevsky, 2012), Convolutional Neural Networks (CNNs) have become dominant and most successful approaches to extract relevant representation from the image. CNNs directly learn the representation from the raw image data and are trained on large image corpora, typically ImageNet (Russakovsky, 2014). Interestingly, after these models are pre-trained on ImageNet, they can typically be adapted for other tasks. This model evaluates how well the most dominant and successful CNN models can be adapted for the Visual Turing Test. Specifically, we evaluate AlexNet (Krizhevsky et al. 2012), VGG (Simonyan and Zisserman, 2014), GoogleNet (Szegedy, 2014).These models, reportedly, achieve more and more accurate results on the ImageNet dataset, and hence, arguably, serve as increasingly stronger models of visual perception.

Serial No.	Paper	Data Set	Image Network	Accuracy
1.	Age and Gender Recognition in the wild with deep attention.	Morph – II + Adience	VGGNet - 16	-
2.	Visual Madlibs: Fill in the blank Description Generation and Question Answering.	MS COCO	VGGNet	-
3.	Image Understanding using vision and reasoning through Scene Description Graph.	ILSVRC + Flickr 8K image + MS-COCO	Reasonable	-
4.	Simple Baseline for Visual Question Answering.	COCO VQA dataset	VGGNet, GoogLeNet	61%

2.10.2 Long Term Short Memory Network

Zao et al. [2] uses a LSTM model was trained to generate the answer after receiving the image features (obtained from a convolutional net-work) and the question as input by Malinowski et al. Wu et al [6] extend a LSTM generation model to use external knowledge that is extracted from DBpedia. Xiong et al [14] a similar model but decouple the LSTMs used for encoding and decoding.

In Goyal's et al [20] paper words are recognized in various portions of the image and combined together with a language model to generate captions. Similarly, Xu et al [50] uses a recurrent network model to disclose salient objects and generate caption words one by one. The model works as an inverse of these caption models at test time by determining the relevant image region given a textual query as input. This enables the model to determine whether a question-answer pair is a good match given evidence from the image.

Serial No.	Paper	Data Set	Image Network	Accuracy
1.	LSTM An Empirical Evaluation of Visual Question Answering for Novel Objects	VQA	ImageNet	-
2.	VQA: Visual Question Answering	VQA	VGGNet	54.6
3.	Yin and Yang: Balancing and Answering Binary Visual Questions	MS COCO +VQA	ImageNet	74
4.	Knowledge Acquisition for Visual Question Answering via Iterative Querying	Visual7W and VQA		Visual7W:0.52 and VQA :0.47
5.	Biometrics and forensics integration using deep multi-modal semantic alignment and joint embedding	Subset of Visual Genome	VGGNet - 16	-

2.10.3 Combining RNNs and CNNs

The task of illustrating visual content like still images as well as videos has been acknowledged with a combination of encoding the image with CNNs and decoding, i.e. predicting the sentence description with an RNN (Donahueet al. 2015; Karpathy and Fei-Fei 2015; Venugopalanet al. 2015b; Vinyals et al. 2014; Zitnick et al. 2013). This is accomplished by using the RNN model that first observes the visual content and is then trained to predict a sequence of words that is a description of the visual content. [2] The proposed model extends this idea to question answering, where a model trained to either generate or classify an answer based on visual as well as natural language input. Kim et al, proposed a model where image and question was separately described by a CNN and by a RNN, and then a Multimodal Residual Network (MRN) was used for combining both modalities. Fukui et al used a CNN for describing the image and a two-layered LSTM for the question.

Serial No.	Paper	Data Set	Image Network	Accuracy
1.	A Deep Learning Approach to Visual Question Answering	Daquar	GoogLeNet	-
2.	Revisiting Visual Question Answering Baselines	Visual7W Telling+VQA Real Multiple Choice	ResNet-101	-
3.	VIBIKNet: Visual Bidirectional Kernelized Network for Visual Question Answering	VQA	VIBIK Net	-
6.	Multi-level Attention Networks for Visual Question Answering∗	VQA + Visual7W	ResNet(152 layers)	70.0
7.	Learning to Reason: End-to-End Module Networks for Visual Question Answering	SHAPES + CLEVR + VQ	VGG -16	83.7
8.	Multi-modal Factorized Bilinear Pooling with Co-Attention Learning for Visual Question Answering	VQA	152 layer ResNet	71.3
9.	Leveraging Visual Question Answering for Image-Caption Ranking	MSCOCO	19 layer VGGNet	Caption retrieval:7.1% Image retrieval:4.4%
10.	Ask Me Anything: Free-form Visual Question Answering Based on Knowledge from External Sources	Toronto COCO-QA + VQA	VGGNet-16	COCO-QA:69.73% VQA:59.44%
11.	Simple And Effective Visual Question Answering In A Single Modality	COCO-QA	GoogLeNet	61

2.11 Multimodal Compact Bilinear Pooling (MCB)

Bolaños et al [7] reckons on Multimodal Compact Bilinear pooling (MCB) to get a joint representation. Bilinear pooling enumerates the outer product between two vectors, which allows, in contrast to element-wise product, a multiplicative interaction between all elements of both vectors. For fine- grained classification of vison only tasks bilinear pooling models (Tenenbaum and Freeman, 2000) have recently shown to be beneficial. However, given their high dimensionality (n2), bilinear pooling has so far not been widely used. This paper upholds the idea from Zhung et al. [40] which depicts how to efficiently compress bilinear pooling for a single modality. This work extensively evaluate the extension to the multimodal case for text and visual modalities.

2.12 Attention Models

Lu et al [6] came across a hierarchical attention mechanism based on parses of the question and the image which they call "question-image co-attention". It models interactions between specific parts of the inputs (image and question) depending on their actual contents. Instead of holistic, image-wide features, the visual input is then typically represented a spatial feature map. The feature map along with the question is used to determine spatial weights that reverts the most applicable regions of the image. This approach uses graph representation, which they equate to subgraph matching, which is similar to weighting operation. Graph nodes representing question words are associated with graph nodes representing scene objects and vice versa. Similarly, the co-attention model of Lu et al determines attention weights on both image regions and question words. Their best-performing approach proceeds in a sequential manner, starting with question-guided visual attention followed by image-guided question attention. In our case, we found that a joint, one-pass version performs better.

Serial No.	Paper	Data Set	Image Network	Accuracy
1.	Adaptive Attention Fusion Network For Visual Question Answering	COCO-QA	19-layer VGGnet	64.7
3.	Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering	VQA	VGGNet	51.88
4.	ABC-CNN: An Attention Based	Toronto COCO-QA, DAQUAR, and VQA	ConceptNet	Coco-Qa: 47 Daquar: 42 Vqa:48

2.13 Survey Conclusion

VQA is an important basic research problem in computer vision and natural language processing that requires a system to do much more than task specific algorithms, such as object recognition and object detection. An algorithm that can answer arbitrary questions about images would be a milestone in artificial intelligence. We believe that VQA should be a necessary part of any visual Turing test.

For this report, we critically reviewed existing datasets and algorithms for VQA. We discussed the challenges of evaluating answers generated by algorithms, especially multi-word answers. We described how biases and other problems plague existing datasets. This is a major problem, and the field needs a dataset that evaluates the important characteristics of a VQA algorithm, so that if an algorithm performs well on that dataset then it means it is doing well on VQA in general.

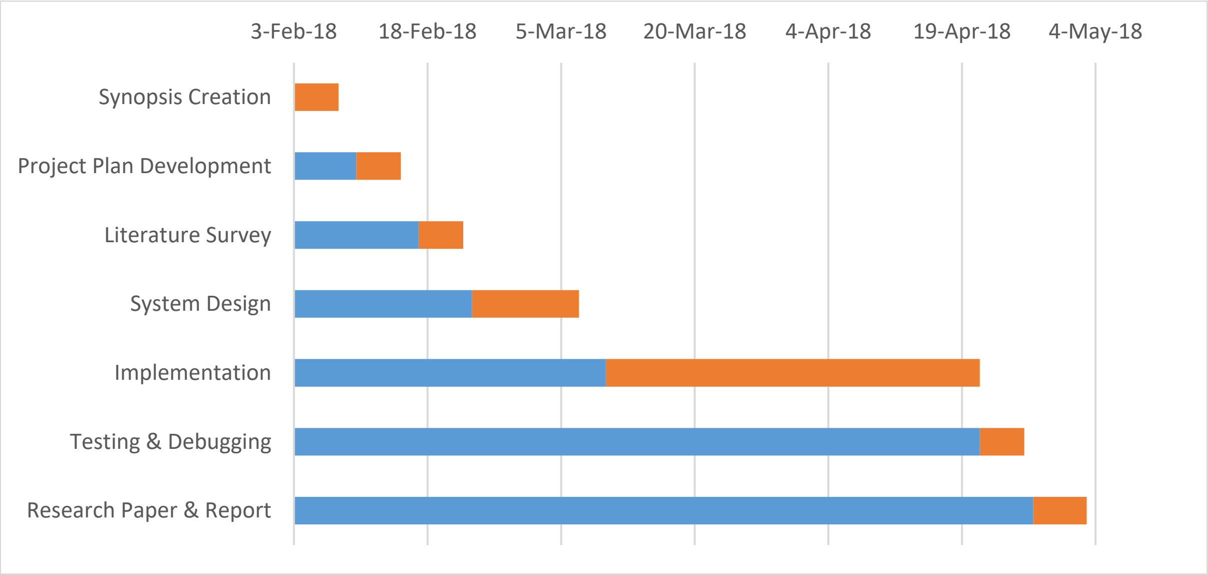
Future work on VQA involves the creation of larger and far more varied datasets. Bias in these datasets will be difficult to overcome, but evaluating different kinds of questions individually in a nuanced manner, rather than using naive accuracy alone, will help significantly. Further work will be needed to develop VQA algorithms that can reason about image content, but these algorithms may lead to significant new areas of research.

4. PROJECT MANAGEMENT PLAN

4.1 Schedule and Budget Summary:

- Development of the system just requires fairly powerful computers that all the group members already posses.
- The Gantt chart is to be followed strictly to meet all the deadlines and the project is to be delivered before the end of April 2018.

4.2 Gantt Chart:



4.3 Managerial Process Plans:

This section of the Project Management Plan specifies the project management processes for the project. This section defines the plans for project start-up, risk management, project work, project tracking and project close-out.

4.3.1 Start-up Plan

Estimates

- The resource and budget are already estimated and achieved for the starting of the project development.
- Based on the previous developed projects, the estimated size of the project will not be more than 5000 LOC.
- All the required estimations will be repeated at each 3 weeks interval to keep a track of budget, schedule and size.

Resource Acquisition

- The data required for the training of the model is open source and already available online.
- All the other essential software and hardware resources are already acquired.

4.3.2 Work Plan

Schedule Allocation

- The components of the project have a linear dependency among themselves i.e. the starting of any process depicted in the Gantt chart requires the previous component to be finished.
- The critical path is identified as follows: Project Plan -> Literature Survey -> Design -> Development -> Testing.
- The development of the research paper and report can start as soon as the design is complete and the implementation has started. Although it can be finished only after the project is complete.
- The deadlines for individual objectives are identified and put in the form of a Gantt chart.

Project Metrics

- Schedule Variance - ((Actual calendar days – Planned calendar days) + Start variance)/ Planned calendar days x 100.
- Productivity - Actual Project Size / Actual effort expended in the project.
- Cost of quality - (review + testing + verification review + verification testing + QA + configuration management + measurement + training + rework review + rework testing)/ total effort x 100.

5. SOFTWARE REQUIREMENT SPECIFICATIONS

5.1 Project Overview:

This project is an attempt to apply deep learning tools and neural networks to enable a computer to answer questions by looking at images pertaining to the question. The answers can be in various forms such as a phrase, a word, a yes/no answer, selection from multiple choice or fill in the blank. Adding a voice interface facilitates the usability and the accessibility of the project and helps it connect better with the user group.

This project is an open-source project that can be used for application, development or research by any individual or organisation with the condition that necessary acknowledgments may be provided.

5.2 General Description:

5.2.1 Product Functions

This project will take an image as input and try to answer questions pertaining to the image. The answers can be in various forms such as a phrase, a word, a yes/no answer, selection from multiple choice or fill in the blank.

5.2.2 Similar System Information

The system that is being developed is a relatively a new project but still there are a large number of similar systems present that use a wide array of different methods to achieve similar goals. The possible strength our system has over the majority is the improved features of the models used in the system.

5.2.3 User Characteristics

The users for the proposed system is anyone who would be interested in using our VQA System for various applications, research or further development. The project has several applications which can lead to normal people being the users of this system. Thus the users do not require high technical skills.

5.2.4 User Problem Statement

The current VQQ systems are not very efficient. There has not been any major project developed that has utilized this idea to its full capabilities. Thus the users want a system that is fast, efficient and dependable.

5.2.5 User Objectives

The user wants a system that give accurate answers when asked a question about any image. Also the system must keep training to increase its efficiency with each problem that the system processes.

5.2.6 General Constraints

Constraints include an easy to use user interface for the program, that runs on a Linux platform with Python and other necessary libraries installed. Also, if possible, compatible with Windows platform. The project must be developed in Python language with the development of necessary packages and libraries if necessary.

5.3 Functional Requirements:

5.3.1 Input provided can be of multiple formats.

- Images that'll be input can be of various formats but should be sharp and clear.
- Questions can be framed any way the user wants to.
- High criticality.
- Limited network availability could present a technical challenge.
- This requirement is the basis of the question part of the project.

5.3.2 The system shall be accessible via simple queries and low technical knowledge.

- Users should be able to easily ask questions on the image that that they input into the system. They should also be able to specify answer constraints.
- Very high criticality
- We do not foresee any technical issues preventing the implementation of this.
- Given the capabilities of the team and resources present, this requirement is able to be satisfied.
- This requirement depends on requirement number one.

5.3.3 The system should be able to be learn from each question.

- Questions and the images input and answer predicted should result in updated knowledge of the model.
- Very high criticality.
- We do not foresee any technical risks involved in this requirement.
- The only factor we can encounter here is the user of the system not being able to use it correctly. We will overcome this by assisting those who'll be using it.
- This requirement is dependent on requirement two.

5.3.4 Voice Interface

- If possible, a voice interface is to be added for input of the question and output of the answer.
- Low criticality.
- We do not foresee any technical risks involved in this requirement.
- This requirement shall utilize already developed APIs thus it should be able to be satisfied.
- This requirement does not depend on any other requirement nor does any other requirement depend of this requirement.

5.4 Interface Requirements:

5.4.1 User Interfaces

- **5.4.1.1 GUI**
  The user interface for this program is to be constructed for the easy access of the functionalities for the user. It can be made platform independent but at bare minimum it should have a basic GUI that runs on any Linux Platform.
- **5.4.1.2 CLI**
  There will be a command line interface only if a GUI is not possible or is very complex.
- **5.4.1.3 API**
  The project will be developed on any good python API such as Anaconda Navigator or can be developed as a Spyder or Jupyter Project.
- **5.4.1.4 Diagnostics or ROM**
  There will be documentation provided if any new libraries are developed.

5.4.2 Hardware Interfaces

The program uses the hard disk and requires a GPU for training. Access to the hard drive and GPU will be managed by the operating system and the developer.

5.4.3 Communications Interfaces

If we decide to implement a Voice Interface with communication networks available for a shared or concurrent use, the developer will program the system to handle those connections automatically.

5.4.4 Software Interfaces

The system will import data from online databases. This functionality is built-in to the API.

5.5 Performance Requirements:

The system is designed to be operated on a normal mid end PC, thus no additional system requirements exist except for the hardware requirements to train the model using a powerful GPU to decrease training time by implying parallel processing.

Basic requirements for our project are as follows:

- GHz processor or higher (Frequency can decrease if number of cores increases).
- 16 GB RAM or higher (32GB Recommended).
- NVIDIA 1060Ti GPU (6GB GDDR5) or higher.
- At least 15GB of Hard Drive Space.
- Windows 7 or Later or any Linux based operating system.

5.6 Non Functional Requirements:

5.6.1 Security

Since there is no obvious information that is of a high security level, the only requirements that could be implemented are encrypting the system and/or making the system access password-protected, at developer's interest.

5.6.2 Binary Compatibility

This system will be compatible with any computer that has Python installed along with the necessary libraries and frameworks present, and will also be designed keeping more than one operating systems in mind.

5.6.3 Reliability

Reliability is one of the key attributes of the system. Back-ups will be made regularly so that restoration with minimal efficiency loss is possible in the event of unforeseen mishap. The system will also be thoroughly tested by all team members to ensure reliability.

5.6.4 Maintainability

The system shall be maintained by the team in case there is any necessity of it.

5.6.5 Portability

The system shall be designed in a way that shall allow it to be run on multiple computers or mobiles.

5.6.6 Extensibility

The system shall be designed and documented in such a way that anybody with an understanding of Machine Learning, Natural Language Processing, Python and Image Processing shall be able to extend the system to fit their needs with the team's basic instructions.

5.6.7 Reusability

The system should be designed in a way that allows the model to be re-used repeatedly with increased efficiency.

5.6.8 Application Affinity/Compatibility

This system requires Python 3.0 or later to be installed on the system. It also requires the libraries and packages used which will be mentioned in the documentation.

5.6.9 Resource Utilization

The resources that will be used in the creation of this system include: A computer with a powerful GPU, and fast internet connection.

5.6.10 Serviceability

The debugging of the system should be able to be sufficiently performed by any person with a basic understanding of Python and Machine Learning.

5.7 Operational Scenarios:

Scenario A: Initial Training

The developer shall use the information available from the online databases for its initial training and evolution. After the model has been trained to a satisfactory level it can be published for normal use and further development.

Scenario B: User Access

The user shall be able to input and image enter a question pertaining to the image, and get answer in the form they want the output to be in. The system will attempt to answer the question accurately and as quickly as possible.

Scenario C: Debugging and Maintenance

In case a problem or a bug arises, any member of the team will debug and try to remove the bug as soon as possible.

6. DESIGN

6.1 Deep neural networks for VQA

The common approach to VQA is to train a deep neural network with supervision which maps the given image and question to a relative scoring of candidate answers. The main idea is to learn a joint embedding of the visual and textual inputs. First, the image and the question are processed independently to obtain separate vector representations. Those features are then are mapped with learned functions to a joint space, then combined and fed to an output stage. We examine each of those elements next.

6.2 Image Encoding

On the computer vision side, the input image x^I is processed with a deep convolutional neural network (CNN) to extract image features described as a vector y^I . This large fixed size vector encodes the contents of the image. This CNN is typically a standard network architecture that has been pre-trained to perform image recognition. The motivation for a pre-trained network is to take advantage of the vast amounts of training data available for image recognition, relative to the amounts of data annotated for VQA. The pre-trained network is used as a generic feature extractor, by discarding the final classification layers, and using the features produced within the CNN prior to this classification. In comparison to classical handcrafted image features such as scale-invariant feature transform (commonly known as SIFT) or histogram of oriented gradients (commonly known as HOG), CNN features provide higher-level representations of the contents of the image, and are naturally produced as a fixed-size vector. The size of this vector is typically in the order of 1,024 or 2,048

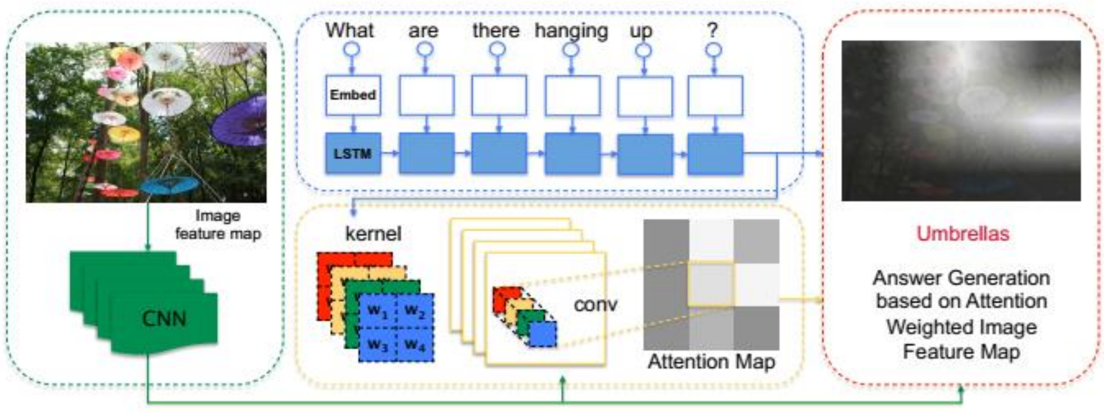
6.3 Question Encoding

On the language side, the input question is also processed to obtain a fixed-size representation of its contents. Initially, the i_{th} word of the question is represented by an index $x^Q{}_i$ in the input vocabulary. Each word is then turned into a vector. This uses a mapping implemented as a lookup table W[.] that associates the index of any word of the input vocabulary to a learned vector. An alternative implementation initially represents each word with a one-hot vector (a vector of all zeros, except for a one at the location of the word index in the vocabulary), which is then multiplied with a dense weight matrix that contains the embeddings of all words. The vectors of all words are then collapsed into a single vector. A simple option for this purpose is to make a bag-of-words (BoW), which corresponds to simply averaging the word vectors, i.e., $y_x{}^Q = (1/N)\sum W[x_i{}^Q]$. Another popular option is to feed the word vectors into a recurrent neural network (RNN) such as a long short-term memory (LSTM). An RNN processes words sequentially and can capture the sequential relationships between them. In comparison, a BoW does not account for word order, and, for example, would produce a same representation for “this man eats a hot dog” and “a hot man eats this dog.”

6.4 Combining the Image and Question Features

The feature vectors y^I and y^Q represent the image and the questions, respectively. They are each passed through a learned function before being combined. The intuition here is to map the features to a joint space, in which distances between both modalities become comparable. The learned functions $f^I(.)$ and $f^Q(.)$ are typically implemented as additional layers of the neural network, e.g., f(y) = ReLU(Wy + b) , where W and b are learned weights and biases, and ReLU is a rectified linear unit that serves as a nonlinearity. The mapped features are then combined before being fed to the output stage. A simple option for this combination is to simply concatenate them as $z = [f^I(y^I). f^Q(y^Q)]$. Alternatively, it is popular to include multiplicative interactions within the neural network to increase its capacity and use $z = f^I(y^I). f^Q(y^Q$, where · is the Hadamard (element-wise) product.

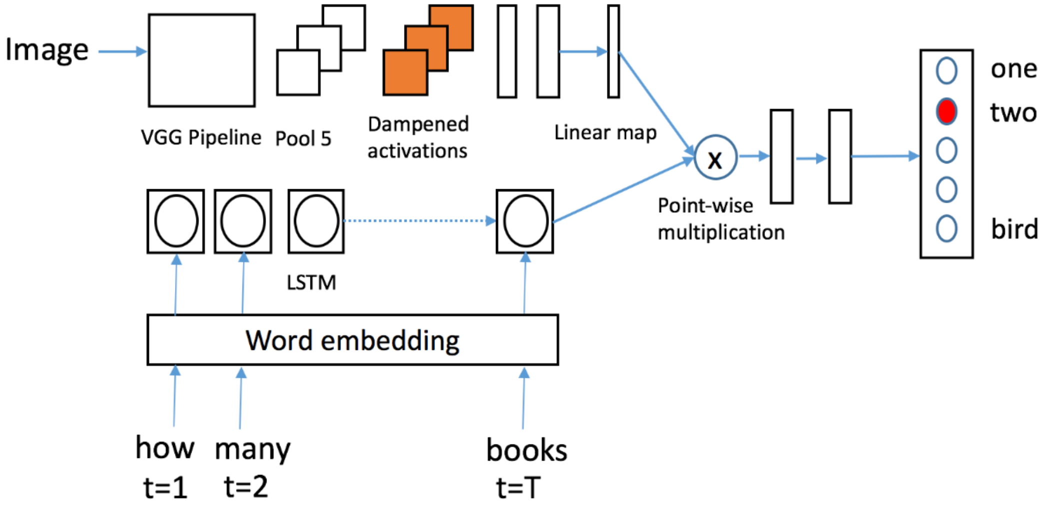
6.5 Attention Based Model



The framework: The green box denotes the image feature extraction part using CNN; the blue box is the question understanding part using LSTM; the yellow box illustrates the attention extraction part with configurable convolution; the red box is the answer generation part using multi-class classification based on attention weighted image feature maps. The orange letters are corresponding variables.

6.6 Proposed Model

Our model uses a LSTM unit to convert the question into a 1024 dimension encoding. The LSTM model takes one-hot encoding for the question words as input followed by a linear transformation to transform the image features to 1024 dimensions to match the LSTM encoding of the question. The question and image encodings are then fused. The fused features are then passed through a multi-layer perceptron neural network classifier with 2 hidden layers and 1000 hidden units. The output will be a 1000 way softmax classifier that predicts one of top-1000 answers in the training dataset.

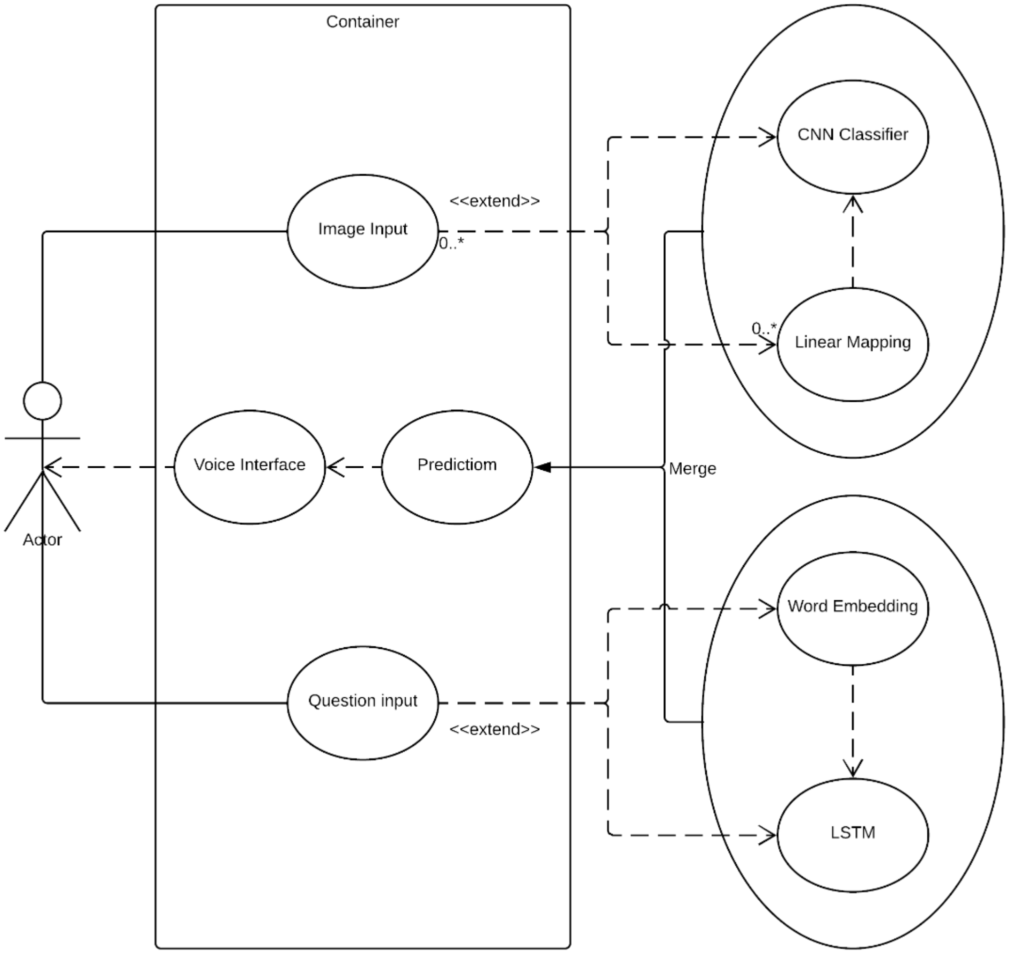


6.7 Generating Output

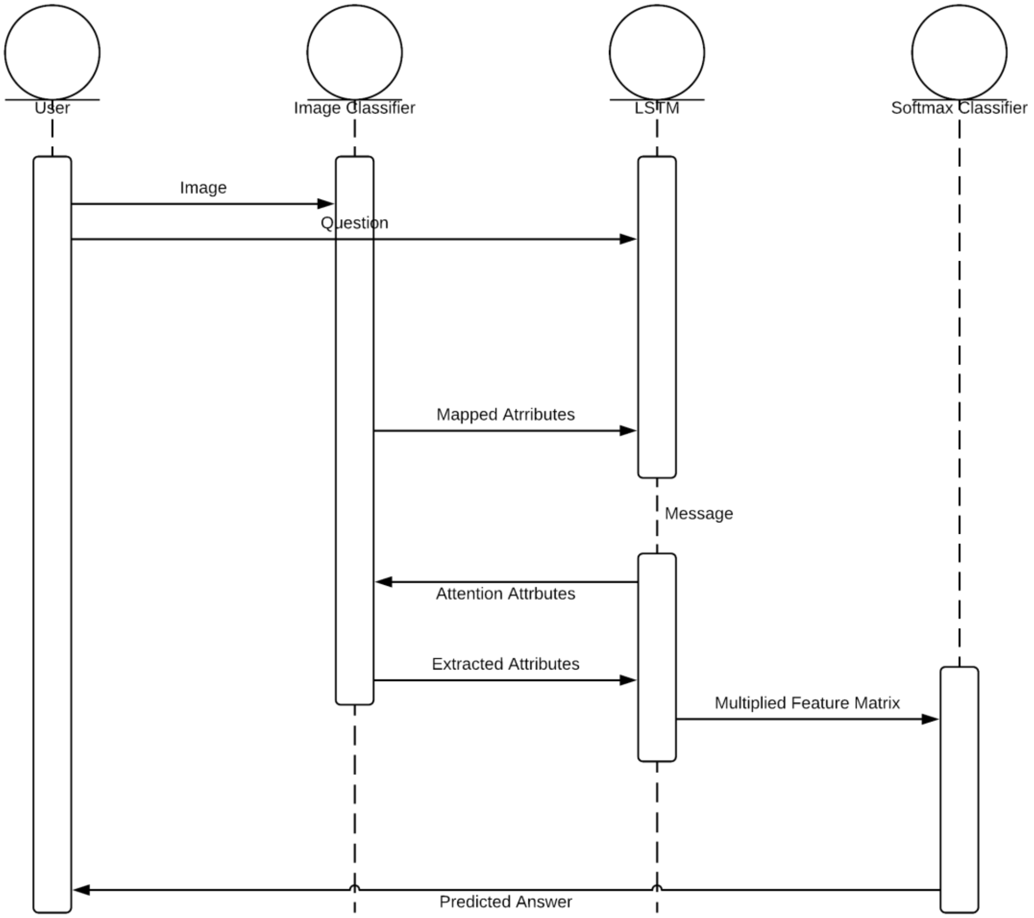
The output stage of a VQA system can be seen either as a generation or as a classification task. The generation of a free-form answer has the advantage of being able to compose complex sentences. In practice however, such a model is difficult to learn. Current data sets are limited to short answers, and a practical alternative is to rather learn a classifier over candidate answers. For this purpose, a large set of candidate answers is predetermined from the most common ones in the training set (typically in the order of 2,000). This inevitably leaves out some infrequent words, but such a set is typically sufficient to answer correctly more than 90% of test questions. This is a non-limiting issue since this figure is well above the accuracy of current systems. The combined features z are passed to a classifier over those candidate answers (a linear layer followed by a softmax or sigmoid transformation). The classifier assigns score to each candidate answer, and the top-ranked one is returned as the final output. In a multiple choice setting, only the scores assigned to proposed choices are considered. For training the model, the classifier is followed by a cross-entropy loss, and the whole network is trained end-to-end by backpropagation to minimize this loss over the set of training examples.

6.8 Preliminary Use Case Models:

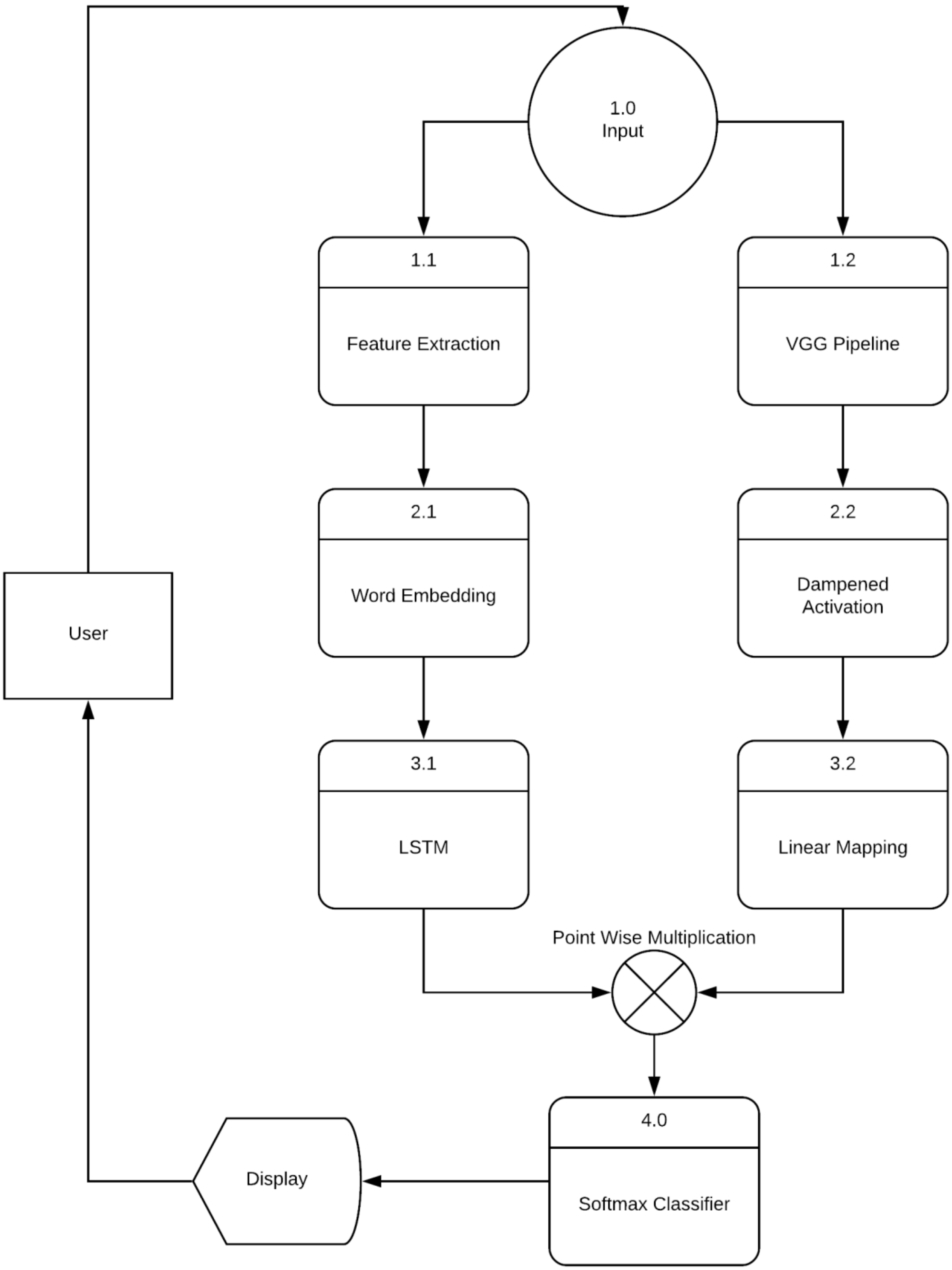
This section presents a fundamental use case diagram with necessary use cases that satisfy the system’s requirements. The purpose is to provide an alternative, "structural" view of the requirements stated above and how they might be satisfied in the system.



6.9 Sequence Diagram



6.10 Dataflow Diagram



7. IMPLEMENTATION

7.1 Baseline

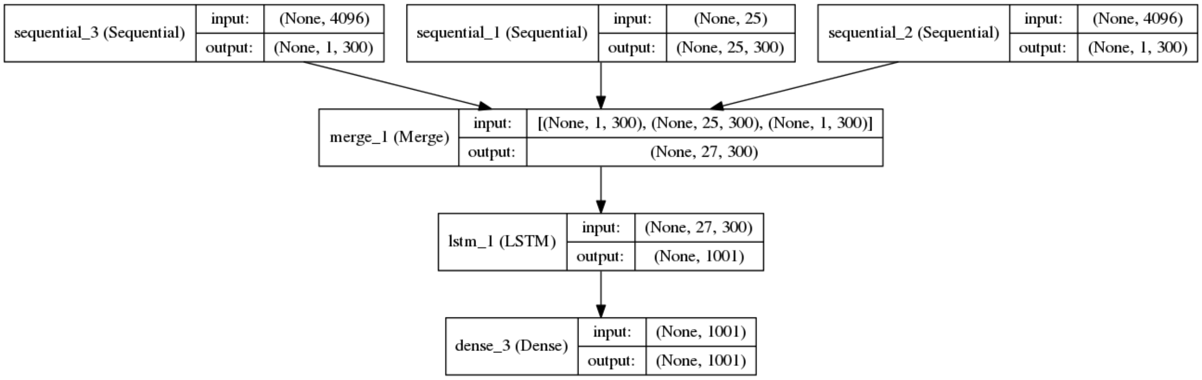
Given the 1-hot encoding of the words of the question $Q = \{q_1. \; . \; . \; q_b\}$, firstly we embed the words to a vector space to get $Q_w = \{q_{w1}, \ldots, q_{wa}\}$. To compute the phrase features, we apply 1-D convolution on the word embedding vectors. Precisely, at each word location, we calculate the internal product of the word vectors with sieves of multiple window sizes: unigram, bigram and trigram. For the given range of the words, the convolution product with window length s is given by

$$q\hat{}\ p_{a,b} = \tanh(q^{w}{}_{a:a-1+b}.W_a), b \in \{1, 2, 3\}$$

Where, W_s is the weight parameter. The word-level features Q are suitably 0-padded before being fed into bigram and trigram window convolutions to preserve the size of the series after the convolution is finished. Provided the convolution results, we then apply max-pooling across various n-grams discretely at every word location to procure phrase-level features

$$q^{p}{}_{b} = \max(q\hat{}\ p_1,b, q\hat{}\ p_2,b, \ldots q\hat{}\ p_3,b), b \in \{1, 2, \ldots, B\}$$

The pooling method we used varies from those used in prior works [1] in that it adaptively chooses discrete gram features at every single time step, while retaining the aboriginal sequence size and form. We then adopt an LSTM to encode the sequence q_p^b after max-pooling. The analogous question-level feature q_s^b is the hidden vector of the LSTM at time t.



7.2 Co-Attention

We nominate two co-attention systems that alter in the order in which question and image attention maps are developed. The first system, which we are calling simultaneous co-attention, produces question and image attention simultaneously. The second mechanism, which can we are calling alternating co-attention, sequentially oscillates between generating question and image attentions.

Alternating Co-Attention. In this mechanism, we sequentially alternate amid generating question and image attention. In brief, it subsists of 3 steps: 1) encapsulate the question into one individual vector say v; 2) attend to the image based on the summary of the question; and 3) attend to the question again based on the attended feature of the image. Thus, we characterize an attention operation as $x\hat{} = A(X; g)$, which takes the question (or image) features X and attention guidance g generated from the image (or question) as inputs, and outputs the attended question (or image) vector. This method can be summarized in the following steps

$$H = \tanh(U.T.W_g g\ + XxWx)^{-1})$$

$$a = \mathrm{softmax}(h_x H.w_B)$$

$$x = \sum a_x^i\, x_i$$

Where, U is a vector with all elements equal to 1. $W_x, W_g \in R_{k\times d}$ and $w_h^x \in R_k$ are the parameters. 'a' is the attention weight of the feature X. The fluctuating co-attention process is explained as the following. At the foremost step of alternating co-attention, X = Q, and g is equal to 0; At the second step, X = V where V is the question feature, and the guidance g is intermediate attended image feature s from the first step; Finally, we use the attended question feature v as the guidance to attend the image again, i.e., X = Q and g = v. Analogous to the parallel co-attention model, this alternating co-attention is also performed at every level of the hierarchy.

Simultaneous Co-Attention - It attends to the image and question concurrently. Similar to [24], we associate the image and question by determining the affinity between question and image features at all pairs of question-locations and image-locations. Explicitly, provided an the question representation $Q \in X_d\times B$, and an image feature map $V \in X_d\times N$, the affinity matrix $C \in XB \times N$ is calculated by

$$C = \tanh(V.Q_B.W_x\ )$$

where $W_b \in Rd\times d$ consists of the weights. After calculating this affinity matrix, one plausible way of calculating the image (or question) attention is by simply maximizing out the affinity over the locations of other modalities, i.e. $k_v[n] = maxi(C_i,n)$ and $n_a[t] = maxj\ (C_b,j\ )$. Rather than selecting the maximum activation, we search if the performance is enhanced if we contemplate this affinity matrix as a feature and learn to predict question and image attention maps by the following method

$$H_v=\tanh((QxW_q)^{-1}C+ W_vV),\ H_q= \tanh((W_vxV\ )^{-1}CxT+ W_qQ)$$

$$a_1 = \mathrm{softmax}(wT\ h_vH_v\ ),\ a_2 = \mathrm{softmax}(wT\ hqHq\ )$$

Where $Wv,Wq \in Rk\times d$, $wv, wq \in Rk$ are the weight parameters. $a_1 \in R_N$ and $a_2 \in R_T$ are the attention probabilities of every discrete image region v_n and word q_b respectively. The affinity matrix C transforms question based attention space to image based attention space. Based on the above attention weights, the question and image attention vectors are computed as the weighted sum of the question features and image features,

$$v = \sum N_{n=i}\, a_{v}.n_{v}^{n-1},\ q = \sum T_{t=i}\, a_{q}^{t}.q_{t-1}$$

This simultaneous co-attention is performed at every single level in the hierarchy, leading to v and q where r ∈ {w, p, s}.

7.3 Setup

We used Keras with TensorFlow to develop our model. We use the Rmsprop optimizer with a base learning rate of 5e-4, momentum 0.98 and weight-decay 1e-7. The batch size is set to be 50 and training is done for up to 200 epochs with early stopping if the validation accuracy has not improved in the last 4 epochs. The size of hidden layer is set to 512. All the other word embeddings and hidden layers are vectors of size 384. We apply dropout with probability **0**.*48* on each layer. Also, rescaling of the image is done to **248** × **248**, and then activation taken from the last pooling layer of VGGNet as its feature.

8. RESULTS

The following diagrams shows results of our model on the COCO-QA test set segregated by the type of parameters. Our model can be said to perform good compared to the state-of-the-art 61.6% (SAN (2, CNN)) model with 62.1% (Ours + VGGNet16) accuracy. We also believe that alternating co-attention will perform better than the simultaneous co-attention system for this model but it has chances in incurring more errors at each level of hierarchy. Both these attention mechanisms have their advantages and disadvantages: simultaneous co-attention is more difficult to train as the dot product between image and text is to be calculated which compresses two vectors into a single value. On the other hand, alternating co-attention might suffer from errors that are being accumulated at every iteration.

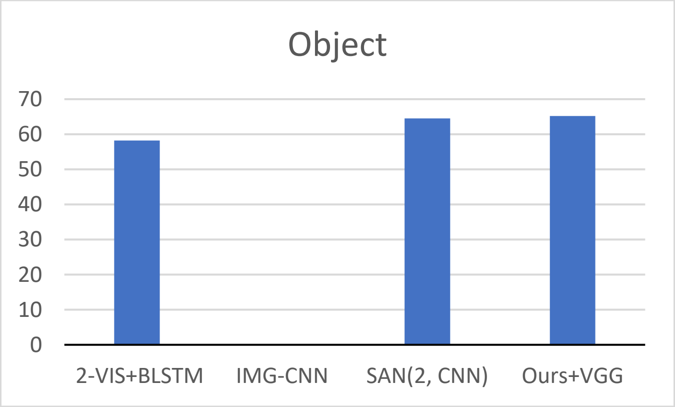


Fig 1 – Performance based on answers being objects

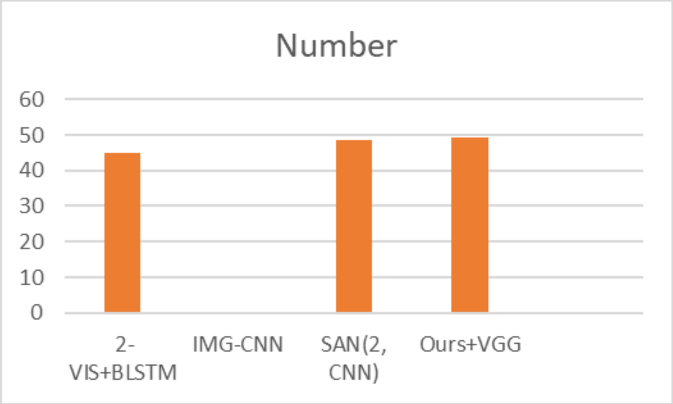


Fig 2 – Performance based on answers being numbers

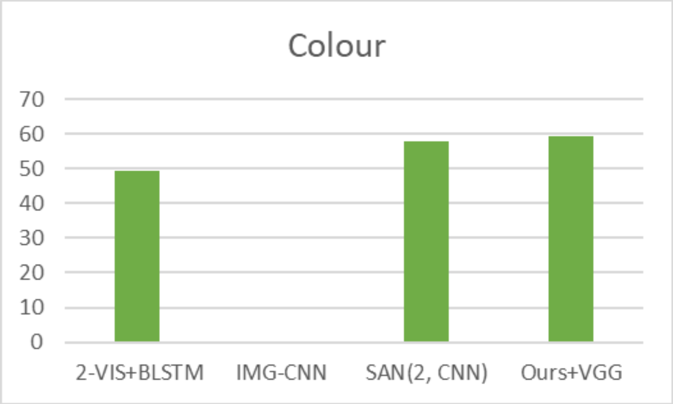


Fig 3 – Performance based on answers being colours

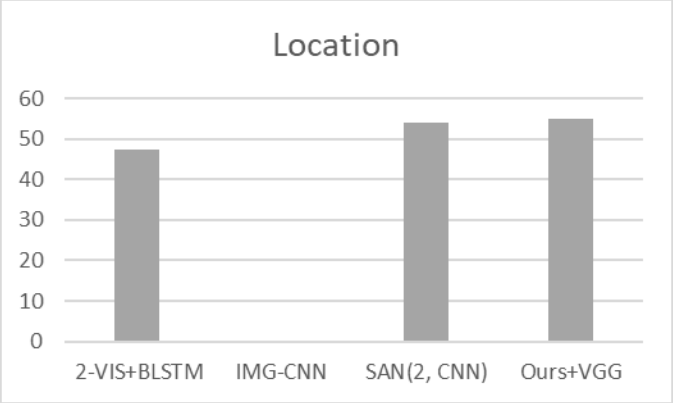


Fig 4 – Performance based on answers being locations

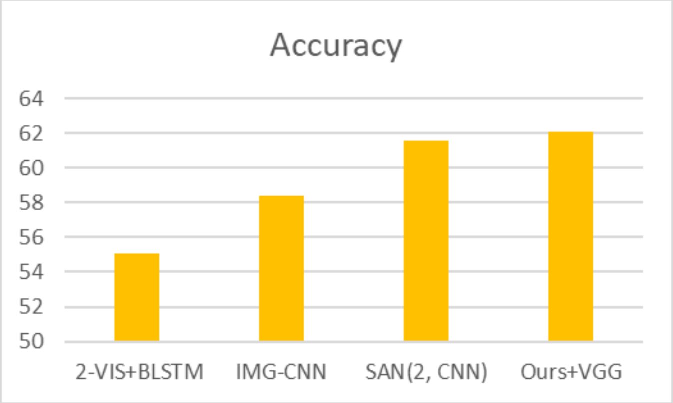


Fig 5 – Performance based on accuracy

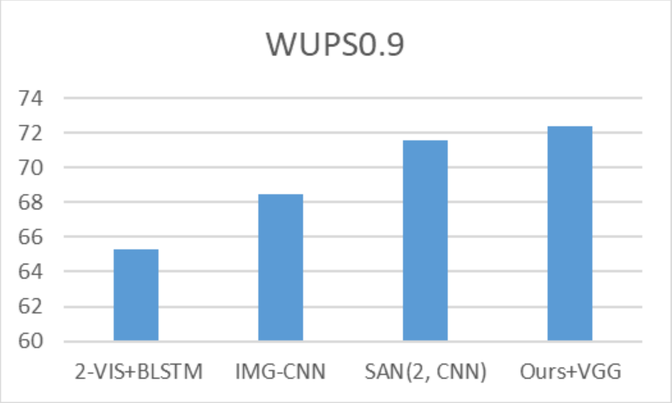
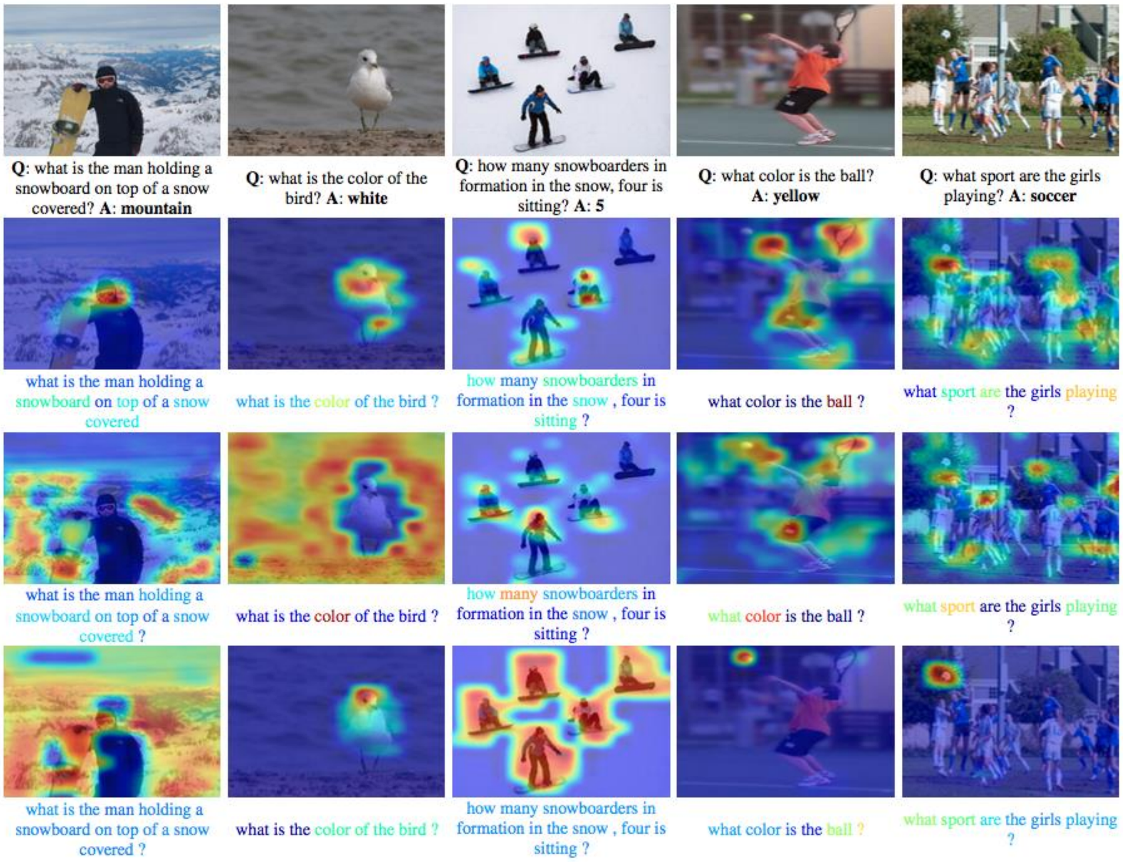


Fig 6 – Performance based on WUPS Scores

8.1 Co-Attention Results

We now try to visualize one of the co-attention maps produced by this method in following image. At the word level, the model attends essentially to the object regions in an image, e.g., heads, animal, face etc.


Q: what is the man holding a snowboard on top of a snow covered? A: mountain
Q: what is the color of the bird? A: white
Q: how many snowboarders in formation in the snow, four is sitting? A: 5
Q: what color is the ball? A: yellow
Q: what sport are the girls playing? A: soccer
what is the man holding a snowboard on top of a snow covered
what is the color of the bird ?
how many snowboarders in formation in the snow , four is sitting ?
what color is the ball ?
what sport are the girls playing ?
what is the man holding a snowboard on top of a snow covered ?
what is the color of the bird ?
how many snowboarders in formation in the snow , four is sitting ?
what color is the ball ?
what sport are the girls playing ?
what is the man holding a snowboard on top of a snow covered ?
what is the color of the bird ?
how many snowboarders in formation in the snow , four is sitting ?
what color is the ball ?
what sport are the girls playing ?

8.3 Conclusion

At the construction level, the image attention has varying patterns over varying images. For the initial 2 images, the attention is seen shifting to the background areas from the objects. For the third image, the attention inclines to become even more focused on the object. We believe that this is induced by the different types of the question. On the other side i.e. the question side, this model is adept at localizing the important phrases of the question, therefore fundamentally discovering the types of the question in the dataset. Let us say for example, this model wants to pay more attention to the phrases like “what color” and “how many snowboarders”. This model efficiently attends to the regions in phrases in the questions and images appropriate for answering the question, e.g., “color of the animal” and animal region. Because this model performs co-attention at 3 different levels, it often obtains complementary information from each of these levels, and then fuses all of them to predict the answer.

We proffered a co-attention model for visual question answering. This inclusion of co-attention enables our model to attend to various regions of the chunks of the question as well as various regions of the image. We modelled the question hierarchically at 3 varying levels to obtain information from various granularities. Ablation studies can further demonstrate the role of question hierarchy and co-attention in visual question answering performance. Through various visualizations, we observed that this model co-attends to interpretable regions of questions and images for predicting then generating the answer. Though this model was evaluated only on visual question answering, it can be potentially applied to various other tasks that involve computer vision and language.

9. Références

[1] Rodríguez, Pau, et al. "Age and gender recognition in the wild with deep attention." *Pattern Recognition* 72 (2017): 563-571.
[2] Malinowski, Mateusz, Marcus Rohrbach, and Mario Fritz. "Ask your neurons: A deep learning approach to visual question answering." *International Journal of Computer Vision* 125.1-3 (2017): 110-135.
[3] Fukui, Akira, et al. "Multimodal compact bilinear pooling for visual question answering and visual grounding." *arXiv preprint arXiv:1606.01847* (2016).
[4] Agrawal, Aishwarya, Dhruv Batra, and Devi Parikh. "Analyzing the behavior of visual question answering models." *arXiv preprint arXiv:1606.07356* (2016).
[5] Jabri, Allan, Armand Joulin, and Laurens van der Maaten. "Revisiting visual question answering baselines." *European conference on computer vision*. Springer, Cham, 2016.
[6] Wu, Qi, et al. "Visual question answering: A survey of methods and datasets." *Computer Vision and Image Understanding* 163 (2017): 21-40.
[7] Bolaños, Marc, et al. "VIBIKNet: Visual bidirectional kernelized network for visual question answering." *Iberian Conference on Pattern Recognition and Image Analysis*. Springer, Cham, 2017.
[8] Yu, Licheng, et al. "Visual madlibs: Fill in the blank description generation and question answering." *Computer Vision (ICCV), 2015 IEEE International Conference on*. IEEE, 2015.
[9] Kafle, Kushal, and Christopher Kanan. "Answer-type prediction for visual question answering." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016
[10] Burt, Ryan, Mihael Cudic, and Jose C. Principe. "Fusing attention with visual question answering." *Neural Networks (IJCNN), 2017 International Joint Conference on*. IEEE, 2017.
[11] Saito, Kuniaki, et al. "Dualnet: Domain-invariant network for visual question answering." *Multimedia and Expo (ICME), 2017 IEEE International Conference on*. IEEE, 2017.
[12] Teney, Damien, Lingqiao Liu, and Anton van den Hengel. "Graph-structured representations for visual question answering." *CoRR, abs/1609.05600* 3 (2016).
[13] Ramakrishnan, Santhosh K., et al. "An Empirical Evaluation of Visual Question Answering for Novel Objects." *arXiv preprint arXiv:1704.02516* (2017).
[14] Xiong, Caiming, Stephen Merity, and Richard Socher. "Dynamic memory networks for visual and textual question answering." *International Conference on Machine Learning*. 2016.
[15] Zhu, Yuke, Joseph J. Lim, and Li Fei-Fei. "Knowledge acquisition for visual question answering via iterative querying." *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 2. No. 3. 2017.
[16] Besbes, Ghada, Hajer Baazaoui-Zghal, and Henda Ben Ghezela. "An ontology-driven visual question-answering framework." *Information Visualisation (iV), 2015 19th International Conference on*. IEEE, 2015.
[17] Gu, Geonmo, Seong Tae Kim, and Yong Man Ro. "Adaptive attention fusion network for visual question answering." *Multimedia and Expo (ICME), 2017 IEEE International Conference on*. IEEE, 2017.

[18] Jang, Yunseok, et al. "TGIF-QA: Toward spatio-temporal reasoning in visual question answering." *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017). Honolulu, Hawaii*. 2017.

[19] Yu, Dongfei, et al. "Multi-level attention networks for visual question answering." *Conf. on Computer Vision and Pattern Recognition*. Vol. 1. No. 7. 2017.

[20] Goyal, Yash, et al. "Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering." *CVPR*. Vol. 1. No. 6. 2017.

[21] Teney, Damien, Qi Wu, and Anton van den Hengel. "Visual Question Answering: A Tutorial." *IEEE Signal Processing Magazine* 34.6 (2017): 63-75.

[22] Hu, Ronghang, et al. "Learning to reason: End-to-end module networks for visual question answering." *CoRR, abs/1704.05526* 3 (2017).

[23] Yu, Zhou, et al. "Multi-modal factorized bilinear pooling with co-attention learning for visual question answering." *Proc. IEEE Int. Conf. Comp. Vis*. Vol. 3. 2017.

[24] Sarhan, Abdullah, Jon Rokne, and Reda Alhajj. "Integrating flexibility and fuzziness into a question driven query model." *Information Sciences* 430 (2018): 349-377.

[25] Fernandez, Alison, and Alexandre Bergel. "A Domain-Specific Language to Visualize Software Evolution.

[26] Kafle, Kushal, and Christopher Kanan. "Visual question answering: Datasets, algorithms, and future challenges." *Computer Vision and Image Understanding* 163 (2017): 3-20.

[27] Lin, Xiao, and Devi Parikh. "Leveraging visual question answering for image-caption ranking." *European Conference on Computer Vision*. Springer, Cham, 2016.

[28] Antol, Stanislaw, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. "Vqa: Visual question answering." In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2425-2433. 2015.

[29] Ares, Gastón, et al. "Visual attention by consumers to check-all-that-apply questions: Insights to support methodological development." *Food Quality and Preference* 32 (2014).

[30] Wu, Qi, et al. "Ask me anything: Free-form visual question answering based on knowledge from external sources." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

[31] Agrawal, Aishwarya, et al. "Vqa: Visual question answering." *International Journal of Computer Vision* 123.1 (2017): 4-31.

[32] Wang, Xueming, Zechao Li, and Jinhui Tang. "Multimedia news QA: Extraction and visualization integration with multiple-source information." *Image and Vision Computing* 60 (2017): 162-170.

[33] Cecílio, José, Karen Duarte, and Pedro Furtado. "BlindeDroid: An information tracking system for real-time guiding of blind people." *Procedia Computer Science* 52 (2015): 113-120.

[34] Aditya, Somak, et al. "Image Understanding using vision and reasoning through Scene Description Graph." *Computer Vision and Image Understanding* (2017).

[35] Chen, Kan, et al. "ABC-CNN: An attention based convolutional neural network for visual question answering." *arXiv preprint arXiv:1511.05960* (2015).

[36] Zhou, Bolei, et al. "Simple baseline for visual question answering." *arXiv preprint arXiv:1512.02167* (2015).

[37] Antol, Stanislaw, et al. "Vqa: Visual question answering." *Proceedings of the IEEE International Conference on Computer Vision*. 2015.

[38] Lin, Yuetan, et al. "Simple and effective visual question answering in a single modality." *Image Processing (ICIP), 2016 IEEE International Conference on*. IEEE, 2016.

[39] Wu, Q., Wang, P., Shen, C., Dick, A., & van den Hengel, A. (2016). Ask me anything: Free-form visual question answering based on knowledge from external sources. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[40] Zhang, Peng, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. "Yin and yang: Balancing and answering binary visual questions." In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, pp. 5014-5022. IEEE, 2016.

[41] Wu, Qi, et al. "Image captioning and visual question answering based on attributes and external knowledge." *IEEE transactions on pattern analysis and machine intelligence* (2017).

[42] Zhu, Chen, et al. "Structured attentions for visual question answering." *Proc. IEEE Int. Conf. Comp. Vis*. Vol. 3. 2017.

[43] Kafle, K., & Kanan, C. (2017, October). An analysis of visual question answering algorithms. In *2017 IEEE International Conference on Computer Vision (ICCV)* (pp. 1983-1991). IEEE.

[44] Zhu, Yuke, Joseph J. Lim, and Li Fei-Fei. "Knowledge acquisition for visual question answering via iterative querying." *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 2. No. 3. 2017.

[45] Edmiston, Pierce, and Gary Lupyan. "Visual interference disrupts visual knowledge." *Journal of Memory and Language* 92 (2017): 281-292.

[46] Toor, Andeep S., and Harry Wechsler. "Biometrics and forensics integration using deep multi-modal semantic alignment and joint embedding." *Pattern Recognition Letters* (2017).

[47] Liu, Mingnan, and Frederick G. Conrad. "An experiment testing six formats of 101-point rating scales." *Computers in Human Behavior* 55 (2016): 364-371.

[48] Wu, Qi, et al. "Visual question answering: A survey of methods and datasets." *Computer Vision and Image Understanding* 163 (2017): 21-40.

[49] Das, Abhishek, et al. "Human attention in visual question answering: Do humans and deep networks look at the same regions?." *Computer Vision and Image Understanding* 163 (2017): 90-100.

[50] Xu, Huijuan, and Kate Saenko. "Ask, attend and answer: Exploring question-guided spatial attention for visual question answering." *European Conference on Computer Vision*. Springer, Cham, 2016.

[51] Lu, Jiasen, Jianwei Yang, Dhruv Batra, and Devi Parikh. "Hierarchical question-image co-attention for visual question answering." In *Advances In Neural Information Processing Systems*, pp. 289-297. 2016.

[52] Sadeghi, Fereshteh, Santosh K. Kumar Divvala, and Ali Farhadi. "Viske: Visual knowledge extraction and question answering by visual verification of relation phrases." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

[53] Shih, Kevin J., Saurabh Singh, and Derek Hoiem. "Where to look: Focus regions for visual question answering." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

[54] Kim, B., and J. Kim. "Question-aware prediction with candidate answer recommendation for visual question answering." *Electronics Letters* 53.18 (2017): 1244-1246.

[55] Geman, D., Geman, S., Hallonquist, N., Younes, L.: Visual Turing test for computer vision systems. Proceedings of the National Academy of Sciences 112(12) (2015) 3618–3623.