CO-ATTENTION MODEL FOR VISUAL QUESTION ANSWERING

by

VAGUL MM **2014103081**
PRINCE MELVIN A **2014103596**
JEGATHEESWARAN M **2014103583**

A project report submitted to the

FACULTY OF INFORMATION AND COMMUNICATION ENGINEERING

in partial fulfillment of the requirements for the award of the degree of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ANNA UNIVERSITY, CHENNAI – 25

APRIL 2018

BONAFIDE CERTIFICATE

Certified that this project report titled **CO-ATTENTION MODEL FOR VISUAL QUESTION ANSWERING** is the *bonafide* work of **VAGUL MM (2014103081)**, **PRINCE MELVIN A (2014103596)** and **JEGATHEESWARAN M (2014103583)** who carried out the project work under my supervision, for the partial fulfillment of the requirements for the award of the degree of Bachelor of Engineering in Computer Science and Engineering. Certified further that to the best of my knowledge, the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or an award was conferred on an earlier occasion on these or any other candidates.

Place: Chennai
Date:

Dr. S.Sudha
Selection Grade Assistant Professor
Department of Computer Science and Engineering
Anna University, Chennai – 25

COUNTERSIGNED

Head of the Department,
Department of Computer Science and Engineering,
Anna University Chennai,
Chennai – 600025

ACKNOWLEDGEMENT

Foremost,we would like to express our sincere gratitude to our project guide **Dr.S.Sudha**, Selection Grade Assistant Professor, Department of Computer Science and Engineering,CEG for her continuous support and guidance which was instrumental in taking the project to successful completion.

We are grateful to **Dr.D.Manjula**, Head, Department of Computer Science and Engineering,CEG for providing conducive environment and facilities for the project.

We express our heartiest thanks to our reviewers, **Dr.V.Vetriselvi**, Associate Professor, Department of Computer Science and Engineering, CEG, **Dr.Geetha Palanisamy**, Associate Professor, Department of Computer Science and Engineering, Anna University and **Dr.R.Arockia Xavier Annie**, Assistant Professor, Department of Computer Science and Engineering, CEG for their valuable suggestions and critical reviews throughout the course of our project

Vagul MM **Prince Melvin A** **Jegatheeswaran M**

ABSTRACT

With tremendous recent progress of computer vision, computers are now capable of dealing with complicated tasks such as object recognition, scene classification, action recognition. Visual Question Answering generally refers to recognizing the text-based questions about an image and to infer the answer for each question. We have developed an co-attention model to map the relation between the image and text.

Added an image-question co-attention to exploit multi-modal relationship between pixels and words, which successfully boosts accuracy.Specifically, not only are we focusing on which region in the image we should look at, but which word in the sentence is of interest.We build Multi-Layer Perceptron (MLP) model with question-grouped training and softmax loss.

Our model has a natural symmetry between image and question which predicts the answer based on the image features extracted and in correspondence to the question asked for the given image.

திட்டப்பணி சுருக்கம்

காட்சியின் கேள்விக்கான பதில். பொதுவாக ஒரு படத்தை பற்றி உரை சார்ந்த கேள்விகளுக்கு அங்கீகாரம் மற்றும் ஒவ்வொரு கேள்விக்கு பதில் தாக்கத்தை குறிக்கிறது. படத்திற்கும் உரைக்கும் இடையேயான தொடர்பைக் கண்டறிவதற்கு நாங்கள் ஒரு கூட்டு கவனத்தை மாற்றியுள்ளோம்.

பிக்சல்கள் மற்றும் வார்த்தைகளுக்கு இடையே உள்ள பல மாதிரி உறவைச் சுரண்டுவதற்கு நாம் பட-கேள்வி இணை கவனத்தைச் சேர்க்கிறோம், இது வெற்றிகரமாக துல்லியத்தை அதிகரிக்கிறது. நாம் கேள்வி-குழு பயிற்சி மற்றும் மென்மையில் இழப்புடன் மல்டி லேயர் பெர்செப்சன் மாதிரியை உருவாக்குகிறோம்.

எங்கள் மாதிரியை படம் மற்றும் கேள்விக்கு இடையில் ஒரு இயல்பான சமச்சீர்நிலை உள்ளது, இதன் அர்த்தம் பட பிரதிநிதித்துவம் கேள்வி கவனத்தை வழிகாட்ட பயன்படுத்தப்படுகிறது மற்றும் கேள்வி பிரதிநிதித்துவம் பட கவனத்தை வழிகாட்ட பயன்படுத்தப்படுகின்றன.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

CBOW Continuous Bag-of-Words

CNN Convolution Neural Network

LSTM Long Short Term Memory

MCB Multi-modal Compact Bilinear pooling

MLP Multi- Layer Perceptron

TDIUC Task Driven Image Understanding Challenge

VGG Visual Geometric Groups

VQA Visual Question Answering

CHAPTER 1

INTRODUCTION

1.1 PROBLEM DOMAIN

Visual Question Answering(VQA) is an exciting computer vision problem that requires a system to be capable of many tasks. Truly solving VQA would be a milestone in artificial intelligence, and would significantly advance human computer interaction. However, VQA datasets must test a wide range of abilities for progress to be adequately measured. With tremendous recent progress of computer vision, computers are now capable of dealing with complicated tasks such as object recognition, scene classification, action recognition. However, the computers ability of understanding semantic details of images still needs to be further evaluated on much complicated tasks. At the same time, the emerging of many successful deep structures in the field of natural language processing has lead to impressive performance in traditional Question-Answering problems. VQA has many potential real-world applications, among which one of the most valuable is to assist visually impaired individuals in understanding contents of images from the web.

Moreover, VQA could be a great system to use in the Visual Turing Test or other computer vision tasks to evaluate the systems ability when comparing with human performance. We propose a novel mechanism that jointly reasons about visual attention and question attention, which we refer to as co-attention.

1.2 PROBLEM DESCRIPTION

The project aims to, answering question about the given image with image attention based on question. First,embedding matrix is extracted for words in questions/answers and promising image features to support any further training. Methods used to extract multi-modal features, image features and text features are from different feature spaces which means it is very difficult to incorporate them. To explore correlation between multi-modal features or build relationships between these features. Bayesian models and visual attention models are most commonly used. It is not enough to only consider image attention, but it is essential to consider image attention based on question embedding matrix is extracted for words in questions/answers and promising image features to support any further training on our proposed model. Also, whatever methods we use to extract multi-modal features, image features and text features are from different feature spaces which means it is very difficult to incorporate them in the same model.

One possible solution is to explore correlation between multi-modal features or build relationships between these features. Bayesian models and visual attention models are most commonly used in state-of-art approaches. image-question co-attention is added to exploit multi-modal relationship between pixels and words, which successfully boosts accuracy. Built a Multi- Layer Perceptron (MLP) model with question-grouped training and softmax loss.

1.3 SCOPE

VQA has many potential real-world applications, among which one of the most valuable is to assist visually impaired individuals in understanding contents of images from the web. Moreover, VQA could be a great system to use in the Visual Turing Test or other computer vi-

sion tasks to evaluate the systems ability when comparing with human performance

1.4 CONTRIBUTION

Some of the common approaches are Bayesian models and visual attention models. We propose a novel mechanism that jointly reasons about visual attention and question attention, which we refer to as co-attention. We add image-question co-attention to exploit multi-modal relationship between pixels and words, which successfully boosts accuracy. 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.

1.5 ORGANIZATION OF THESIS

Chapter 1 discussed about the problem domain,problem description,scope and contribution. Chapter 2 discusses the various VQA approaches in detail and assumptions made in the implementation of the system. Chapter 3 explains the overall system architecture and the design of various modules along with their complexity.Chapter 4 elaborates on the results of the implemented system and gives an idea of its efficiency. It also contains information about the dataset used for testing and other the observations made during testing. Chapter 5 concludes the thesis and gives an overview of its criticisms. It also states the various extensions that can be made to the system to make it function more effectively.

CHAPTER 2

RELATED WORK

This chapter gives a survey of the possible approaches to visual question answering. Extracting the specific dataset for the model plays an important role. For example, for co-attention model we have used Visual7w dataset that includes questions,answers and images. The accuracy depends on the model and dataset we use.Thus,this survey helped us to choose the co-attention model that boosts accuracy.

2.1 VISUAL QUESTION ANSWERING

Joulin et al.,[6] have analysed the performance of both baseline and stateof-the-art VQA models, including Multi-modal Compact Bi-linear pooling (MCB), neural module networks, and recurrent answering units.

Sukhbaatar et al.[4] have established how attention helps certain categories more than others, determine which models work better than others.Analyze existing VQA algorithms using a new dataset called the Task Driven Image Understanding Challenge (TDIUC), which has over 1.6 million questions organized into 12 different categories.

Hallonquist et al.[7] have introduced questions that are meaningless for a given image to force a VQA system to reason about image content and proposed new evaluation schemes that compensate for over-represented question-types and make it easier to study the strengths and weaknesses of algorithms.Analyzed the performance of both baseline and state-of-the-art VQA models, including MCB, neural module networks, and recurrent answering units.

2.2 SIMPLE BASELINE FOR VISUAL QUESTION ANSWERING

Batra J et al. [8]have claimed that artificial data can be used to address visual question answering as a complement to current practice. With the help of existing deep linguistic processing technology that enable to create challenging abstract datasets, which enabled to investigate the language understanding abilities in detail.The simple, abstract domain and the controlled generation process based on randomly sampling microworlds makes such data comparatively unbiased and greatly reduces the possibility of hidden complex correlations.

S. Ren K. He et al.[3] have described a semantic link between textual descriptions and image regions by object-level Grounding Enables a new type of question answering with visual answers,in addition to textual answers used. The VQA tasks in a grounded setting with a large collection of 7W multiple-choice question answering pairs.A great progress in basic perceptual tasks such as object recognition and detection. However, AI models still fail to match humans in high-level vision tasks due to the lack of capacities for deeper reasoning. Recently the new task of visual question answering has been proposed to evaluate a model's capacity for deep image understanding. Previous works have established a loose, global association between question answering sentences and images. However, many questions and answers, in practice, relate to local regions in the images that establish a semantic link between textual descriptions and image regions by object-level grounding. It enables a new type of question answering with visual answers, in addition to textual answers used in previous work.The visual question answering tasks in a grounded setting with a large collection of 7W multiple-choice question answering pairs. Furthermore,evaluates human performance and several baseline models on the question answering tasks. Finally,proposed

a novel Long-Short Term Memory(LSTM) model with spatial attention to tackle the 7W QA tasks.

K. Kafle et al.[1] have described a very simple bag-of-words baseline for visual question answering. This baseline concatenates the word features from the question and CNN features from the image to predict the answer Evaluated on the challenging VQA Dataset, it shows comparable performance to many recent approaches using recurrent neural networks.Recent advances in computer vision have brought us close to the point where traditional object-recognition benchmarks such as Imagenet are considered to be solved. These advances, however, also prompt the question how it can move from object recognition to visual understanding; that is, how it can extend todays recognition systems that provide us with words describing an image or an image region to systems that can produce a deeper semantic representation of the image content. Because benchmarks have traditionally been a key driver for progress in computer vision, several recent studies have proposed methodologies to assess our ability to develop such representations. These proposals include modeling relations between objects, visual Turing tests, and visual question answering.

2.3 DEEP RESIDUAL LEARNING FOR IMAGE RECOGNITION

R. Szeliski et al. [9] have proposed a residual learning framework to ease the training of networks that are substantially deeper than those used previously. Explicitly reformulate the layers as learning residual functions with reference to the layer inputs, instead of learning not referenced functions.When deeper networks are able to start converging, a degradation problem has been exposed: with the network depth increasing, accuracy gets saturated (which might be unsurprising) and then de-

grades rapidly. Unexpectedly, such degradation is not caused by over fitting, and adding more layers to a suitably deep model leads to higher training error.

M. Sabou et al.[2] have described a procedure for constructing and learning neural module networks, which compose collections of jointly-trained neural modules into deep networks for question answering. Decomposes questions into their linguistic substructures, and uses these structures to dynamically instantiate modular networks neural module networks,which provide a general-purpose framework for learning collections of neural modules which can be dynamically assembled into arbitrary deep networks. M. Bernstein et al.[5] have demonstrated that this approach achieves state-of-the-art performance on existing datasets for visual question answering, performing especially well on questions answered by an object or an attribute. Additionally, it had introduced a new dataset of highly compositional questions about simple arrangements of shapes.

CHAPTER 3

SYSTEM DESIGN

3.1 DATASET DESCRIPTION

The dataset for the proposed system consists of 43700 images with almost 3 questions per image and its corresponding possibilities of answers. The total count of questions sums to 248349 questions in the training dataset and 121512 questions in the testing dataset.The dataset is downloaded from the visual7w dataset where the images are from MSCOCO dataset.

3.2 PROPOSED SYSTEM

We propose a novel co-attention mechanism for VQA that jointly performs question-guided visual attention and image-guided question attention. We propose 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 be adaptive and select the phrase sizes whose representations are passed to the question level representation. Finally, we evaluate our proposed model on two large datasets, Visual 7w and COCO. We also perform ablation studies to quantify the roles of different components in our mode. Softmax is done over the output scores of all. Compared with the binary logistic

loss, softmax with cross-entropy loss is demonstrated to train better and converge more quickly in our experiments.

3.3 SYSTEM ARCHITECTURE

This model is based on Facebooks baseline model but with several key differences, as follows.Figure 3.1 and 3.2 shows the architecture of our implemented model.The dataset for the proposed system consists of 43700 images with almost 3 questions per image and its corresponding possibilities of answers. The total count of questions sums to 248349 questions in the training dataset and 121512 questions in the testing dataset.The dataset is downloaded from the visual7w dataset where the images are from MSCOCO dataset.

Firstly, questions and answers are represented by Bag-Of-Words feature of pre-trained Word2Vec Embeddings which are then fine-tuned in the training process.The image features are extracted using VGGnet and the dimension of image features is 2048. The multi-layer perceptron model is constructed with two hidden layers, 1024 hidden units in each layer and dropout layers in the middle for regularization. The final layer is a softmax layer, and is responsible for generating the probability distribution over the set of possible answers. We have used the categorical_crossentropy loss function since it is a multi-class classification problem. The rmsprop method is used for optimization.

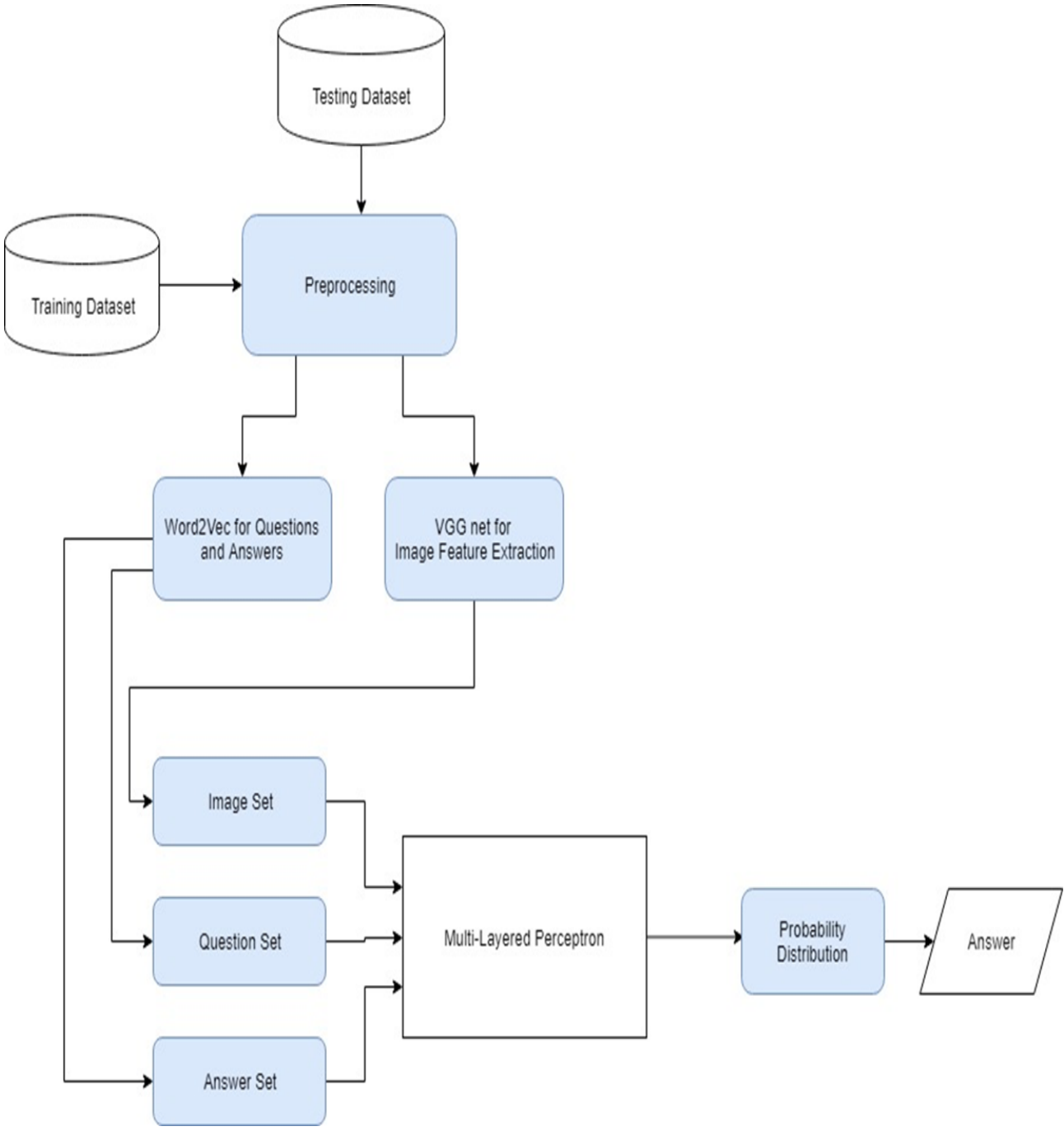


Figure 3.1 Architecture Diagram

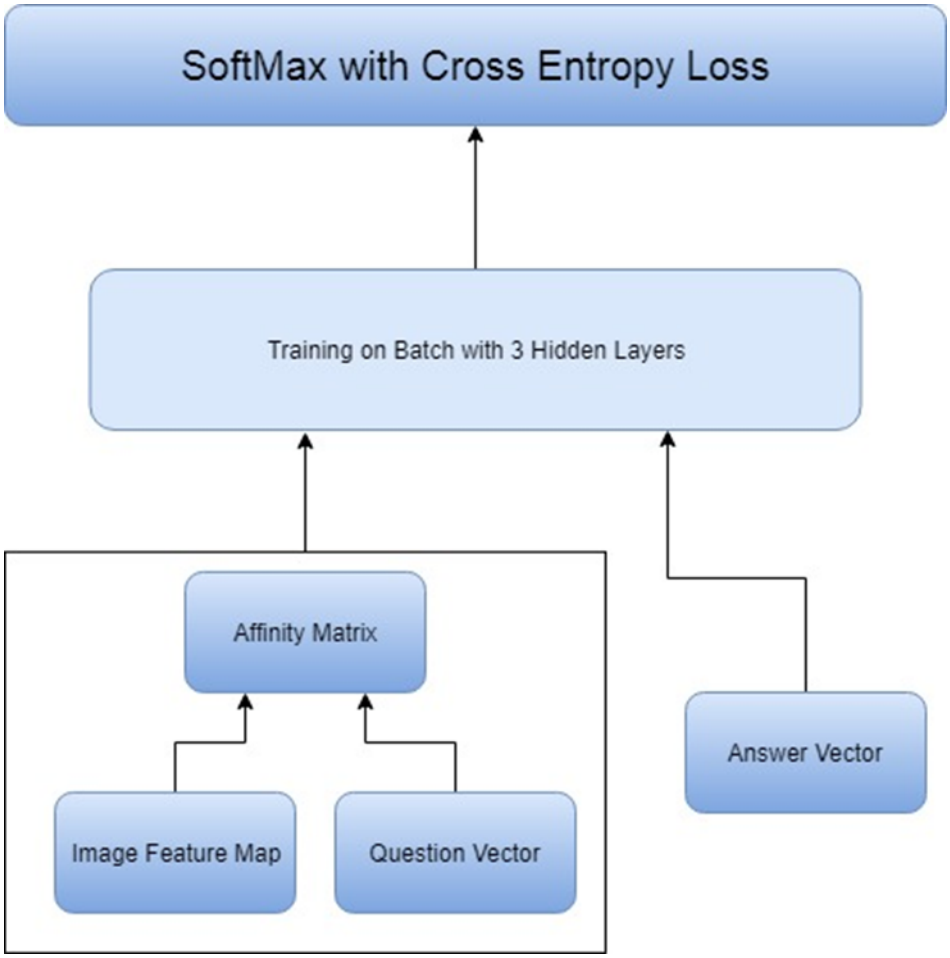


Figure 3.2 Architecture Diagram

3.4 LIST OF MODULES

1. Preprocessing
    1.1 Vector Representation
2. Feature Extraction
3. Co-Attention Model
    3.1 Affinity Matrix
    3.2 Attention Vector
    3.3 Weighted Image and Question Map
4. Score Generation
    4.1 Image-Question-Answer pair
    4.2 Activation Function

3.5 MODULE DESIGN

3.5.1 Preprocessing

We evaluate our proposed model on Visual7W Telling dataset . This dataset includes 248349 training questions, 121512 test questions. Each question starts with one of the six W s, what, where, when, who, why and how, and has four answer choices. The three wrong answers are human-generated on a per-question basis. And the performance is measured by the percentage of correctly answered questions. All the questions and its corresponding answers are taken from the Visual7w questions dataset and Visual7w annotations dataset and joined together.
Input: Raw dataset with images , questions and answers
Output: Modified dataset with image path,question id,question and answers

Algorithm 1 explains about the preprocessing.

Algorithm 1 Preprocessing

Load annotations dataset
while *each question_id* **do**
| retrieve image_id store array list of answer objects in answers
end
while *each answer* **do**
| find max occurence;
end
store array list of answer objects in answers

Vector Representation

Word2vec is a group of related models that are used to produce word embeddings. These models are shallow, two-layer neural networks that are trained to reconstruct linguistic contexts of words. Questions

and answers are represented as Bag of Words and converted to vectors using Word2Vec Embeddings.

3.5.2 Feature Extraction

Feature extraction involves extracting a higher level of information from raw pixel values that can capture the distinction among the categories involved. This feature extraction is done in an unsupervised manner wherein the classes of the image have nothing to do with information extracted from pixels. After the feature is extracted, a classification module is trained with the images and their associated labels.The Visual Geometric Groups (VGG) convolutional layers are followed by 3 fully connected layers. The width of the network starts at a small value of 64 and increases by a factor of 2 after every sub-sampling/pooling layer. Deeper neural networks are more difficult to train. We present a residual learning framework to ease the training of networks that are substantially deeper than those used previously. Images are represented as the penultimate layer feature of VGGnet.

Input: Images , Questions and Answers

Output: Image Features and Vector representation for questions and answers

Algorithm 2 explains about the Feature extraction.

Algorithm 2 Feature Extraction

For every image input
Image passed through Deep Convolutional Network
Activation function extracted in second layer
Size of feature vector is 4096

3.5.3 Co-Attention Model

We propose 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 image co-attention terms are being incorporated into our model to represent the relationships between image and question. Specifically, in the image we would decide Where to look at using the spatial attention terms, and as for the questions, we will have a sense of What are we asking inorder to choose informative words within the sentence.

Input: Image features, Question Vectors

Output: Weighted image and question map.

Affinity Matrix

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.,the affinity matrix C is calculated by the equation 3.1

$$C = tanh(Q^T W_b V) \tag{3.1}$$

where W_b contains the weights.

Attention Vector

We will train the attention vector of image feature av and the attention vector of question feature aq using an one-hidden layer neural network a v [n] = maxi(C_i,n) and a q [t] = maxj (C_t,j). Instead of choosing the maxactivation, we find that performance is improved if we consider this affinity matrix as a feature and learn to predict image and question attention maps via the following equations 3.2 and 3.3

$$H^v = tanh(W_v V + (W_q Q)C), H^q = tanh(W_q Q + (W_v V)C^T) \quad (3.2)$$

$$a^v = softmax(w_{hv}^T H^v), a^q = softmax(w_{hq}^T H^q) \quad (3.3)$$

where W_v,W_q , w_{hv}, w_{hq} are the weight parameters. a v R_N and a q R_T are the attention probabilities of each image region v_n and word qt respectively.The affinity matrix C transforms question attention space to image attention space

Weighted Image and Question Map

We compute the weighted image feature and question feature as in the equations 3.4 and 3.5:

$$\hat{v} = \sum_{n=1}^{N} a_n^v v_n \quad (3.4)$$

$$\hat{q} = \sum_{t=1}^{T} a_n^q q_t \quad (3.5)$$

After that, the new weighted feature v and q are concatenated with the original answers. Once the weighted image and question feature are obtained we can concentrate on the parts of the image based on the questions asked. So when a question is asked , the relationship between the question and image is obtained using the co-attention model and the specific part of the image is concentrated.

3.5.4 Score Generation

Image-Question-Answer Pair

MLP model takes image-question-answer triplet as input,and outputs the probability of this triplet being true. Both questions and answers are represented as Bag-Of-Word features of pre-trained word2vec embeddings . Images are represented as the penultimate layer feature of VGG net. These features are concatenated and fed into an MLP model with one hidden layer

Activation Function

Softmax is done over the output scores of all four answers. Softmax with cross-entropy loss is demonstrated to train better and converge more quickly in our experiments. Equations 3.6, 3.7 and 3.8 represents the softmax loss.Each image-question-answer triple goes through the following MLP:

$$z_1 = W_1 x_{iqa} + b_1 \tag{3.6}$$

$$h_1 = max(0, z_1) \tag{3.7}$$

$$s = W_2 h_1 + b_2 \tag{3.8}$$

CHAPTER 4

RESULTS AND DISCUSSION

4.1 RESULTS

We proposed a hierarchical co-attention model for visual question answering. Co attention allows our model to attend to different regions of the image as well as different fragments of the question.We implemented a basic MLP structure as our baseline and further explored some tweaks to improve the models performance. Then co-attention terms are computed as weights on question features, image features. We model the question hierarchically at three levels to capture information from different granularity. Through visualizations, we can see that our model co-attends to interpretable regions of images and questions for predicting the answer. Though our model was evaluated on visual question answering, it can be potentially applied to other tasks involving vision and language

4.2 IMPLEMENTATION DETAILS

4.2.1 Dataset

The dataset from Visual7w containing question set and annotation set for a particular set of image is shown in Figure 4.1 and Figure 4.2

{"info": {"description": "This is v1.0 of the VQA dataset.", "url": "http://visualqa.org", "version": "1.0", "year": 2015, "contributor": "VQA Team", "date_created": "2015-10-02 19:28:08"}, "task_type": "Open-Ended", "data_type": "mscoco", "license": {"url": "http://creativecommons.org/licenses/by/4.0/", "name": "Creative Commons Attribution 4.0 International License"}, "data_subtype": "train2014", "questions": [{"question": "What shape is the bench seat?", "image_id": 487025, "question_id": 4870250}, {"question": "Is there a shadow?", "image_id": 487025, "question_id": 4870251}, {"question": "Is this one bench or multiple benches?", "image_id": 487025, "question_id": 4870252}, {"question": "Is this a modern train?", "image_id": 78077, "question_id": 780770}, {"question": "What color is the stripe on the train?", "image_id": 78077, "question_id": 780771}, {"question": "What is on the other side of the train?", "image_id": 78077, "question_id": 780772}, {"question": "Is the bus driver on any kind of antidepressant medication?", "image_id": 501867, "question_id": 5018672}, {"question": "Is the bus moving?", "image_id": 501867, "question_id": 5018670}, {"question": "What color is the bus?", "image_id": 501867, "question_id": 5018671}, {"question": "Are these items for sale?", "image_id": 529524, "question_id": 5295240}, {"question": "What is for sale under this tent?", "image_id": 529524, "question_id": 5295241}, {"question": "Is this a grocery store?", "image_id": 529524, "question_id": 5295242}, {"question": "Is the dog on the bed?", "image_id": 497494, "question_id": 4974940}, {"question": "What is in the frame over the bed?", "image_id": 497494, "question_id": 4974941}, {"question": "Where is the dog?", "image_id": 497494,

Figure 4.1 VQA Question Dataset

{"answer": "twisting", "answer_confidence": "no", "answer_id": 9}, {"answer": "curved", "answer_confidence": "maybe", "answer_id": 10}], "image_id": 487025, "answer_type": "other", "question_id": 4870250}, {"question_type": "is there a", "multiple_choice_answer": "yes", "answers": [{"answer": "yes", "answer_confidence": "yes", "answer_id": 1}, {"answer": "yes", "answer_confidence": "yes", "answer_id": 2}, {"answer": "yes", "answer_confidence": "yes", "answer_id": 3}, {"answer": "yes", "answer_confidence": "yes", "answer_id": 4}, {"answer": "yes", "answer_confidence": "yes", "answer_id": 5}, {"answer": "yes", "answer_confidence": "yes", "answer_id": 6}, {"answer": "yes", "answer_confidence": "yes", "answer_id": 7}, {"answer": "yes", "answer_confidence": "yes", "answer_id": 8}, {"answer": "yes", "answer_confidence": "yes", "answer_id": 9}, {"answer": "yes", "answer_confidence": "yes", "answer_id": 10}], "image_id": 487025, "answer_type": "yes/no", "question_id": 4870251}, {"question_type": "is this", "multiple_choice_answer": "1", "answers": [{"answer": "1", "answer_confidence": "yes", "answer_id": 1}, {"answer": "multiple", "answer_confidence": "yes", "answer_id": 2}, {"answer": "1", "answer_confidence": "maybe", "answer_id": 3}, {"answer": "1", "answer_confidence": "yes", "answer_id": 4}, {"answer": "multiple",

Figure 4.2 VQA Annotation Dataset

4.2.2 Preprocessed Dataset

The question, image and answers are extracted from the dataset and stored separately as shown in Figure 4.3 and Figure 4.4

```
What is the table made of?
Is the food napping on the table?
What has been upcycled to make lights?
Is this an Spanish town?
Are there shadows on the sidewalk?
What is in the top right corner?
Is it cold outside?
What is leaning against the house?
How many windows can you see?
Is this in a park?
Is there a bicycle in this picture?
Is the person feeding the birds?
Is this a Girl Scout?
What uniform is she wearing?
What color is the fence?
What color is the linoleum?
Is the water running in the sink?
How is the floor made?
What is the teddy bear sitting on?
Do children like this object?
What is written on the teddy bear's feet?
Is the weather warm in this picture?
How many people are in this photo?
Why would the snowmobiler be riding up the mountain for the
skier?
How many people can the red buses hold?
Are the red buses identical?
How many double-decker buses are in the picture?
Where is this picture?
```

Figure 4.3 Preprocessed Questions Dataset

curved
yes
1
no
white
trees
no
yes
red
yes
vegetables
no
yes
picture
on bed
gold
right
widow
stool
kitten
mouse
cheese
cheese
yes
venice
5 mph
no
brown
black
white
calm

Figure 4.4 Preprocessed Answers Dataset

4.2.3 Feature Extraction

Questions and answers are represented as Bag of Words and converted to vectors using Word2Vec Embeddings as shown in Figure 4.5

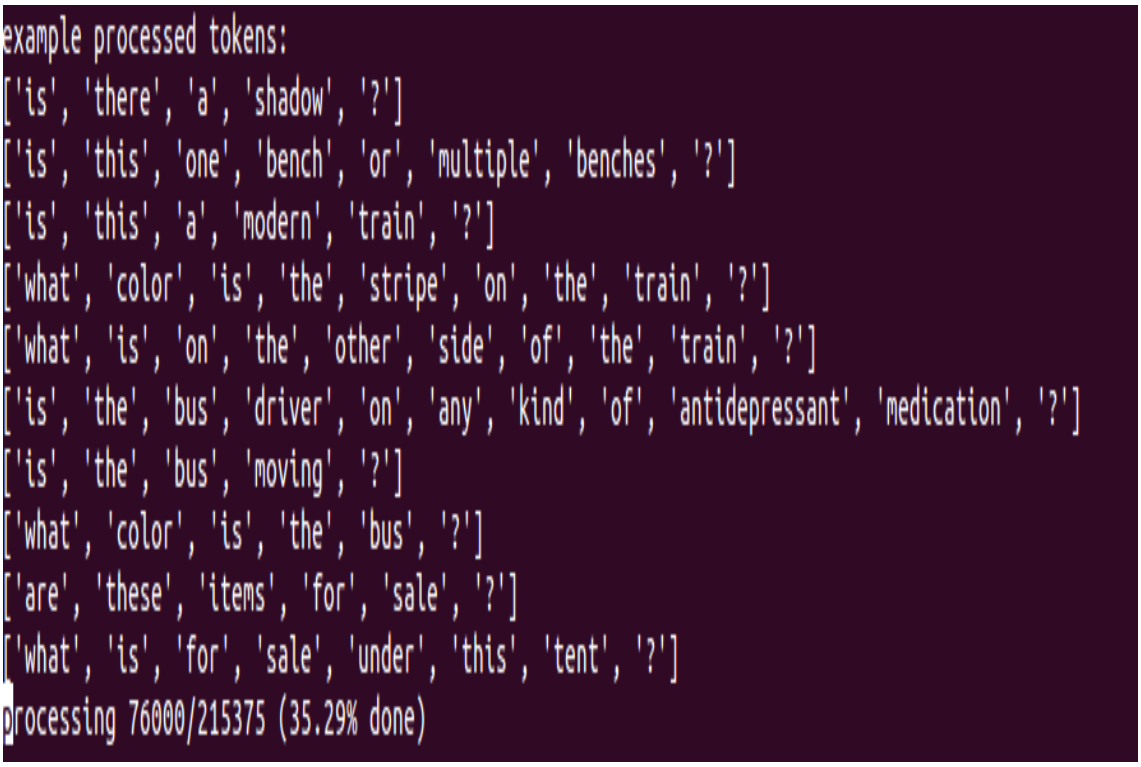


Figure 4.5 Word Features

We present a CNN to ease the training of networks that are substantially deeper than those used previously. Image Features are represented using the VGG net as shown in Figure 4.6

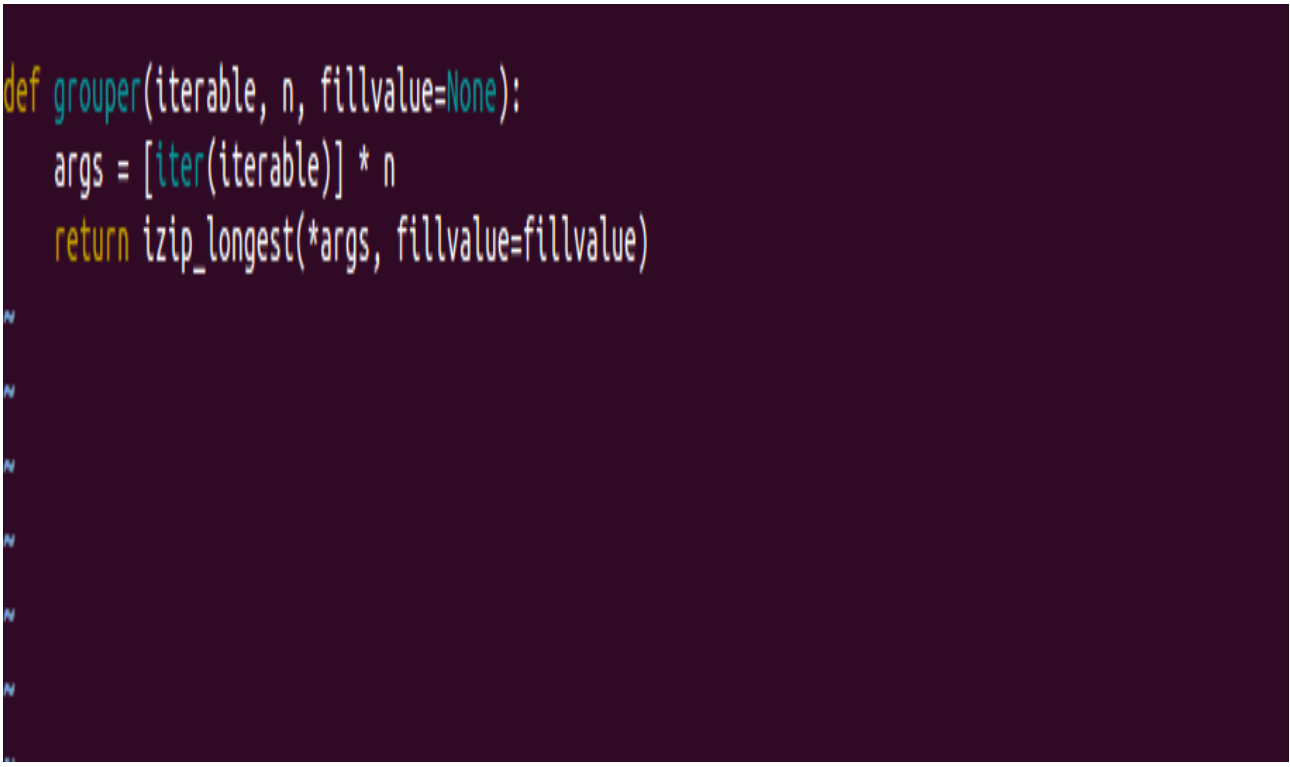
Table

	0	1	2	3	4	5	6	7	8	9	10	11	
0	0.728793...	0.986336...	-0.00986...	-0.2207372	-1.3585073	1.0168936	-0.9551489	-0.9307153	1.048407	0.176789...	0.2832436	-1.5647954	0.4
1	1.0756949	0.075593...	-1.4424669	-1.2479157	-2.2220876	-1.3448822	0.293678...	0.9360543	-0.87165...	-1.9137027	-0.8161369	1.8530742	-0.
2	-0.33970...	-0.5766505	-0.83542...	-1.7762364	0.5455329	-0.7219636	-0.42064...	-0.7813213	0.329485...	-1.0450093	0.7598525	-0.06151...	0.7
3	-0.947105	-0.7927845	-0.3722799	-1.1096088	0.8305281	-1.1630827	0.663290...	0.201222...	-1.4560167	-1.0355661	0.8718845	-1.2742928	1.2
4	-0.3061524	-0.9991422	2.2137175	0.132905...	0.6108385	-0.49633...	-0.2205112	-0.52788...	-0.45920...	0.496966...	0.7372744	-2.4245508	0.5
5	-0.47932...	0.6819535	2.1548762	0.058969...	0.8258646	2.6130865	0.358633...	-1.0336216	0.7820809	-0.4745762	-1.7337455	-0.38960...	-1.
6	0.254095...	0.436597...	-0.48289...	-0.4220609	2.2068504	0.6865043	0.5096441	-0.8140911	-0.53059...	-0.64893...	-0.82898...	0.450097...	1.4
7	1.1275936	-0.8463132	-0.00585...	-0.22642...	-1.0263612	-0.8517554	0.9081781	-0.19461...	-0.3317587	0.030766...	-2.0987537	0.283378...	-0.
8	-0.3719247	-0.59280...	-0.00764...	-0.40142...	0.3412581	0.460686...	0.8152571	-0.568858	0.3675996	-0.3784035	-0.51603...	-0.73838...	-0.
9	0.101241...	-0.5172637	-0.8206364	-1.1783596	1.5739114	-0.04229...	0.166688...	-0.04273...	-1.2916782	0.8158739	0.151775...	-0.5788078	1.1
10	-0.61941...	1.2645993	-1.2397602	-0.46991...	-0.91012...	0.6774856	-0.7833695	-0.33768...	-0.5483232	-1.2421031	-0.54416...	-0.39480...	0.9
11	-2.347987	-0.33657...	0.7060952	0.494454...	-0.12533...	0.004012...	1.5026324	-1.1384739	0.188797...	-1.0093808	1.5835283	-0.01109...	1.6
12	-0.19026...	1.5152558	-1.3855714	-0.91772...	-0.82300...	-0.2720406	-0.37178...	-0.27066...	0.210711...	0.081323...	-0.5729728	-0.13151...	-0.
13	0.480084...	0.303997	0.103391...	-0.087477	-1.2191426	-0.13600...	-1.1991309	0.5251754	0.1588409	-1.3279618	0.559543...	0.4510318	-0.
14	0.2680723	-0.08363...	2.0356555	-0.41446...	-0.35971...	-1.3216816	0.4367033	-0.8578966	1.456285	-0.7517315	-0.663311	-0.42335...	0.0
15	0.206633...	0.328613...	-0.09082...	1.1383455	0.578490...	-2.3114262	-0.49417...	0.217640...	-0.6229095	0.4187681	1.7133201	0.235138...	0.2
16	-0.11217...	0.431923...	1.0004141	-0.357587	0.246263...	-0.9309398	0.086595...	0.616662...	0.391240...	0.150778...	-0.20327...	-1.1154056	-0.
17	1.321974	0.9259951	-0.09815...	0.044597...	-0.51658...	0.9781574	-0.8217583	-0.01791...	1.0716587	0.9436129	1.1509969	-1.7267301	-1.
18	-1.0718209	0.7509072	2.5338135	0.754566	-0.8802726	-0.47299...	-0.73642...	-0.0625853	-0.26954...	-0.7195768	-0.09554...	0.397623...	-1.
19	-0.38373...	0.319992...	-0.09038...	-0.45492...	-0.27803...	-0.02160...	-0.9915986	0.578574	0.5364598	-0.3182197	1.4610752	0.861560...	0.3
20	-0.41851...	-0.32315...	0.3683172	0.7197046	-0.48920...	0.195462...	-0.936833	-0.08769...	-1.6094257	1.3323743	-2.0101244	-1.1098034	-0.
21	-1.8075566	-1.1092081	0.099963...	1.5462774	0.5705457	-0.85852...	-0.44409...	-0.26421...	-0.23917...	0.9082774	-0.05381...	0.9845942	0.5

Figure 4.6 Image Features

4.2.4 Train Model

Figure 4.7 and Figure 4.8 show the training code of our implemented model



```
def grouper(iterable, n, fillvalue=None):
    args = [iter(iterable)] * n
    return izip_longest(*args, fillvalue=fillvalue)
```

Figure 4.7 Grouper Code

```
print 'Training started...'
for k in xrange(args.num_epochs):
        index_shuf = range(len(questions_train))
        shuffle(index_shuf)
        questions_train = [questions_train[i] for i in index_shuf]
        answers_train = [answers_train[i] for i in index_shuf]
        images_train = [images_train[i] for i in index_shuf]
        progbar = generic_utils.Progbar(len(questions_train))
        for qu_batch,an_batch,im_batch in zip(grouper(questions_train, args.batch_size, fillvalue=questions_train[-1]),
                                              grouper(answers_train, args.batch_size, fillvalue=answer
s_train[-1]),
                                              grouper(images_train, args.batch_size, fillvalue=images_
train[-1])):
                X_q_batch = get_questions_matrix_sum(qu_batch, nlp, c_id,q_id)
                X_i_batch = get_images_matrix(im_batch, id_map, VGGfeatures)
                X_batch = np.hstack((X_q_batch, X_i_batch))
                                                                                95,1-8        85%
```

Figure 4.8 MLP Training Code

4.2.5 Output

Figure 4.9 shows the final output of the model.

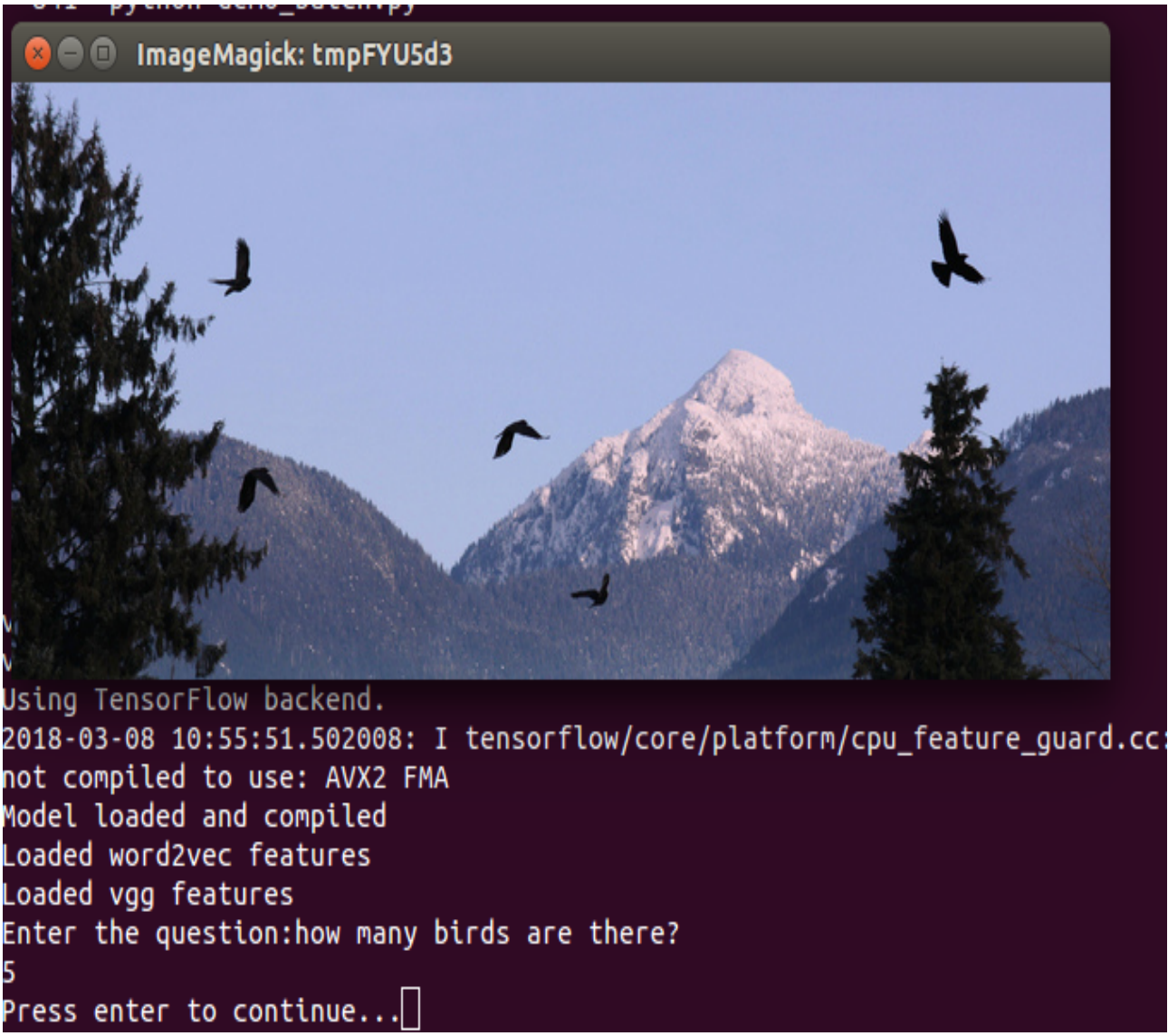


Figure 4.9 Final Output

4.3 PERFORMANCE EVALUATION

The Performance is evaluated using the standard parameters given below.

4.3.1 Precision Score

Precision is the number of correct positive results divided by the number of all positive results returned by the classifier as in equation 4.1. Figure 4.10 shows the precision score of the system.

$$Precision = \sum TruePositive \div \sum PredictedConditionPositive \quad (4.1)$$

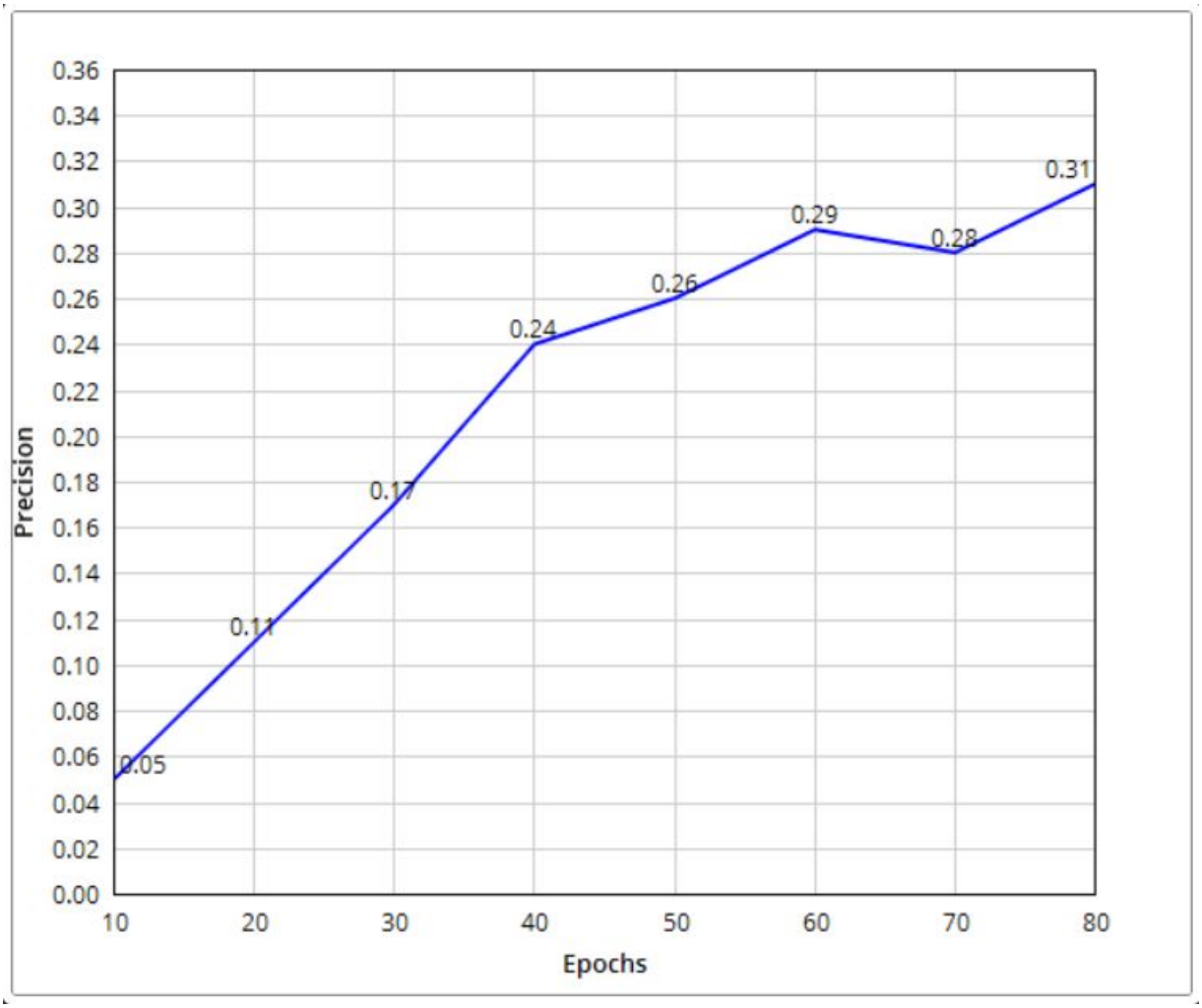


Figure 4.10 Precision Score for the proposed system

4.3.2 Recall Score

Recall r is the number of correct positive results divided by the number of all relevant samples (all samples that should have been identified as positive as in equation 4.2. Figure 4.11 shows the recall score of the system.

$$Recall = \sum TruePositive \div \sum ConditionPositive \tag{4.2}$$

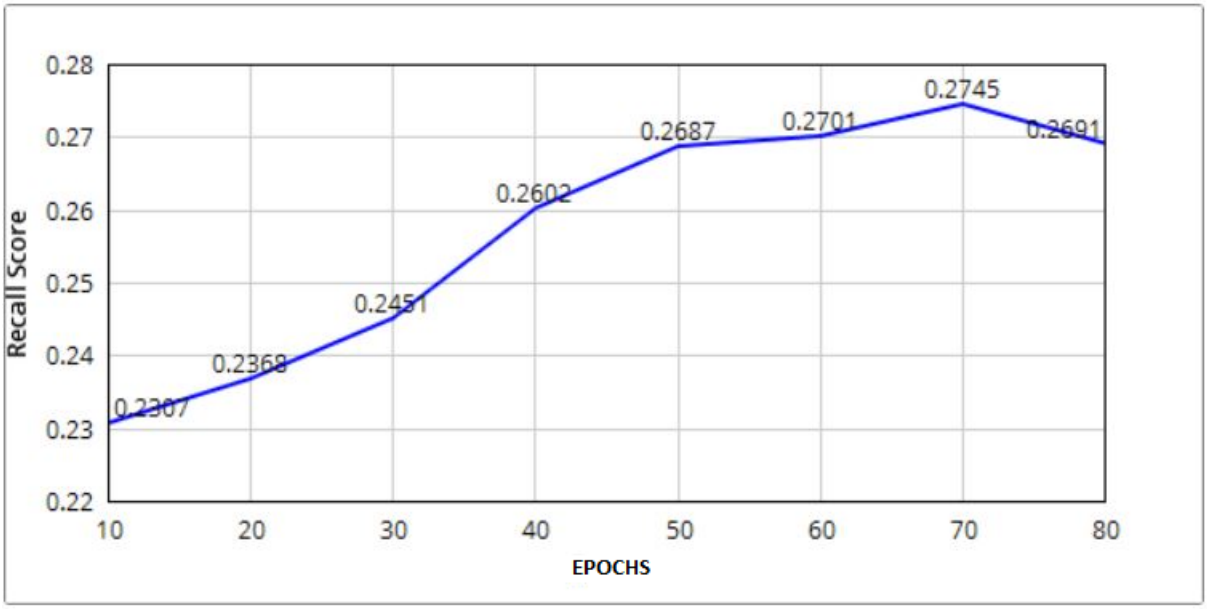


Figure 4.11 Recall Score for the proposed system

4.3.3 F1 Score

The F1 score is the harmonic average of the precision and recall as in equation 4.3, where an F1 score reaches its best value at 1 (perfect precision and recall) and worst at 0. Figure 4.12 shows the F1 score of the system.

$$F1Score = 2 * precision * recall \div precision + recall \tag{4.3}$$

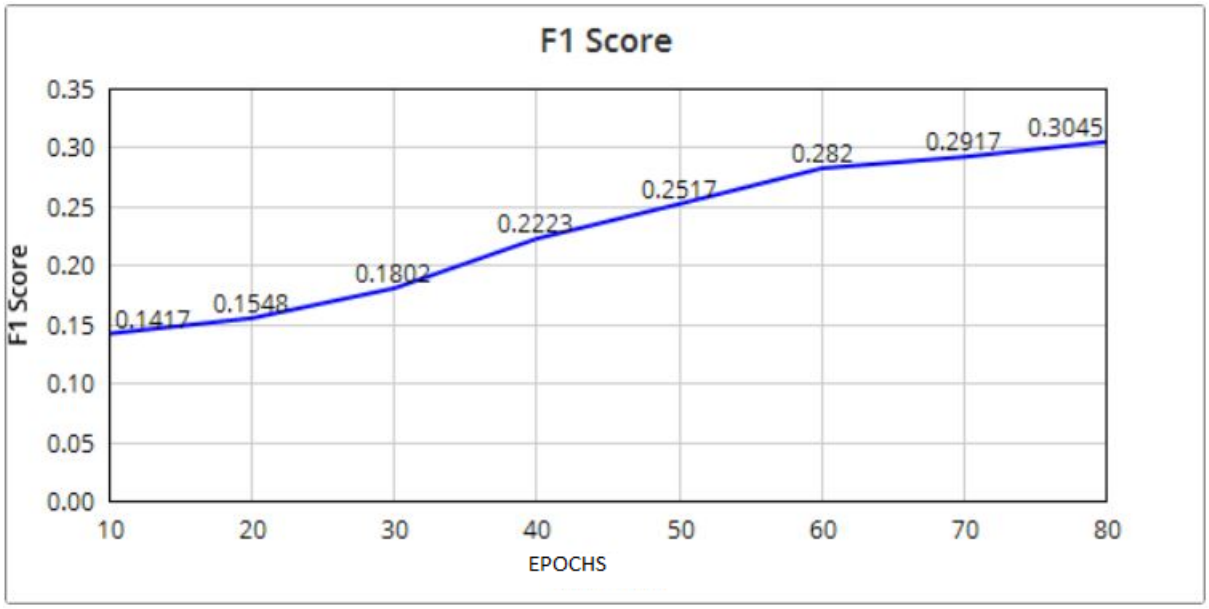


Figure 4.12 F1 Score for the proposed system

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 CONTRIBUTIONS

Many challenges lie in this complicated problem. First, we have to extract embedding matrix for words in questions/answers and promising image features to support any further training on our proposed model. Also, whatever methods we use to extract multi-modal features, image features and text features are from different feature spaces which means it is very difficult to incorporate them in the same model. One possible solution is to explore correlation between multi-modal features or build relationships between these features. Bayesian models and visual attention models are most commonly used in state-of-art approaches.

The project focuses on the task of Visual Question Answering where text-based questions are generated about an given image, and the goal is to produce. We implemented a basic MLP structure as our baseline and further explored some tweaks to improve the models performance. Then co-attention terms are computed as weights on question features, image features and trained along with other parameters.

5.2 FUTURE WORK

It may be helpful to find a way to incorporate intra-sentence attention and spatial attention into the model instead of only considering co-attention terms. This has many real world implications as an aid to visually impaired humans. It can be used to describe the visual scenery

as seen by a camera and used to describe it to the person. It can also be used to help such persons to view whatever is displayed on the screen.

REFERENCES

[1] A. Joulin A. Jabri and L. van der Maaten, "Revisiting visual question answering baselines", *European Conference on Computer Vision,*, pp. 727-739, Springer,2016.

[2] S. Sukhbaatar A. Szlam B. Zhou, Y. Tian and R. Fergus, "Simple baseline for visual question answering", *arXiv preprint arXiv:1512.02167*, 2015.

[3] N. Hallonquist D. Geman, S. Geman and L. Younes, "Visual turing test for computer vision systems", *Proceedings of the National Academy of Sciences,*, vol. 112, no. 12, pp. 3618-3623,2015.

[4] D. Batra J. Lu, J. Yang and D. Parikh, "Hierarchical questionimage co-attention for visual question answering", *in Advances In Neural Information Processing Systems*, pp. 289 - 297,2016.

[5] S. Ren K. He, X. Zhang and J. Sun, "Deep residual learning for image recognition", *Proceedingsof the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778,2016.

[6] K. Kafle and C. Kanan, "Visual question answering: Datasets, algorithms, and future challenges", *arXiv preprint arXiv:1610.01465*, 2016.

[7] R. Szeliski, "Computer vision: algorithms and applications", 2010.

[8] M. Sabou V. Lopez, V. Uren and E. Motta, "Is question answering fit for the semantic web?:a survey", *Semantic Web, vol. 2, no. 2, pp.*

125-155, 2011.

[9] M. Bernstein Y. Zhu, O. Groth and L. Fei-Fei, "Visual7w: Grounded question answering in images", *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.