Ali Hamza & Raja Muneer
01-134181-009 & 01-134181-056

Visual Question Answering Using Deep Learning

Bachelor of Science in Computer Science

Supervisor: Dr.Imran Siddiqui

Department of Computer Science
Bahria University, Islamabad

June, 2021

Certificate

We accept the work contained in the report titled “Visual Question Answering using deep learning ”, written by Mr Raja Muneer Ali Hamza as a confirmation to the required standard for the partial fulfilment of the degree of Bachelor of Science in Computer Science.

Approved by . . . :

Supervisor: Dr.Imran Siddiqui

Internal Examiner: Name of the Internal Examiner (Title)

External Examiner: Name of the External Examiner (Title)

Project Coordinator: Dr. Moazam Ali

Head of the Department: Dr. Arif Ur Rahman

November, 2021

Acknowledgments

In the name of Allah, the most beneficent and the most merciful. Foremost, we would like to express our sincere gratitude to our supervisor Dr. Imran Siddiqui for his continuous support during project and for making us understand the importance of deep learning as well as in enhancing our understanding of the project Visual Question Answering. We would like to thank him his for his patience, motivation, enthusiasm and immense support and knowledge. His guidance has helped us throughout the project and in writing this project report.We could not have imagined a better supervisor and mentor for our project.

Besides, our supervisor we would like to extend our sincere thanks to all the faculty members who have helped us throughout this project for their effort, time and their dedication and also class mates who have been a constant support through thin and thick times.

Finally, we would like to extend our heartiest thanks to our families and especially our parents for supporting and believing in us throughout our lives.

RAJA MUNEER & ALI HAMZA
Islamabad, Pakistan

November, 2021

“AND MY SUCCESS CAN ONLY COME FROM ALLAH. IN HIM I TRUST, AND UNTO HIM I RETURN”

AL-QURAN

Abstract

The problem of answering questions about an image is commonly known as visual question answering. It is a well-established problem in computer vision. The Visual Question Answering(VQA) task requires the understanding of both text and vision. Given an image and a question in natural language, the VQA system tries to find the correct answer to it using visual elements of the image and inference gathered from textual questions. These types of models are really helpful for visually-impaired people to get information about the surrounding environment or certain set of images.

In the recent years there are many VQA systems developed by using different techniques of computer vision, Natural language processing, and deep Learning. Mostly these VQA systems are developed on scenes images and very few models work on textual features of images but learning textual features plays an important role in predicting an answer. Computer vision experts can develop an AI system for blind or visually-impaired people, so they can ask questions regarding any particular scene and environment. Using textual features in the model can help in predicting more accurate answers.

There is a lot of textual data which are present in images, which can be used for very useful predictions.However, most of the VQA methods does not utilize the text often present in the images. These "texts in images" provide additional useful cues and facilitate better understanding of the visual content.

In our project to develop a VQA system, we approached such a method that uses textual features along with visual features to predict an answer. In our project we use a books cover dataset which contains 207k images of the book cover and contains more than 1 million questions-answers. Our model detected text from images, objects from images and created a co-related mechanism to represent their features. We have used pre-trained CNN and Bilstm. OCR is used to retrieve text from images, pre-trained CNN is used to represent images and LSTM is used for Question representation. We have used OCR results for Name entity recognition, word to vector from the text and combined them with the results of CNN, LSTM to feed a composite vector into fully connected layers. We used different deep learning methods, functions, and approaches to build our VQA system. The experimental results and rigorous analysis demonstrate various challenges present in this dataset leaving ample scope for the future research.

Contents

List of Figures

List of Tables

Acronyms and Abbreviations

VQA	Visual Question Answering
CNN	Convolution Neural Network
RNN	Recurrent Neural Network
LSTM	Long short term memory
Bi-LSTM	Bi-directional long short term memory
VGG	Visual Geometry Group
EAST	An Efficient and Accurate Scene Text Detector
OCR	Optical Character Reader
DNN	Deep Neural network
FCC	Fully connected layer
VGG	Visual Geometry Group
MSE	Mean Square Error
Tf	Tensorflow
NER	Name entity Recognition
DL	Deep learning
ML	Machine Learning
API	Application Programming Interface
UI	User Interface
UX	User Experience

Chapter 1

Introduction

1.1 Introduction

Visual Question Answering (VQA) is the task of answering questions about a given piece of visual content such as an image, video, or infographic. In a general way, we can define a VQA system as a system that takes as input an image and a natural language question about the image and generates a natural language answer as the output. This is by nature a multi-discipline research problem. We need NLP for at least two reasons: to understand the question and to generate the answer. As seen in Figure 1.1 the image with a natural language question is given as an input and the deep learning model needs to process the image with a question to answer it correctly. The answer could be in any form of the following: numbers, phrases, words, or a sentence. an initial model that combined the image and question representation and merge the results to predict the answer.

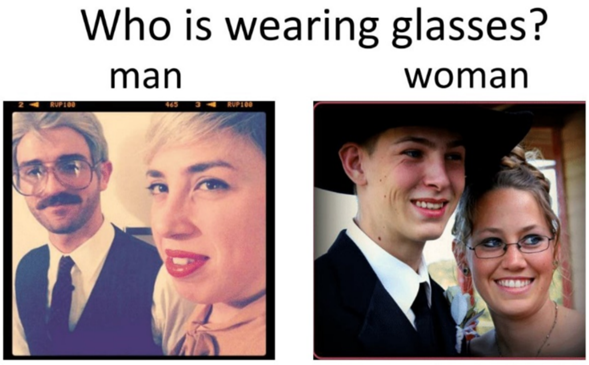


Figure 1.1: Sample Image VQA Problem

1.2 Problem Description

These days a lot of data is available in the form of images, videos and text. Today this data is used to solve many problems. A lot of data is present in images which can be used to solve many problems in the field of computer vision, natural language processing. In our project, we are using the text data from images and by using this data we are making a Visual question answering system. These "texts in images" provide additional useful cues and facilitate a better understanding of the visual content. In our project, we are using the data which is present in images, and we are primarily focused on text data that is present in images. By using this data, we are making a VQA model. There are good datasets available for visual question answering. These data sets are more focused on real-world scenes. But we are using a data set that is more focused on textual data which is present in images. Many problems in the field of computer vision and natural language processing can be solved using this data. As humans, it is easy for us to see an image and answer any question about it using our commonsense knowledge. However, there are also scenarios, for instance, a visually-impaired user or an intelligence analyst, where they want to actively elicit visual information given in an image. The objective of our project is to develop an artificial intelligence system that can answer the question asked by users. Most of the previous studies show that visual question answering models only focused on visual features and ignored the textual features in a given scene of an image. We need an optimal system that considers the features of text along with the visual features.

Image features are one of the most important pieces of information for a VQA system to output the correct answer. The exact location of the object or text of an image plays a vital role in the prediction of an accurate answer. In our project computer vision and deep learning techniques are used to identify the relevant part of the image and extract its features.

The dataset which are we going to use in our project is OCR-VQA-200K [1]. It is an open-source dataset which is developed by the Indian Institute of Science, Bangalore, India. This dataset comprises 207K images of the book cover and contains more than 1 million question-answer pairs about these images. The book's cover consists of texts, figures, and tables. by using this data set our VQA system predicts the name of a given book, the name of the author, publish date of the book, and other things which are mentioned in the book cover. This type of diverse information in images makes it more challenging for the VQA community. An example image from our dataset is shown in the Figure 1.2.

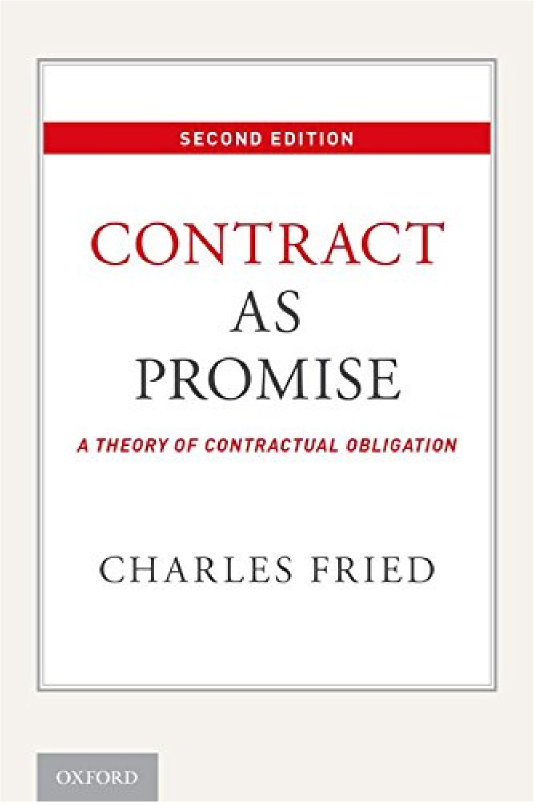


Figure 1.2: Sample Image from OCR-VQA 200K

There are different types of questions that can be used to question an image. Some examples of questions and their predictive answer are given below:

Q: What is the title of this book?

A : CONTRACT AS PROMISE

Q: Who is the author of this book?

A : CHARLES FRIED

Q: What is the edition of this book?

A : SECOND EDITION

1.3 Deep Learning

Deep learning is a subset of machine learning. Machine learning is a subset of artificial intelligence. A deep neural network is simply a shallow neural network with more than one hidden layer. Each neuron in the hidden layer is connected to many others. The word 'deep' in deep learning is attributed to these deep hidden layers and derives its effectiveness from it. Selecting the number of hidden layers depends on the nature of the problem and the size of the data set. There are some popular approaches to deep learning are included Deep Feedforward Neural Networks (D-FFNN), Convolutional Neural Networks (CNNs), Deep Belief Networks (DBNs), Autoencoders (AEs), and Long Short-Term Memory (LSTM) networks. Deep learning has a plethora of applications in almost every field such as health care, finance, and image recognition. Our problem VQA also uses deep learning. Our proposed solution has used core algorithms and techniques of deep learning for image understanding and text understanding.

1.4 Project Objectives

The objectives of this project include:

- To propose a predictive model that can extract textual features of an image to predict an answer
- To learn the visual and textual knowledge from the inputs (image and question respectively)
- To combine the computer vision and Natural language processing

1.5 Proposed Method

In this project, we have proposed a method that uses modern techniques of computer vision, natural language processing, and deep learning. Our proposed model takes a question and image as an input and predicts an answer as an output. In our proposed method we use textual features as well as visual features to predict an answer. Our model detects text in images and creates a co-related mechanism to represent the features. We have used different deep learning techniques on our model to train. We have used pre-trained CNN, BiLstm, OCR to represent images, questions and retrieve text from images. We apply word2vec, NER-tagging on a text which we retrieve from OCR and combine the results of word2vec, NER-tagging with the results of CNN and Bilstm to feed a composite vector into a fully-connected layers.Our proposed performed well and achieve good accuracy on this dataset irrespective of its font and layout analysis.

1.6 Document Organization

The Project report comprises seven chapters.
Chapter 2 includes the literature work and research work done by other researchers in proposing the VQA model.
Chapter 3 includes the requirements, user requirements, system requirements, functional, non-functional requirements of our proposed system.
Chapter 4 includes the architecture and details of our proposed methodology.
Chapter 5 includes System implementation in which tools, languages are mentioned which are used in our Proposed model.
Chapter 6 comprises of testing of our proposed model. The last chapter concludes our project work and future work.

Chapter 2

Literature Review

VQA has gained huge interest in recent years. Improvement in image classification and introduction of large-scale visual question answering benchmarks, such as VQA and VQA v2.0 have played important role in triggering a push in VQA research. Different methodologies have been proposed to improve the performance of models.

2.1 Datasets for Visual Question Answering

VQA is an emerging field in terms of research and development in computer vision and Deep learning. For further development, different datasets are made and released by the VQA community. Different work is done by the VQA community on different datasets but every dataset has some limitations and problems. We will mention few of them in Table 2.1 and will see what are the problems and limitations of these datasets.

Table 2.1: Existing Datasets For VQA

	DAQUAR	COCO-QA	COCO-VQA	Visual-7W
Total Images	795 train, 654 Test	82783 Train,40504 Test	82783 Train,81434 Test	47300
QA pairs	10620 Train, 5970 Test	78736 train,38948 Test	248349 Test,244302 test	327939
Distinct Answers	968	430	145172	25553
Longest Question	25 Words	24 Words	32 Words	24 Words
Longest Answer	7 Word List	1 Word	17 Words	20
Answer Format	Words	Single Word	Words, Phrases	Sentences
Image Source	NYUDv2	COCO	COCO	COCO
QA Generation	Human + Algorithm	Algorithms	Human	Human
Answer Types	Color, Object,	object, Number,location	Yes/No,number	7-W

2.1.1 DAQUAR

DAQUAR (The Dataset for Question Answering on Real-world images) is a collection of question-answer pairs for the NYU Depth V2 dataset. The dataset is available in two configurations. DAQUAR-FULL consists of 6795 (train) and 5673 (test) question-answer pairs. DAQUAR-37 has 37 object categories, and 3825 (train) and 297 (test) question-answer pairs. One of the limitations of DAQUAR is that it contains exclusively indoor scenes, which constrains the variety of questions available. Moreover, the images often contain significant clutter and numerous small objects, making some questions very difficult to answer. Even people have significant difficulty with this dataset, with humans only achieving 50.20 percent accuracy on DAQUAR-FULL. It is concluded that the dataset is very challenging and it is very difficult to train the machine on this dataset and achieve good accuracy.

2.1.2 COCO-QA

COCO-QA contains 78,736 QA pairs for training and 38,948 pairs for testing. All questions belong to one of four categories: object (69.84 percent), number (7.47 percent), color (16.59 percent), and location (6.10 percent)

2.1.3 COCO-VQA

COCO-VQA is the subset of The VQA Dataset that has been created from real-world images drawn from COCO.COCO-VQA consists of 614,163 human-generated free-form questions with 6,141,630 human responses (10 responses per question). Question-answer pairs are often complex and rich, and answering many of them requires object recognition or activity recognition as well as knowledge-based reasoning.

2.1.4 Visual7W

The Visual7W dataset contains 327,939 questions for 47,300 COCO images. Visual7W consists of multiple-choice question-answer pairs consisting of six types of ’W’ questions (what, when, who, why, where, how).In addition to the question-answer pairs, Visual7W has annotated bounding boxes that refer to the objects referred to in question-answer pairs.

2.2 OCR-VQA 200k

This dataset is open source dataset which contains 207K images and around 1 million question Answer pairs.This dataset contains the images of books covers.In this dataset authors proposed a method to use a text which is present in images. This was the first dataset in VQA that that reads a text from images. The summary of OCR-VQA 200k is given in table Table 2.2

Table 2.2: Summary of OCR-VQA–200K statistics:

Number of Images	207,572
Number of QA pairs	1,002,146
Number of unique authors	117,378
Number of unique titles	203,471
Number of unique answers	320,794
Number of unique genres	32
Average question length (in words)	6.46
Average answer length (in words)	3.31
Average number of questions per image	4.83

2.3 Related Work

Although VQA is a new problem, different algorithms for VQA are already being deployed. Most existing papers on VQA have used Long-Short-Term-Memory (LSTM) neural networks.Authors [2], [3], [4], and [5] used LSTM networks to encode the question and combined the question encoding with image features from a deep convolutional neural network (CNN).A similar approach is followed by other researchers with minor changes in the VQA model. Most of the authors used LSTM neural networks. In [15], a similar approach was taken, with the main difference being that they fed CNN features to the LSTM as the first "word," followed by vectors encoding each word of the sentence, and then finally the last word was the CNN features once more. In a variant of this approach, In authors[12] sequentially gave their LSTM network concatenated CNN and word features at every time step. In [4], separate LSTMs were used for the question and answer, but they had a shared word embedding layer, and CNN image features are fused at the end of the LSTM. Their model was able to output more than one-word answers or lists, and it could generate coherent sentences.

From the above study, we found that VQA has gained a lot of attention from the researchers. Different techniques of computer vision, natural language processing, and deep learning have been implemented. In another work[6] authors a transformer-based architecture for VQA Model. The architecture that is proposed in this paper was a localization-aware answer prediction network(LaAP-Net). The purpose of this architecture was to improve the positions of image texts and push the network to better exploit visual features. This transformer-based model takes the questions embedding and OCR embedding as input. In this model, questions are fed through pre-trained model BERT which is a transformer-based model. There are three main components of LaAP-Net architecture which is utilized to develop a VQA system. These are the context-enriched OCR representation, the localization-aware predictor, and the transformer with the simplified decoder. They used three main challenging datasets in the experiment of their proposed architecture. These datasets were TextVQA, ST-VQA, and OCR-VQA. Their proposed network LaAP-VQA network performs state of art on these main datasets of VQA.

In another study[7] researchers of Mircosoft proposed a model for VQA. Their proposed model in this paper was Relation-aware Graph Attention Network(ReGAT), which encodes each image into a graph and models multi-type inter-object relations via a graph attention mechanism, to learn question-adaptive relation representations. Their experiments in this paper showed that their proposed model performed outclass rather than that approaches which were used on both VQA2.0 and VQA-CP v2 datasets. Different VQA models were implemented on the VQA2.0 dataset but they could not reach the accuracy which they achieved by using the model. The accuracy of their proposed model on VQA2.0 was 70.58% which is considered a good score in the VQA community.

2.4 Visual Question Answering in Medical Domian

In one of the paper[8] authors proposed a VQA model on medical images. In this paper, they used the VQA-Med dataset. In their proposed model, they used the Inception-Resnet-v2 model to generate features of images. They apply some enhancement techniques on images as well on their dataset. In pre-processing, they also remove some useless information. In their proposed architecture medical images are transformed into the features through the ResNet-v2 network while questions are fed through the embedding layer and Bi-LSTM layer. After concatenating the features of images and questions add another Bi-LSTM layer in the network. While in the fully connected layer(FC) they apply softmax activation function for output prediction.

2.5 Doc VQA

Visual question Answering has a diversity. There are different types of data set in the VQA community. After scene detection, the VQA community introduce a dataset on Documented images. It gained a lot of attraction from the research community. The "Document Visual Question Answering" (DocVQA) challenge, focuses on a specific type of Visual Question Answering task, where visually understanding the information in a document image is necessary to provide an answer. Doc-Vqa dataset is used that contains images of scanned documents and question-answer pairs with the results. They have used LayoutLM for the document layout segmentation and CNN for image feature extraction. VQA is an emerging and very effective approach to artificial intelligence. Different computer vision, Machine learning, and deep learning techniques have been used by researchers to implement VQA purposed predictive models. Different techniques have different results in VQA based systems[9]. There is a lot of textual data present in images that can be used and can be very helpful. There is very little work done on the text which is present in images. Now we will see the work which is done to make such VQA models that consist of such data which is present in images. This research paper [10] focuses on the text which is present in images and designed such a VQA model which can read the text in images and answer questions by reasoning over the text and other visual content. Several existing datasets which we have seen in Table 2.1 study text detection or parsing in natural everyday scenes. These do not involve answering questions about the images or reasoning about the text. This paper[8] was more focused on the text which was present in images. They used the TextVQA dataset which contains 34602 training images, and 5000,5734 validation, test images respectively. While this dataset contains 45,336 questions and 453,360 answers. They also introduce Look, Read, Reason and Answer (LoRRA), a novel model architecture for answering questions based on text in images.LoRRA reads the text in images, reasons about it based on the provided question, and predicts an answer from a fixed vocabulary or the text found in the image. Now there are some research starts on such type of VQA models which works on textual data that is present in an image and give a detailed description of the scene which is present in the image.

2.6 Discussion

As we can see from the above discussion about different architectures and models of computer vision and deep learning used by the research community of VQA to predict an answer about the image for a given question. We can see various computer vision models are used to identify and locate features of the image and few have focused on textual features in the image to jointly use them for answer prediction. Different pre-trained models like Vgg16, Vgg19, ResNet50 are used for the image classification while LSTM, RNN, and Bi-LSTM are used for question representation. Due to advanced development in deep learning, deep learning techniques are getting popular in research and development.

Chapter 3

Requirement Specifications

3.1 Existing System

There are many VQA systems which are built in recent years but they focus on scene images and in recent years many VQA systems are developed on-scene images. For VQA systems there are many good data sets that are released. Some of the datasets are MS COCO, VQA v1.0, VQA v2.0, Visual7w, DAQAUR. But these datasets mainly focus on scenes images. These data set-based VQA Systems focus on scenes and recognize different objects in an image. But these VQA systems ignore a Text in images. This text in images can be very useful and can be used to make an intelligent System.

3.2 Proposed System

Our project visual question answering will focus on the text in images. We are dealing with text which is present in images and we have developed such a VQA system that deals with the text in images rather than scenes that are identified in an image. For example, from a real-world image, we can ask a VQA system a question like who is wearing glasses. VQA system will answer man if a man is wearing glasses. These VQA systems are mainly focused on scenes of images. In our project, we are mainly focused on the text which is present in images. In our project, we have data on books covers. We have 200k images of the book cover. Our VQA system will give all the information which is present on the book cover. For example, the user asks who is the author? title of the book? what type of book? our system will give the above answers. Our system functions in such a way that at first, we have used pre-trained CNN features for visual representation. Then, we developed an OCR which detects text from images using an East text detector and creates a bounding box around a text which is present in the image, the detected text will then pass through a tesseract and it gives a text which is present in the image. Then comes a natural language processing part. For which we used a bi-directional LSTM to represent questions.A generic network architecture of our proposed method is shown in Figure 3.1.

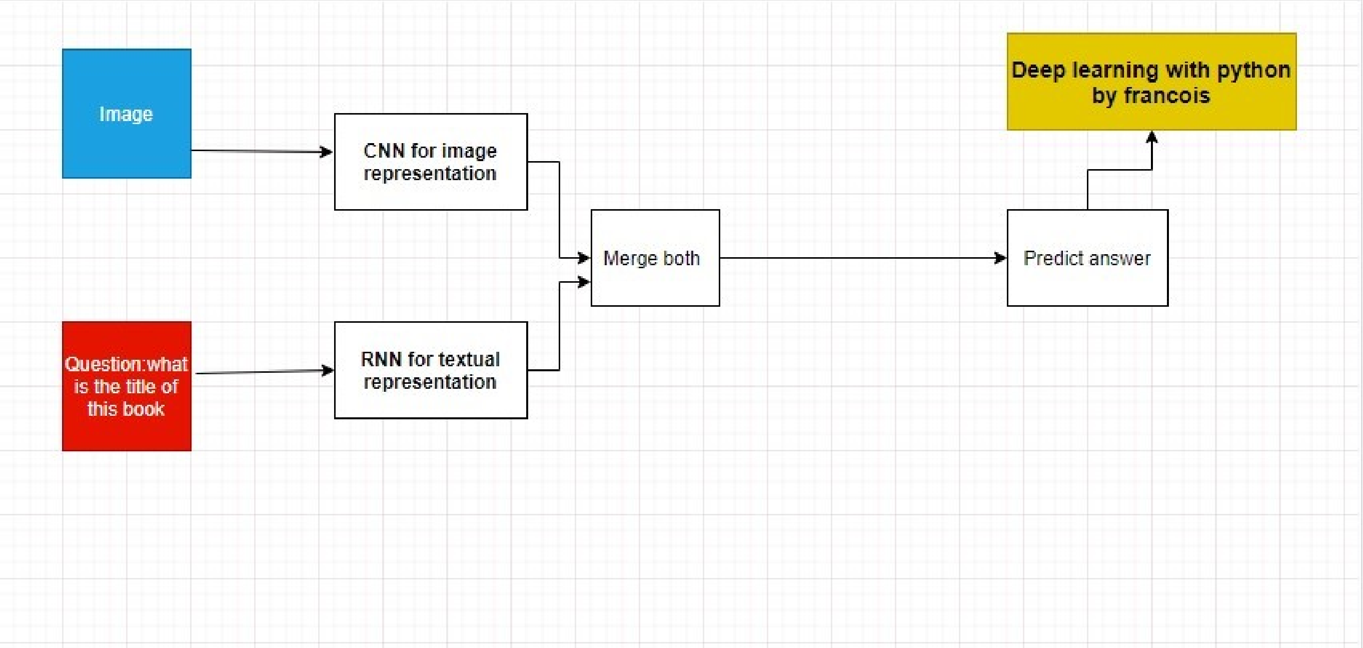


Image
CNN for image representation
Question:what is the title of this book
RNN for textual representation
Merge both
Predict answer
Deep learning with python by francois

Figure 3.1: Network Architecture Diagram

3.3 Functional Requirements

Following are the functional requirements of the system:

- Our system will detect text areas where text is present in the image.
- Our VQA system will recognize a text from a given image using an OCR
- The recognized text will pass through to our LSTM model.
- Our system will answer that question which is asked by a user.

3.4 Non-Functional Requirements

Following are the non-functional requirements of the system:

- Books cover images will be provided
- The user interface must be easy to understand for the user
- The images which will be given to the system should be in good quality
- The system should be efficient

- The system should be accurate
- The system will perform without any delay in time

3.5 Use Case Diagrams

3.5.1 Use Case Diagram 1

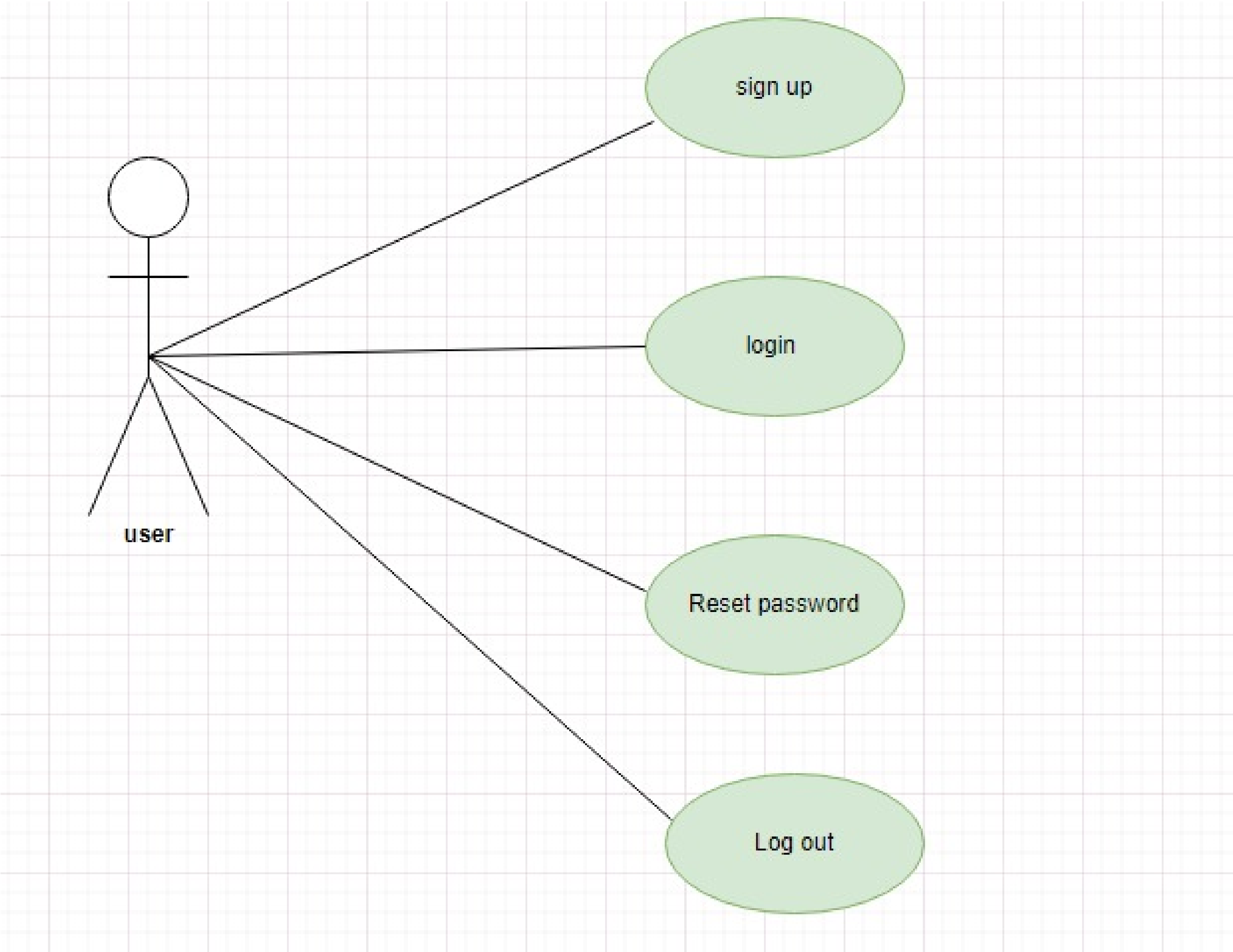


Figure 3.2: Use Case 1

3.5.2 Use Case Diagram 2

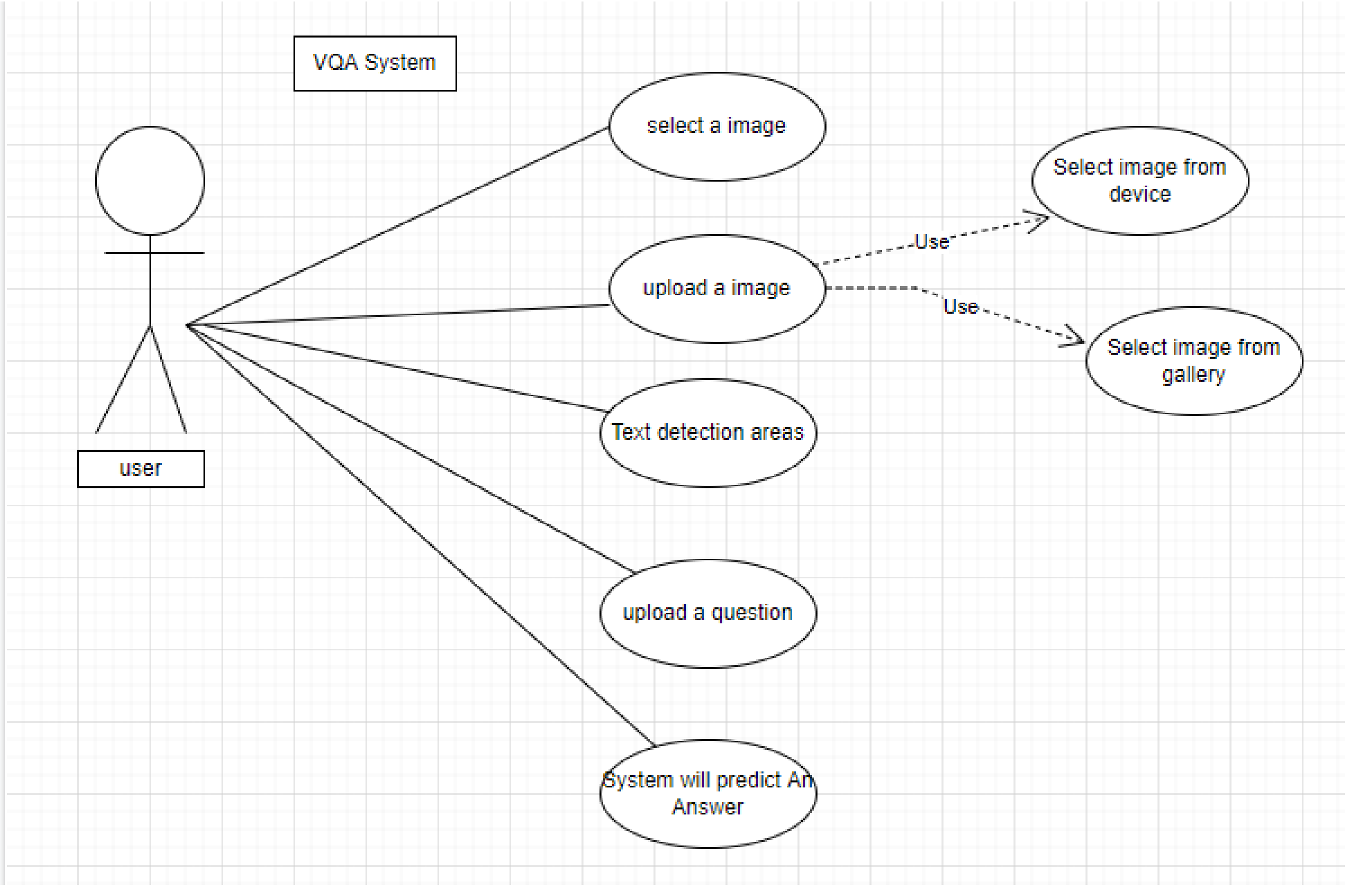


Figure 3.3: Use Case 2

3.6 Use Cases

This section will present the detailed used cases.

Table 3.1: Sing Up UC

Use Case Id:	UC-1
Use Case Name:	Sign Up
Actors:	User
Descryption;	The user will register into an application and want to make an account have to go through signup process
Trigger	A user indicates that he/she wants to sign up into the system
Preconditions:	PRE-1. The user is not already logged in
Postconditions:	POST-1. The user is shown a number of signup options. POST-2. The user is taken to the homepage after successful signup POST-3. The account details of new user are added into the database.
Normal Flow:	1. The new user opens the app 2. The app shows options either to log in or signup. 3. The user selects the signup option. 4. The user has the option to register locally, for this, he has to add credentials. 5. After successful sign up the system takes the user to the homepage.
Alternative Flow:	1. The new user opens he app 2. The app shows option either to log in or signup. 3. The user selects the login option. 4. The user enters the account credentials. 5. The account credentials do not match. 6. The credentials are wrong. 7. The user will not be able to log in. 8. The user will be requested to sign up first.
Exceptions:	1. The new user opens the app 2. The app shows option either to log in or signup. 3. The user selects the login option. 4. The user enters the account credentials. 5. The account credentials do not match. 6. The credentials are wrong. 7. The user will not be able to log in. 8. The user will be requested to sign up first.
Business Rules:	None
Assumptions:	None

Table 3.2: Login UC

Use Case Id:	UC-2
Use Case Name:	Login
Actors:	User
Descryption;	The users who want to login into this application using a local account will go through this use case.
Trigger	A user indicates that he/she wants to login into the system by entering credentials and clicking the login button.
Preconditions:	PRE-1. The user is not already logged in. PRE-2. The user has already signed up.
Postconditions:	POST-1. The user enters correct credentials for logging in to the application. POST-2. The user is redirected to home page after successful login.
Normal Flow:	1. The new user opens the app 2. The app shows options to enter credentials for login and a button to signup in case of new user. 3. The user enters credentials and logins to the system. 5. After successful login the system takes the user to the homepage.
Alternative Flow:	1. The user opens the app. 2. The app shows options to enter credentials for login. 3. The user enters credentials for login. 4. The user is unable to login because user has not register yet. 5. The user is requested to sign up first. 6. The user signup to the application by entering credentials. 7. The user now able to login.
Exceptions:	1. The user opens the app 2. The app shows option either to log in or signup. 3. The user enters credentials for login. 4. The login credentials are wrong. 5. The login credentials do not exist or match. 6. The user is resulted back to step 3.
Business Rules:	None
Assumptions:	None

Table 3.3: Text Detection UC

Use Case Id:	UC-3
Use Case Name:	Text Detection
Actors:	User
Descryption;	The user after successful login has option to choose from two options. 1. Text Detection 2. Visual Questioning The user chooses Text Detection.
Trigger	A user indicates that he/she wants to detect text from an image.
Preconditions:	PRE-1. The user already logged in. PRE-2. The user has already chosen the text detection option.
Postconditions:	POST-1. The user uploads an image. POST-2. The user is directed to final text detection screen.
Normal Flow:	1. The user has successfully logged into the system. 2. The user has selected the text detection option from choose screen. 3. The user uploads an image by browsing his system or selecting from default images. 4. The user after selecting the image clicks detect text button. 5. The detected text from the image is displayed to the user on the final screen.
Alternative Flow:	1. The user has successfully logged into the system. 2. The user has selected the visual questioning mistakenly. 3. The user wants to go back to text detection. 4. The user clicks the back button on top to go back to choose screen. 5. The user selects the text detection option. 6. The user chooses an image and then clicks on detect text button. 7. The detected text from image is displayed to the user.
Exceptions:	1. The user is logged into the system. 2. The user chooses the Text Detection Option. 3. The user doesn't upload an image and clicks detect text button. 4. The user will be prompted to upload an image.
Business Rules:	None
Assumptions:	None

Table 3.4: Visual Questioning UC

Use Case Id:	UC-4
Use Case Name:	Visual Question Answering
Actors:	User
Descryption;	The user after successful login has the option to choose from two options. 1. Text Detection 2. Visual Questioning The user chooses Visual Questioning.
Trigger	A user indicates that he/she wants to question an image.
Preconditions:	PRE-1. The user already logged in. PRE-2. The user has already chosen the Visual Questioning Option.
Postconditions:	POST-1. The user uploads an image. POST-2. The user writes a question related to the image. POST-2. The user is directed to the final Visual Questioning screen.
Normal Flow:	1. The user has successfully logged into the system. 2. The user has selected the Visual Questioning option from choose screen. 3. The user uploads an image by browsing his system or selecting from default images. 4. The user after selecting the image writes the question in the question box. 4. The user after selecting the image and writing the question clicks on Predict Result Button. 5. The Predicted Answer from the system is displayed to the user on the final vqa screen.
Alternative Flow:	1. The user has successfully logged into the system. 2. The user has selected the text detection option mistakenly. 3. The user wants to go back to visual Questioning. 4. The user clicks the back button on top to go back to choose screen. 5. The user then selects the visual questioning option. 6. The user chooses an image and writes a question then clicks on Predict Result Button. 7. The Model Predicted answer is displayed to the user.
Exceptions:	1. The user is logged into the system. 2. The user chooses the Visual Questioning Option. 3. The user doesn't upload an image or lefts the question field empty and clicks Predict Result button. 4. The user will be prompted to upload an image and write a question.
Business Rules:	None
Assumptions:	None

Chapter 4

Design

4.1 System Architecture

VQA system will give an answer against a question about an image from the user using Visual Question Answering using deep learning techniques. This system or these type of systems can help a lot for blind people. The VQA system helps in predicting an answer within a limited about the question asked by a system or a person.The VQA system uses the advance techniques of deep learning , natural processing and computer vision.Visual Question Answering system predict an answer with the combination of image and text .The architecture of our proposed model is shown in the Figure 4.1.

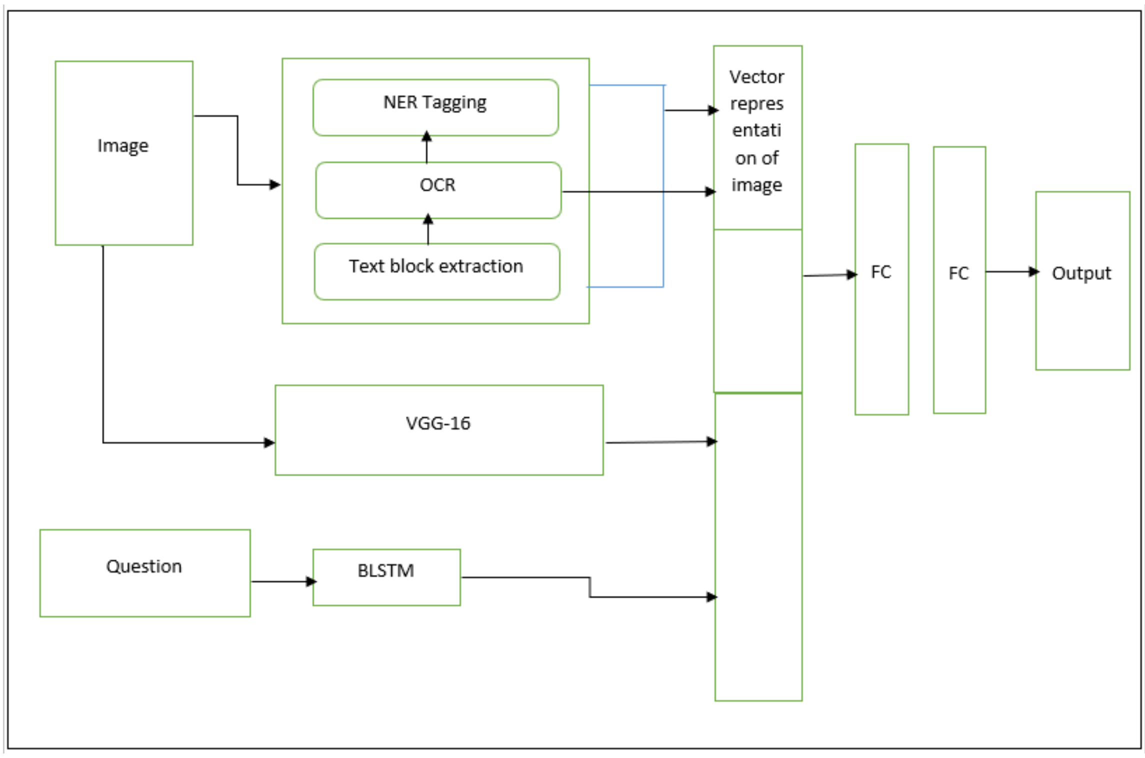


Figure 4.1: Network Architecture Diagram

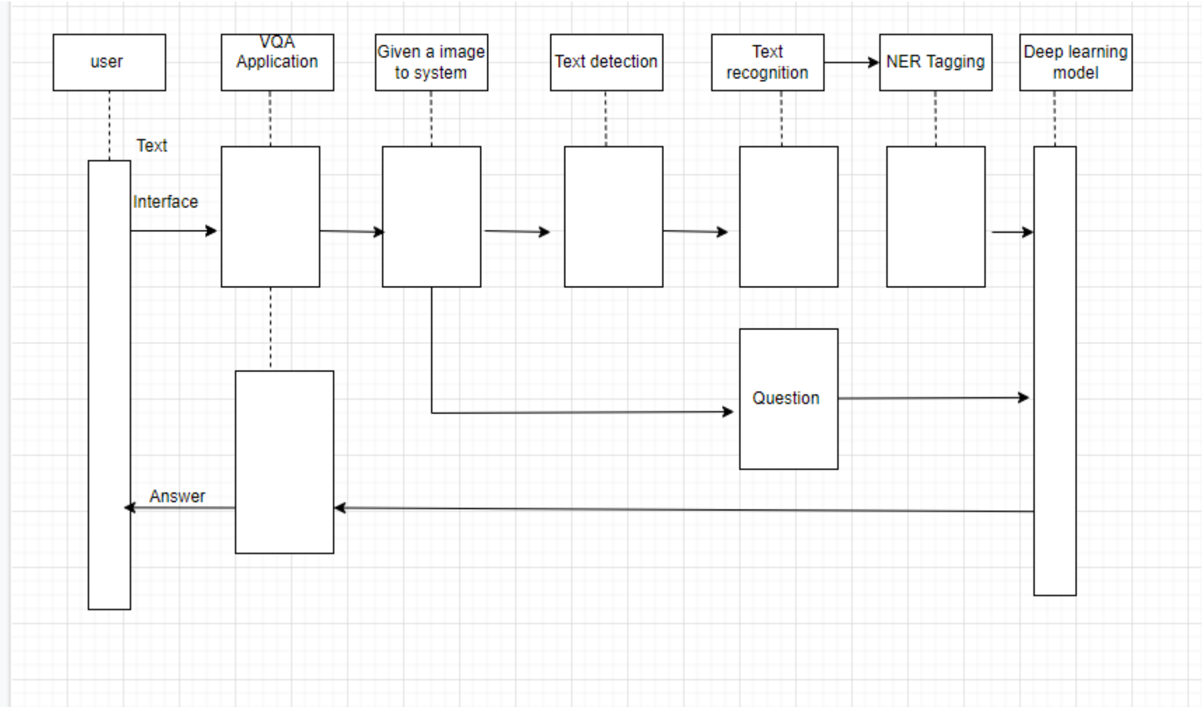


Figure 4.2: Sequence Diagram

To predict an answer from image we have the following steps and methods in our visual question answering system

- Text detection and Extraction
- Optical character Recognition
- Word2Vec
- NER tagging
- Answer Prediction

4.2 Text Detection

In our project of VQA,the first step is to detect the text regions which are present in an image.These text which are present in a image and their corresponding text is represented as features for VQA task.There are many approaches to detect text regions in computer vision like tesseract text block segmentation, EAST and VGG text detector.We use EAST text detector for the detection of text in our VQA task.EAST is a deep learning text detection based system. It is a fast and accurate text detection method.It made a bounding boxes where text is found and get these bounding boxes.We choose maximum five text blocks only, if more than five text box found in a image then we calculate area and pick only those five boxes which are are largest area wise.The results of EAST text detection are shown in Figure 4.2.

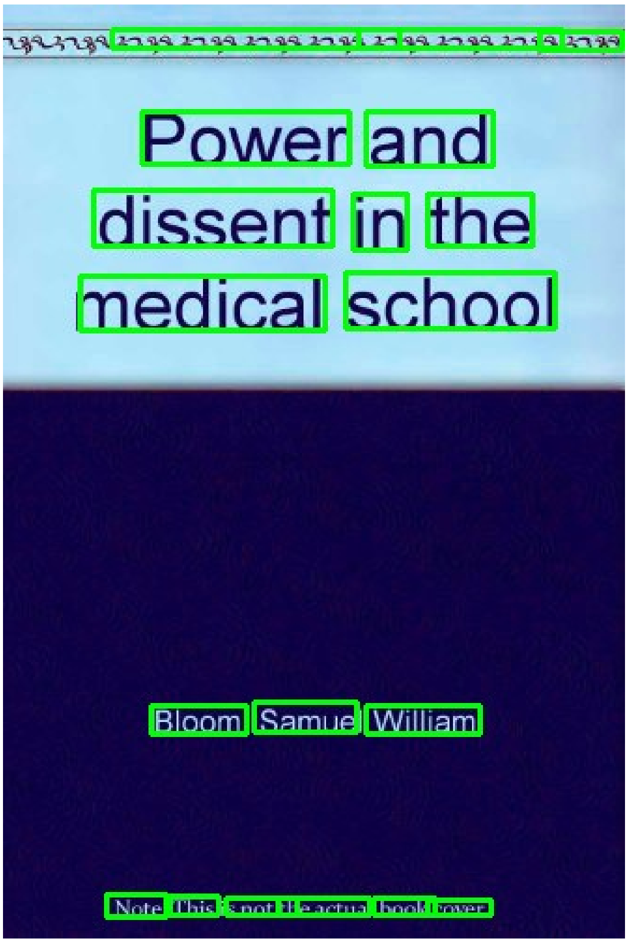


Figure 4.3: EAST Text Detector Results

4.3 Optical Character Recognition

Optical Character Recognition(OCR) allows to extract text from different types of images,such as images of documents, invoices, bills, financial reports, photos of street signs , products and more. There is no OCR which can work perfectly in every task , image .OCR varies problems to problems. In our case OCR is used to extract text from images.Once text blocks are identified using EAST text detector , then we fed these identified text blocks to our OCR Tesseract.The text of these identified blocks are shown in Figure 4.3.

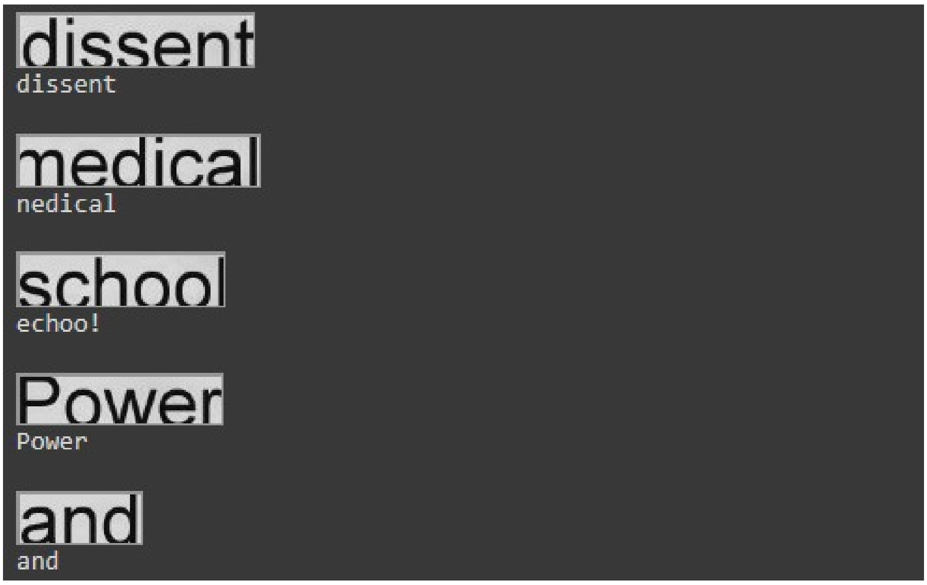


Figure 4.4: Optical Character Recognition

4.4 Word2Vector

After the OCR detect and recognize a text from an image, We have picked the Top-5 bounding boxes area-wise. If the image contains more than 5 bounding boxes we pick only Top-5 bounding boxes area-wise.We represent each text box into the Five-dimensional vector. The vector contains index number of block, top-left x-y coordinates and bottom-right x-y coordinates of text-block location.We have used ”word2vec-google-news-300” gensim pre-trained model. In order to get w2v of word at average location of detected text and this w2v model gave vector representation of word in 300-dimension.The w2v representation is used to present whole text using its average location word.

4.5 NER-Tagging

Named entity recognition (NER) is a well-studied problem in natural language processing. Named entity recognition is a method and a NLP technique to detect the entity from the text.It is used to the retrieve the name entity and categories such as the name of the person, year, edition etc. We use here five dimension vector which we pass word2vec and append it 1 if person name is found otherwise append it with 0, append the vector with 1 if we found year on text block otherwise with 0, and appended with edition number in numerical form, e.g., first to 1, 2nd to 2, etc. if found , otherwise with 0. After the Ner-tagging the vector became 8-dimension for each text-block.Now All of the 8-dimension vectors are then concatenated to form a 40-dimension vector.

4.6 Question Representation using Bi-LSTM

LSTM(Long short term memory) is a type of recurrent network.LSTM can learn the patterns of sequence and predict the target value in each time-step.We represent questions using bidirectional long short term memory (Bi-LSTM) .The question is presented using two layers of BiLstm with 300 units in each and then fed into the dense layer with 1024 units.

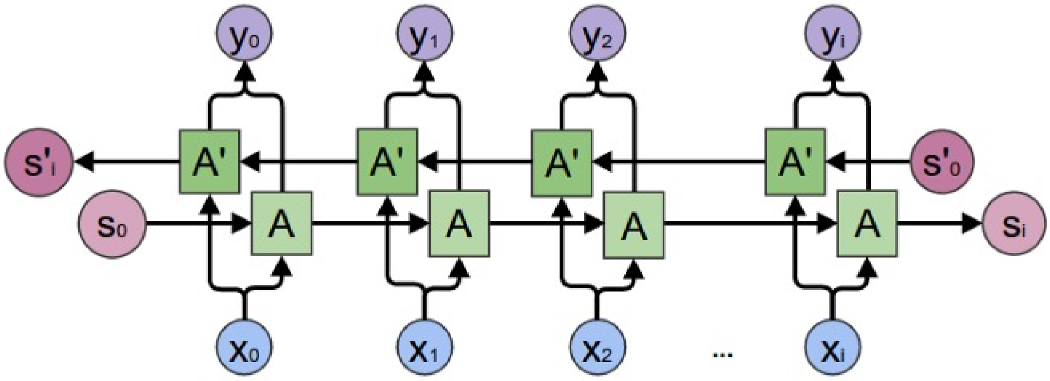


Figure 4.5: BI LSTM

4.7 Network Architecture

Visual question Answering system is implemented using pre-trained Vgg16 for the image representation and Bi Lstm for the question representation. We have used East text detector to extract the text blocks and tesseract ocr engine to recognize the text from detected text blocks. Our network has three inputs for two fully connected layers.

4.8 Model Training

We divide dataset into training, testing and validation part.The training part of the data consists 800k samples. The training is done using 50 epochs with batch size of 90. Here, the goal is to locate and classify named entity mentions in unstructured text into predefined categories such as the person names, organizations, geo-political entities, year, etc. The summary of our model is shown in table 4.1.

Table 4.1: Model Summary

Layer	Output Shape	No of parameters
Vgg16-Block-Dense	(None,4096)	102764544
BLOCK(input layer)	[(None,24)]	0
word2vector(input layer)	[(None, 300)]	0
Bidirectional-lstm	[(None,300)]	0
Joint features FC	[(None, 1024)	1049600]
Fully connected Dense layer1 (FC1)	[(None, 1024)]	4834304
Fully connected Dense layer2 (FC2)	[(None,1024)]	1049600
Dense-softmax	(None,59)	(60475)

4.8.1 Hyper-parameters

The list of hyper-parameters are shown in table 4.2.

Table 4.2: Hyper Parameters

Hyper-parameters	Value
VGG-16 filters	512
Bi-Lstm	300
FClayer Activation function	RELU
FC layer units	1024
output laye activation function	Softmax
Batch size	90
learning rate	0.01
Loss Function	Cross-Entropy

4.9 GUI Interface

4.9.1 Prototype GUI Interface

The application interface is prototyped in figma.The prototype shows that there are two iterations depending upon two options provided in the application which are text detection and Visual Questioning. Iteration 1 is run for only text detection whereas Iteration 2 is run for Visual Questioning.

- **Iteration 1** This iteration will be run for text detection option.

- First Application is launched and the user will be directed to splash Screen.Figure 4.5 shows the prototype of splash screen.

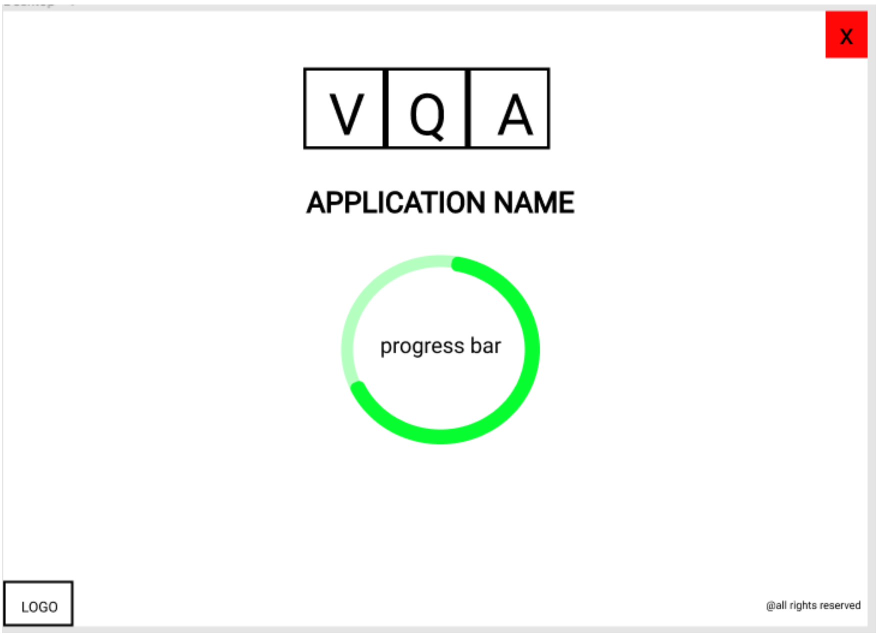


Figure 4.6: Splash Screen

- After Splash Screen the user will need to enter the input credentials to gain access to choose screen. The Choose screen will have two options to select from. User can select any of the choice depending upon the functionality he needs to perform.Figure 4.6 shows choose screen of the application.

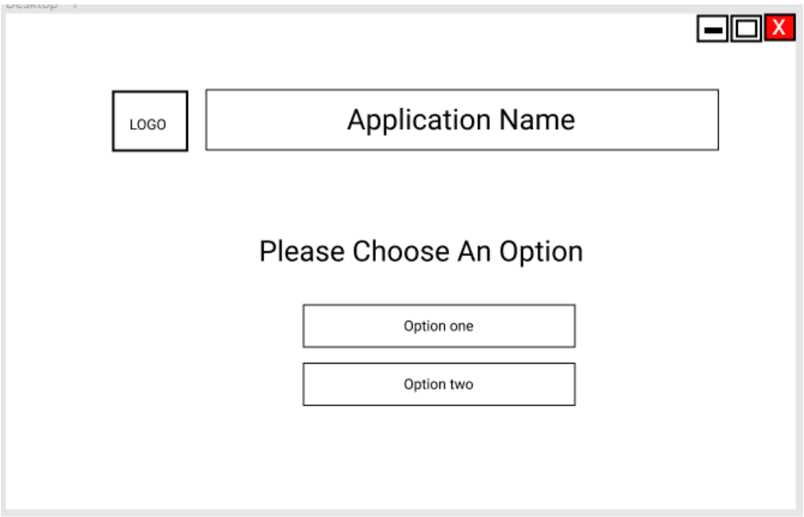


Figure 4.7: Choose Option Screen

- If user chooses option 1 the user will be directed to text detection Screen.Figure 4.7 shows the text detection screen.

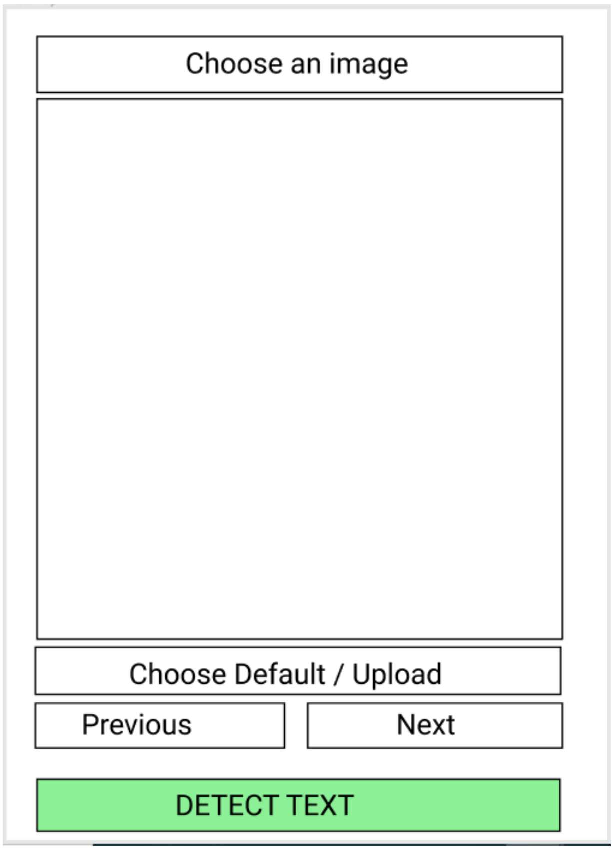


Figure 4.8: Text Detection Screen

- The text detection screen needs an image to be uploaded to detect text from it. The user will have to click on Choose Default/Upload button to select from default images or browse computer to select an image. Figure 4.8 shows default images screen with some images to select from.

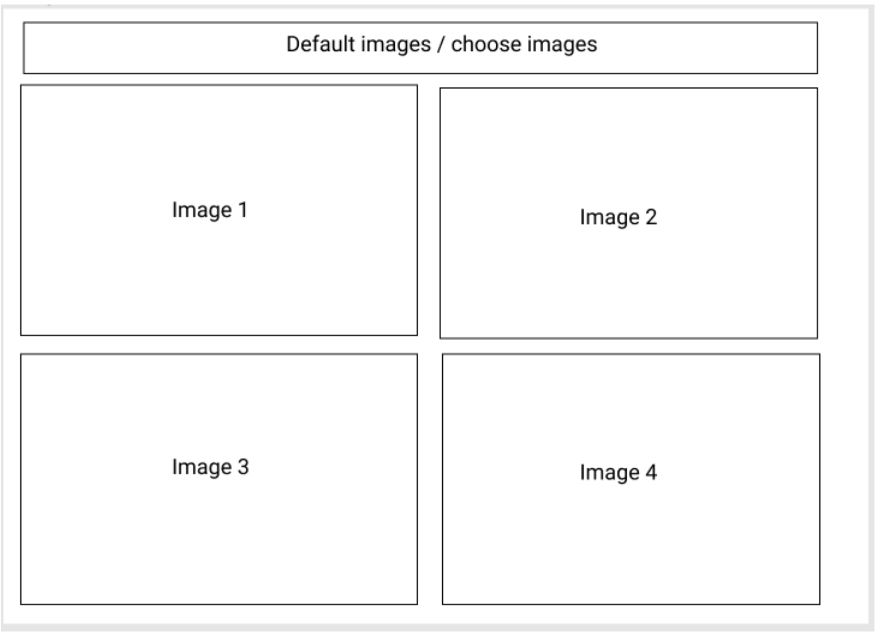


Figure 4.9: Choose /upload image

- After selecting an image the user will have to click on detect text button to detect the text form the image. The final screen will output text detected from the image as shown in the figure 4.9.

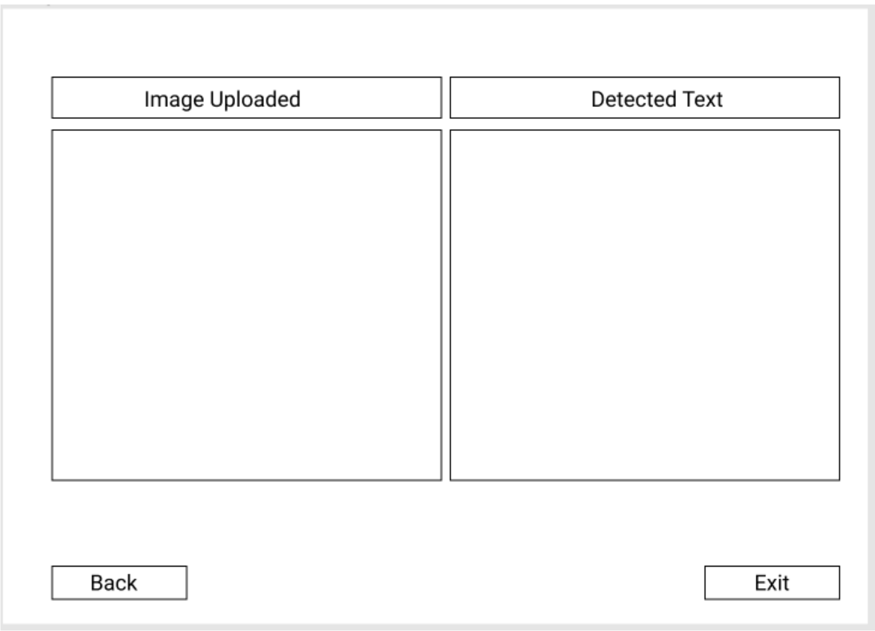


Figure 4.10: Final Text Detection Screen Prototype

- **Iteration 2** This Iteration will be run for Visual Question Answering option from choose Screen.

- The user will be directed to splash screen first and then login screen and after successful login the user will be directed to choose screen. The choose screen will display two option for the user.The choose screen is shown in figure 4.10.

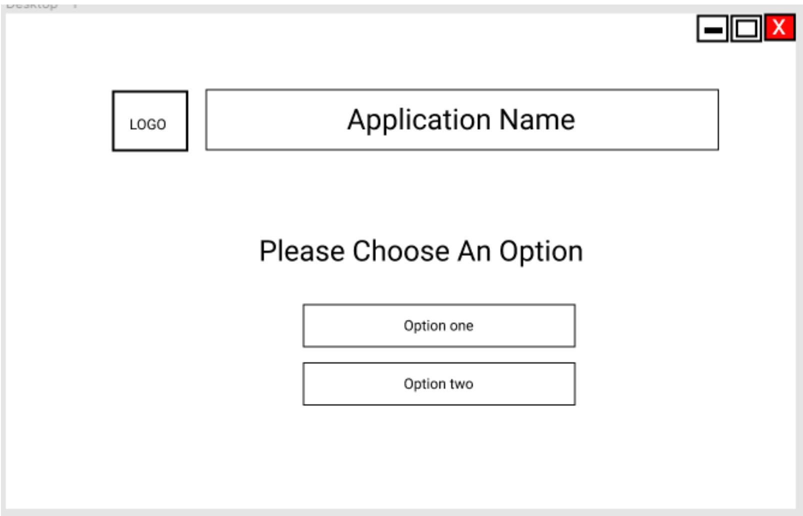


Figure 4.11: Choose Option Screen

- The user selects option 2 that is Visual Questioning Answering and user will be redirected to Visual Question Answering Screen.Figure 4.11 shows Visual Questioning Screen.

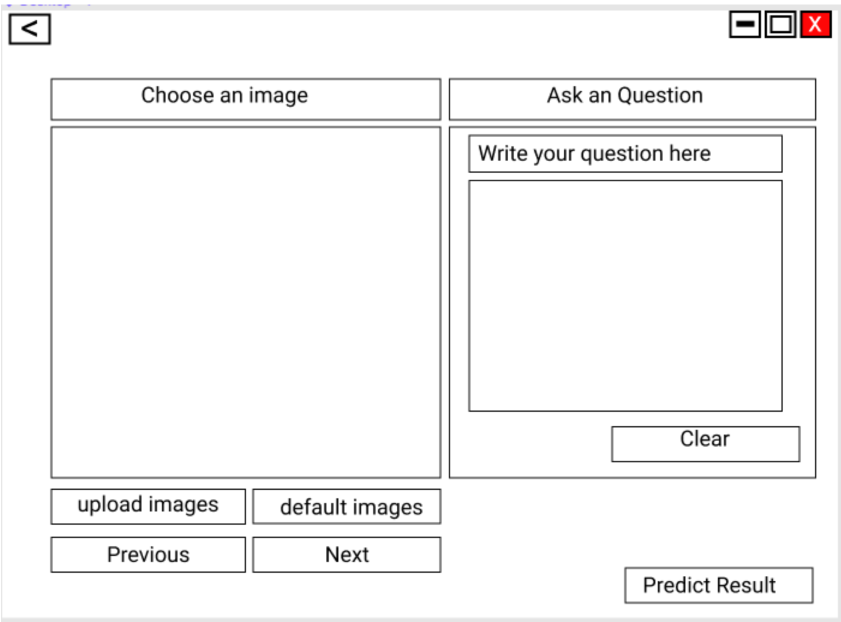


Figure 4.12: Visual Questioning Screen Option 2

- The user will select/upload an image from the image screen and will also write a query about image.Figure 4.12 shows the default images screen to choose from. User can also browse and upload any image by clicking on upload image button from Visual Question Answering Screen.

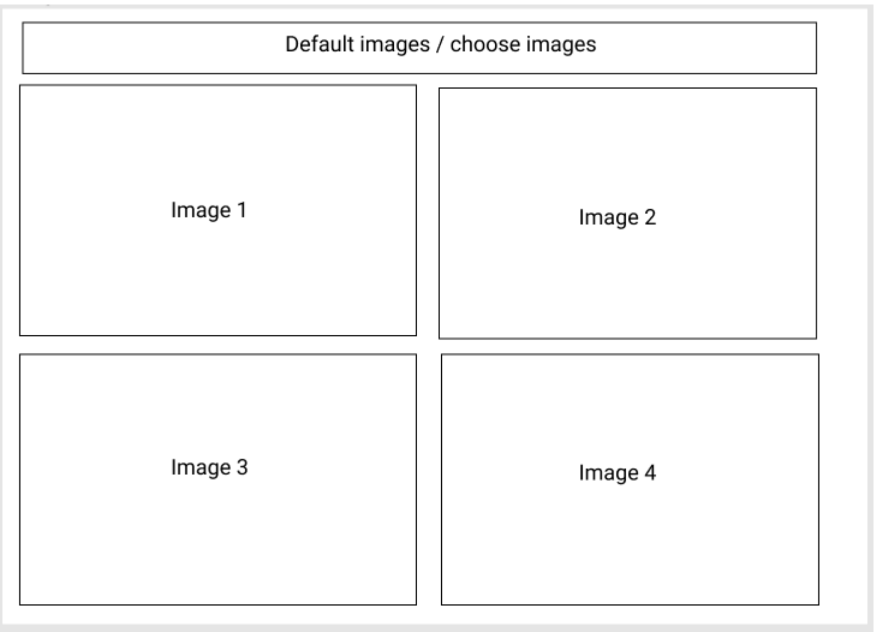


Figure 4.13: Upload images / Default Images screen

- The final Screen will Display the result of the user query about the image.Figure 4.13 shows the final output of the Visual Questioning Answering Screen.

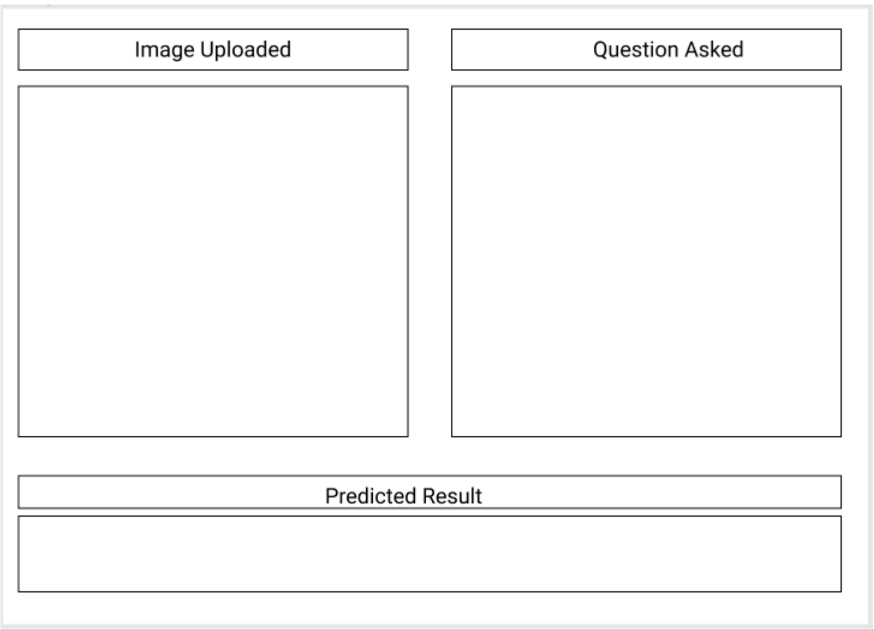


Figure 4.14: Final Visual Questioning Output Prototype

4.9.2 Actual GUI Interface

The actual gui output of the application is given below in the form of images. As our application has two options to select from. One text detection and second Visual Questioning Answering so the gui shows iterations for both. Iteration 1 for Text Detection Module and Iteration 2 for Visual Questioning.

- **Iteration 1 - Text Detection**

- The user will be directed to splash screen after launching the application.The splash screen is shown in Figure 4.14

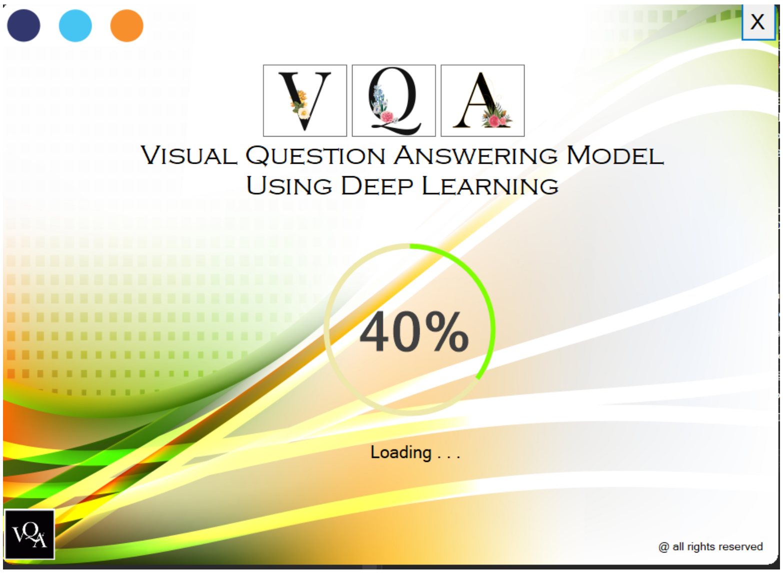


Figure 4.15: Desktop Splash Screen

- The user will then after successful login will be directed to choose screen and user will have option to select from text detection and visual question answering.The choose screen is shown in Figure 4.15.

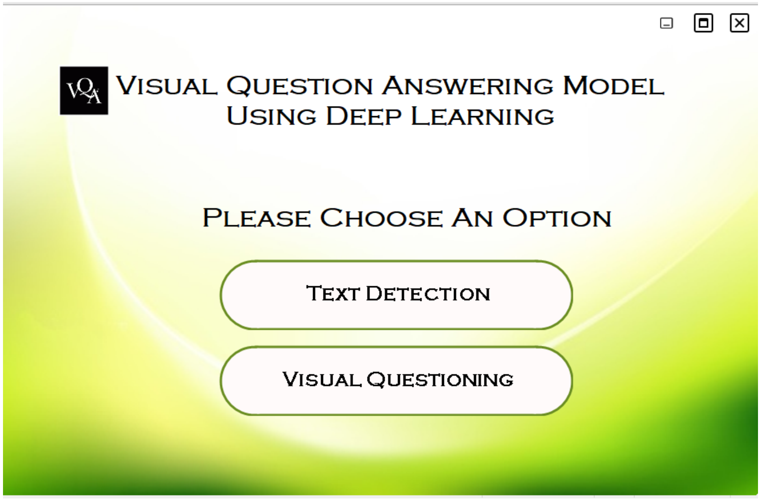


Figure 4.16: Choose Option Screen

- The user selects Text Detection Option as 1st iteration is for text detection.Figure 4.16 shows that user is selecting text detection option.

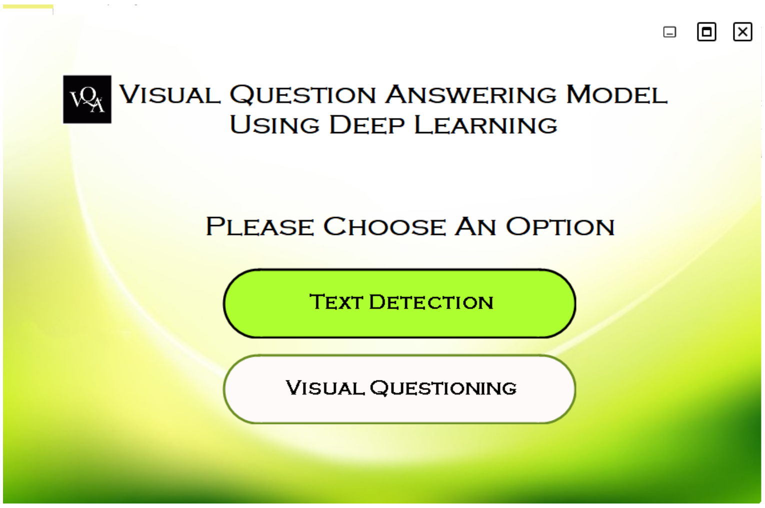


Figure 4.17: Option Text Detection Choosed

- The user will be directed to Text Detection Screen.The text detection screen is shown in Figure 4.17.

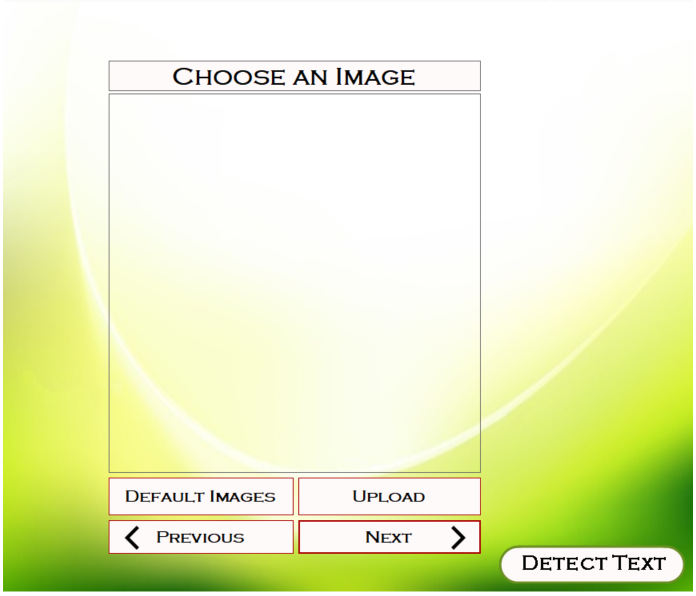


Figure 4.18: Text Detection Screen

- The user need to upload an image to detect text so he browse/ selects from default images.The default Image Screen will be displayed if user chooses a default image for text detection. The user can select any image from the default images.The default images screen is shown in Figure 4.18.

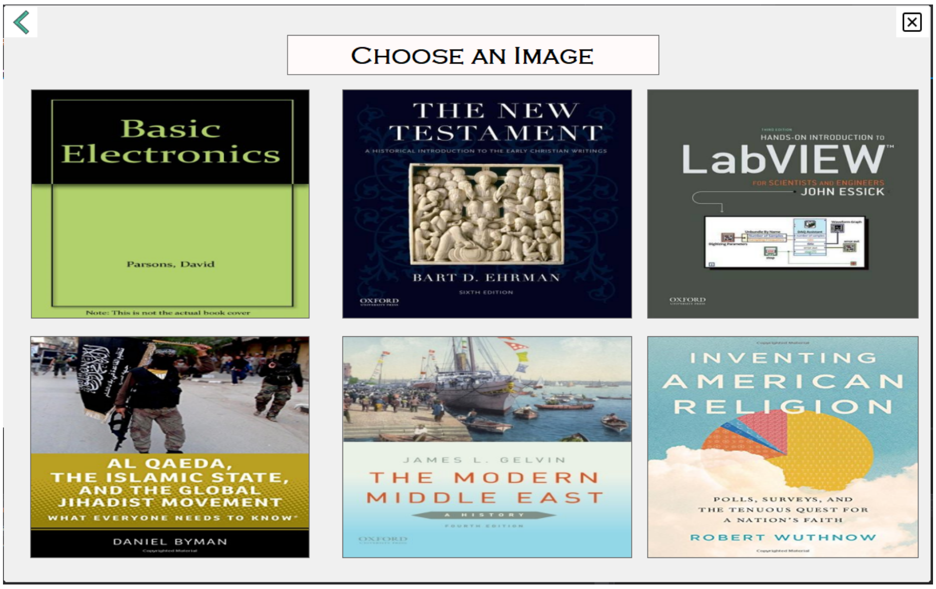


Figure 4.19: Default Images Screen

- The user selects an image and the selected images is displayed on the screen.The image selected from default image is shown in text detection screen as we can see in Figure 4.19.

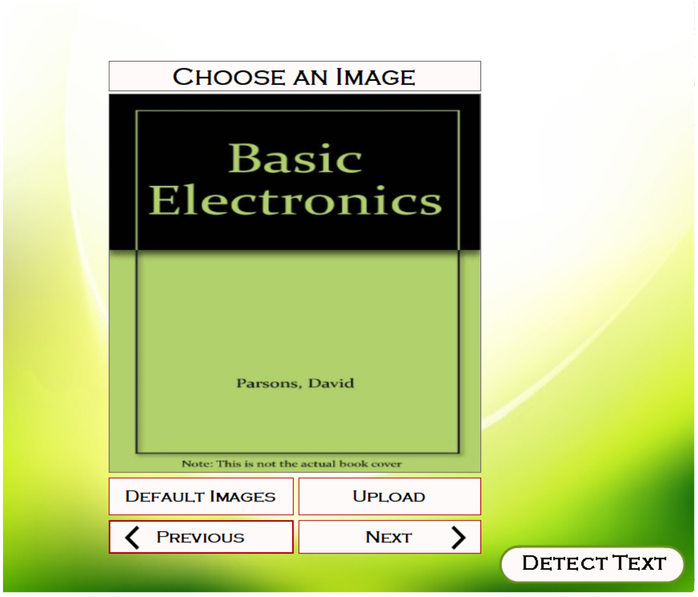


Figure 4.20: After Choosing Image

- User after selecting image click on the Detect Text Button.Figure 4.20 shows that user is clicking on detect text button.

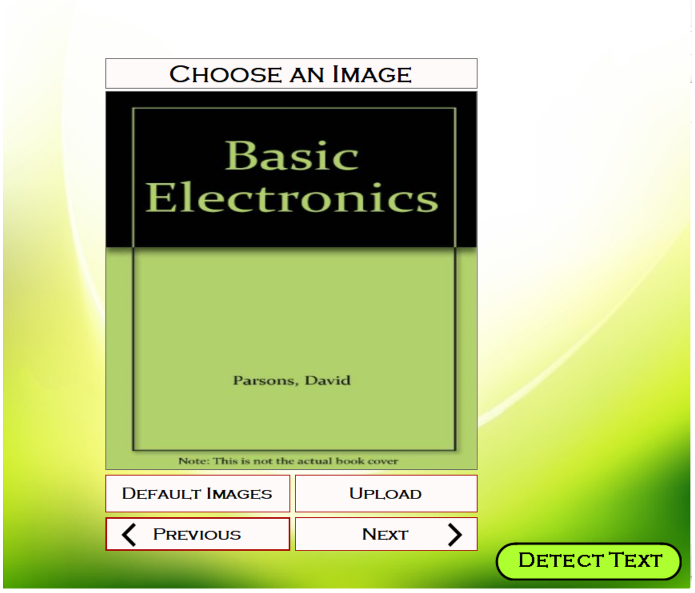
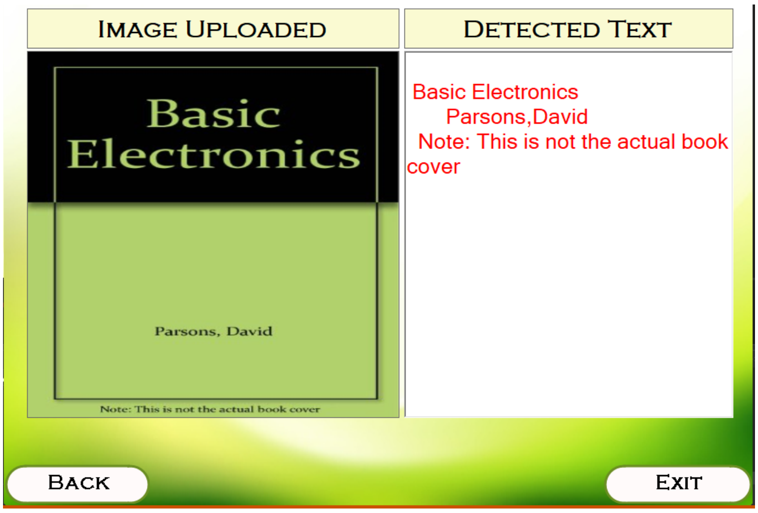


Figure 4.21: Detect Text Button Clicked

- The final Screen displays the text from the image user selected.Figure 4.21 shows the detected text from the selected image.

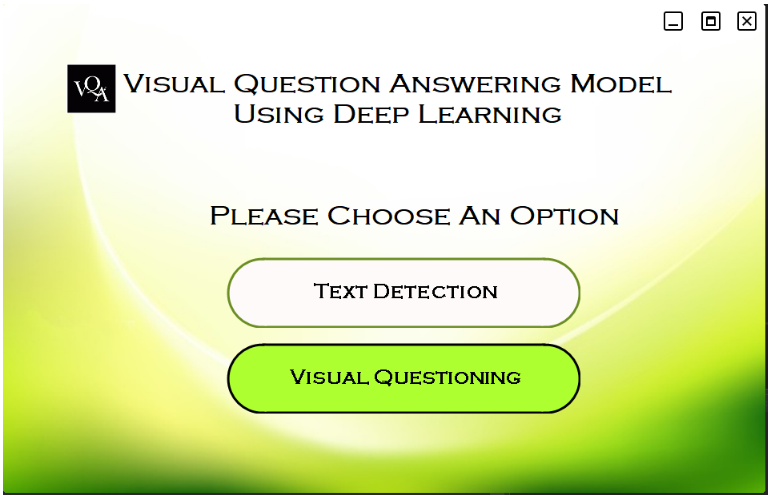


Basic Electronics
Parsons,David
Note: This is not the actual book cover

Figure 4.22: Text Detection Final Output

- **Iteration 2 - Visual Questioning Actual Output** This iteration is run for Visual Question Answering Option from choose screen.

- The user chooses Visual Questioning Option and will be directed to Visual Questioning Screen.Figure 4.22 shows that user is selecting Visual Questioning Option.



VISUAL QUESTION ANSWERING MODEL
USING DEEP LEARNING

PLEASE CHOOSE AN OPTION

Figure 4.23: Option Visual Questioning chosen

- The user needs to select an image to question. So he selects/browse an image. Figure 4.23 shows Visual Questioning Screen.

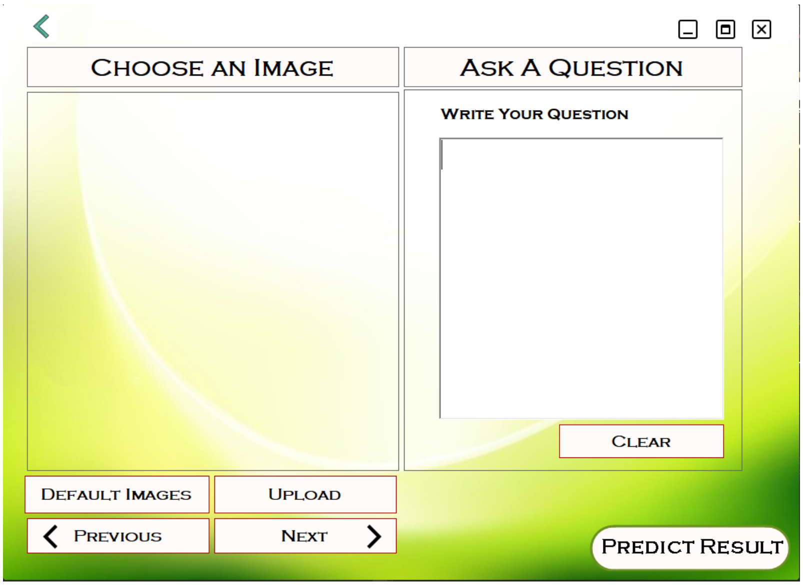


Figure 4.24: Visual Questioning Screen

- On Clicking Default images the user will be direct to choose from default images screen where user can choose from some of the default images. Figure 4.24 shows that user has clicked on default images button and a set of default image is displayed to the user.

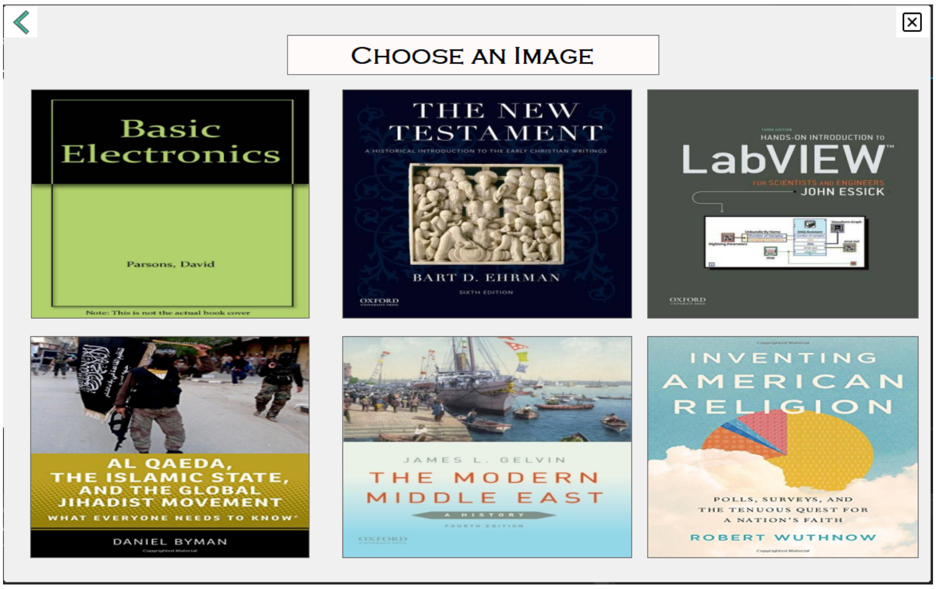
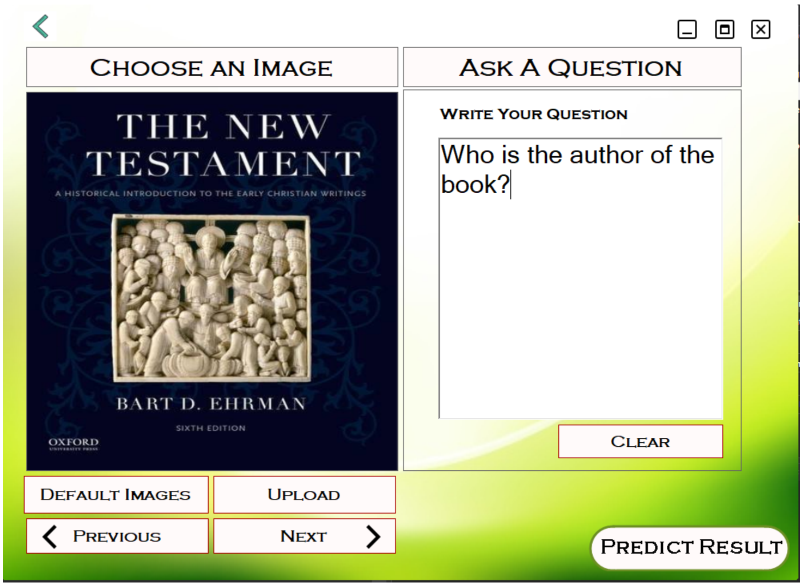


Figure 4.25: Default Images screen

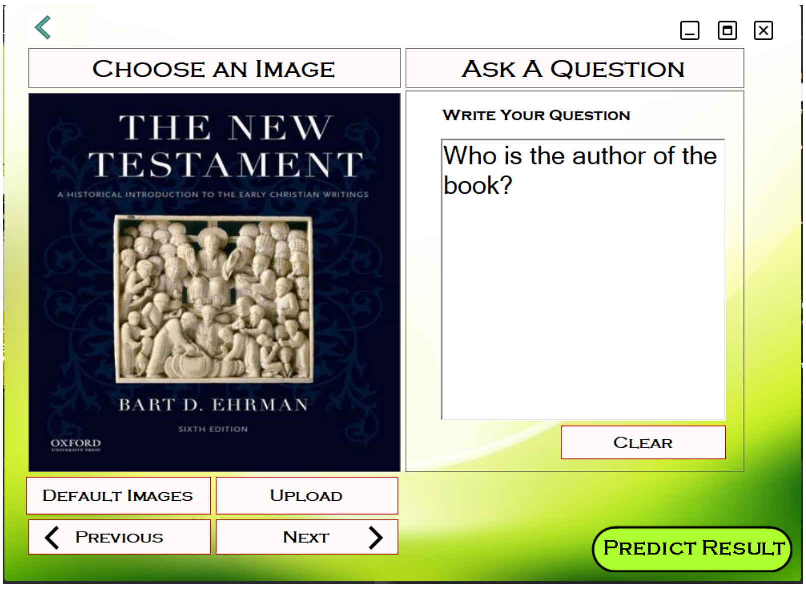
- After selecting an image the user needs to write a question in the question box for questioning image. Figure 4.25 shows that user has selected an image and has written a question.



Who is the author of the book?

Figure 4.26: After choosing Image and writing question

- After Selecting an image and writing the question the user click Predict Results button to find the result.Figure 4.26 shows that user is clicking on the predict result button to predict the model output.



Who is the author of the book?

Figure 4.27: VQA Question And Image Screen

- The final Screen outputs the user input and displays the result.Figure 4.27 shows the model predicted result corresponding to user selected image and user asked question.

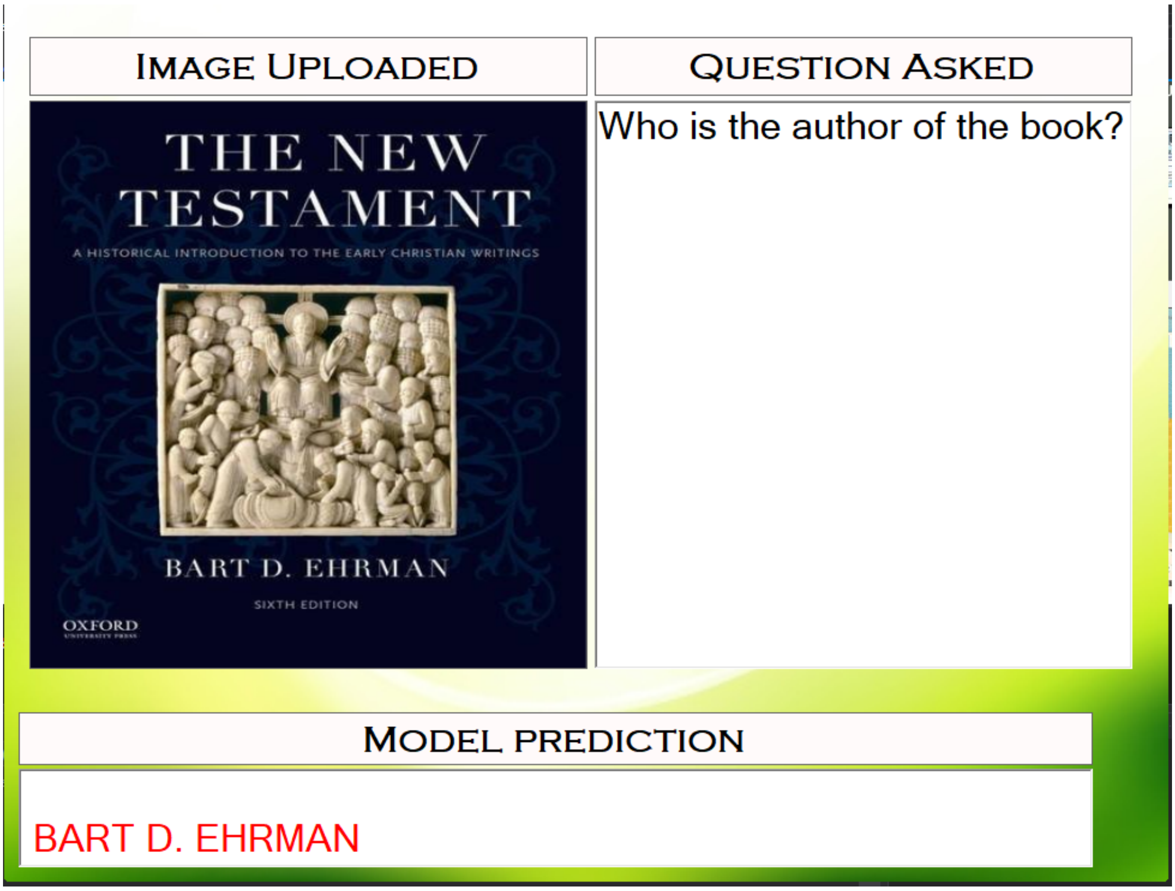


IMAGE UPLOADED

QUESTION ASKED

Who is the author of the book?

MODEL PREDICTION

BART D. EHRMAN

Figure 4.28: Final VQA output

4.10 Diagram

4.10.1 Activity Diagram

An activity Diagram is a graphical representation of workflow of stepwise activities. The flow of control in our application is shown through activity diagram give in Figure 4.28.

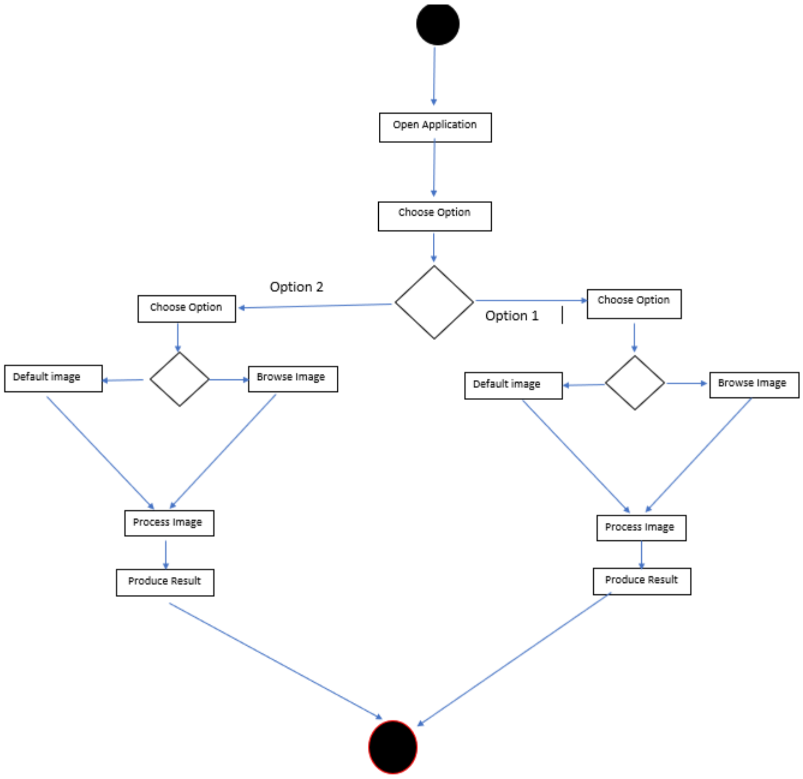


Figure 4.29: Activity Diagram

Chapter 5

System Implementation

In this chapter, languages, tools, and technologies are shown that are being used for implementation.

5.1 Dataset

In our project we use OCR-VQA–200K dataset.This dataset is open source dataset which contains 2 lac images and around 1 million answer questions.This dataset contains the images of books covers.The images of the dataset are split into training, validation part and test part. T

5.1.1 Data Set Images Examples

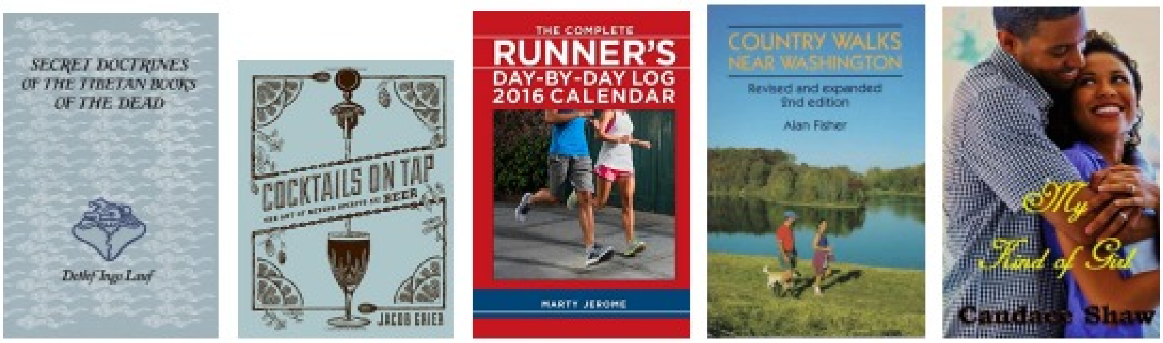


Figure 5.1: Dataset Images Example

5.2 Product Architecture

This product essentially focuses on developing such a system which predicts an accurate answer against a question about an image using a Visual question Answering (VQA) models.The systems front-end implementation registers a user on a system and user will upload a image and question related to the image , which information he wants to extract from image and in the return system will predict an accurate answer.We have applied different computer vision and deep learning techniques to develop this product.Text detection is done by in a image using a pre-trained model EAST text detector and Text understanding is done of a deep learning Model LSTM. Visual question answering is an desktop application that provides an interface for a user to detect text in the images as well as provides functionality to question the image. The VQA model will then predict the answer based on the text in the image and the question asked by the user.
In addition user can browse gallery to upload an image or can select from default set of images and can perform desired functionality depending upon the user need. Deep learning based detection model EAST detect a text from image and OCR will extract a text from detected regions of image.

5.3 System Internal Components

This section lists the internal components of the system with their functionality, their logic and their implementation, treating them as different sub-phases of the implementation phase of the project life cycle.

5.3.1 Text detection and recognition implementation

The focus of this project is to use the text that is present in images and develop an application which can do question answering from images by using that text. For this purpose first we detect a text which is present in images. we use pre-trained deep learning based model (EAST) for the detection of text in images. By using this model we make a bounding boxes where text is present .Once the text is detected and bounding boxes is created then we pass the these bounding box to tesseract that recognize a text which is present in image.

5.3.2 Texual features representation

After the text detected and recognized then we apply NER tagging to our text.First we represent a bounding box in to the vector and value of vector is bounding box coordinates.Vector is five dimensional which contains index number of box top left x, top left y, bottom left x, bottom right y. Then we make a dictionary and key of dictionary is vector that contains bounding box coordinates while value of that key is the text of bounding box.After that we make a method for NER tagging where we import a library Spacy library. Spacy is used in Natural language processing for text pre-processing.In NER tagging method spacy read our text , if person is found in text, then we append a vector by 1 if not found then we append it by zero.same for edition and year if they are in bounding boxes text ,we append it by 1 otherwise we append it by zero. After that, five dimensional vector becomes 8 dimensional vector. Then we pass this vector into NER tagging module where vector will become 8 dimensional.

5.3.3 combine features representation of Scene and Textual data

In our Project VQA after textual representation , we used vgg16, and LSTM to get the depth features of image and Question.vgg16 gave us the vector representation of 4096 , which we passed the fully connected layer with 1024 units.For question representation we used two layers of LSTM while each layer of LSTM contain 300 units and then we fed it with dense layer with 1024 units.

5.3.4 Application Working

Visual Question Answering Using Deep Learning is a desktop application which allows user to select or upload an image from the computer and then performs text detection on the image or provides user with functionality to question images based on the text in the image.

As the application front end is developed using Window forms in .NET Framework by using C sharp.To integrate the front end with the back-end of the code which is written in python we had two ways to implement it. One by calling an API and other by calling a python script from within Visual Studio by making processes to execute the python script. Our system can be implemented by using both options but we have chosen second option in which we have trained our model and then used that model script in our front end by making processes to executed the model script written in python using iron python library in .NET Frame Work. On button clicks the images is delivered to the model which performs the desired operation(i.e text detection/visual questioning) and produces an output which is then displayed to user using our window form application.

5.3.5 Features

The application comes with some features that helps user to interact with application with much ease and with less effort.

- Provides an option to choose image detection or question image

- For user's ease default images are provided so that user can know how to interact with the system and how it works
- The application can run without any Internet
- Friendly environment for the user to perform operations just on button clicks without any complexity

5.3.6 Product features

Our product Visual Question Answering is designed on Books cover images. This system can work on other type of images as well like medical images, Travelling images or any other type of images. Our product is trained only on books cover images and it will address only books cover images.Our product address the different types of questions like binary , year , edition ,title of book?

- what is the genre of the book?
- Is this a book related to travelling?
- what is the edition of this book?

5.3.7 Product limitations

Our Product Visual question answering using deep learning works only on Books cover dataset. This can only perform on books cover images. We can use this product on other images as well if we trained on other type of images.

5.3.8 Functionality

After opening the application, a flash screen appears which turns then to a choose screen which displays two options in the form of buttons for the user to select from. User can select any option and then is redirected to the main page of that option where for text detection option he can upload/browse an image from the computer or can select form default images. After the image is selected the user is then requested to press the detect text button which will eventually detect the text in the image and will display it to the user. In the case of visual question answering option the user will be redirected to the vqa main page where he will be requested to select/browse an image and then will be asked to write any queries related to the image then the user will be requested to click predict result button which will eventually produce the result and display it to the user on the final screen.

5.4 Tools and Techniques

- Visual Studio 2019
- Windows
- Figma
- Tensor Flow

5.4.1 Desktop Libraries Used for Implementation

- Window Forms C-Sharp and Controls in it
- Bunifu Framework Extension
- Bunifu Drag Control
- Bunifu Ellipse
- Round Button Class
- Circular Progress Bar Extension

5.5 Languages

- Python Back End -Text Detection, Natural Language Procession and integrating both in python
- C-Sharp Front End - Application Design and interface developed in Windows Visual Studio by using C- Sharp as a front end language.

5.6 Problem Faced

- **Front and Back End Integration**
  There comes many difficulties while integrating C- sharp code with python in Visual Studio. Several dependencies issues arose during the integration.
- **Application Design**
  There is not built in functionality in the windows forms to drag control the form. So either the developer has to write it for himself/herself or he can look for an online solution in the internet.A person having less C-Sharp knowledge will have difficulties in writing the code himself. But there are certain frameworks in visual studio which can be installed in visual studio to perform this task. One good example

is bunifu framework which comes with its own toolbox which provides many good functionalites like drag control , or shaping a form and so on. Hence, one needs to install the bunifu extension on one's project and can enjoy the benefits of it.

Chapter 6

System Testing and Evaluation

6.1 Dataset

In our project visual question answering from images using deep learning we used open source dataset OCR-VQA-200k.This dataset contains 207k books cover images while more than 1 million open ended Questions Answers pairs.The dataset is divided into training, testing and validation part. 80 % of the data for the training part, 10 % for the validation part and 10 % of the data is used for the testing of the model.

6.2 System performance

There are many approaches to implement a VQA model. Different approaches have different results.In our project we have follow that approach which is able to achieve high accuracy as compare to other models. Our model uses VGG16, Ner-tagging ,w2v, joint features representation and bilstm for the prediction of Answer. There are results on training and validation part of dataset by our model is shown in table.

Table 6.1: Results on Train and Validation data

Training Data	**Validation Data**
54.2 % Accuracy , 0.2% Loss	54.2 % Accuracy , 0.7% Loss

6.3 Results

There are different types of questions in our dataset.Accuracy of our model is different on different types of questions.The accuracy of our model on different types of question is mentionded on table 6.2.

Table 6.2: Model Performance on different Question Types

Questions- Type	Model Accuracy
Binary	68%
Edition Number	40%
Author Name	15%
Book Title	65%
Year	67%

In table name the results are mentioned.There are some cases in which our proposed solution performs very well like in binary types of questions,Book title and the year in which it has published. But in some cases our proposed model could not perform very well like author name, edition number etc. The lower performance is because of the wide variation in scale, layout and fonts style of text. And variations in question asked (e.g., paraphrasing) and questions related to genre of book also reduce the performance of model.

6.4 Graphic User Interface testing

User interface testing has many different aspects that need to be considered. These, along with visual design, include functionality and user experience analysis. In the case of our system, the front-end has been tested separately for both GUI analysis and usability and is ensured that every component from button to a text box is properly working.

6.5 Usability testing

Usability testing for this system has been aligned with usability heuristics principles of Jakob Nielson. To ensure consistent look and feel, one color scheme has been set for the entire application major components. By doing this, when the user looks at components with the same color scheme, he/she can identify the relation of similarity and importance of those components. Contemporary colors have been used to divert attention to actions happening on the application. Functionality status of all buttons and input fields is visible without delay. Graphics used in the interface are clear visualizes in order to form an understanding for the user. The application has been developed with minimum user interface components that carry the functionality. This is to avoid clutter and distraction of navigation and to ensure guided functionality and aesthetic appeal.

6.6 System Limitations

The system, by using object detector and WordNet synonym generation is automated. With this, it is limited to the results generated by python models. All objects in an image may not be identified and all label synonyms may not be meaningful.

6.7 Unit Testing

It was important to ensure proper working of the system to make sure that the project fulfills its objects and meets its requirements. Each component of the system mentioned has been tested during and after development.

6.7.1 Unit Test 1

Unit Testing 1: Add new User
Testing Objective: To ensure that a new user can register in the application
Test Case Id: TC-001
Test Case Description: Test the signup functionality
Test Scenario: Verify that on entering correct Name, email, password, confirm password and the user can sign up successfully.

Table 6.3: Sign Up Successful

No.	Test Case	Test Data	Expected Result	Actual Result	Pass / Fail /NA
1.	Verify that user enters the required input field for signup process	Name: User's Full Name Email: Any Valid Email Password: Must be eight characters long with one special letter Confirm password: Must match with the password	User is success_ fully registered and can now login to the system	As expected	Pass
2.	Verify that on inserting the correct credentials user can sign up	The user has inserted correct user name and password	sign up successful	As expected	Pass

Table 6.4: Sign Up Unsuccessful(wrong email)

No.	Test Case	Test Data	Expected Result	Actual Result	Pass / Fail /NA
1.	Verify that user does not provide a valid email	Email: Any invalid Email provided e.g. (abcgmail.com)	Inform the user that the email provided is not valid and must provide a valid email.	As expected	Pass

Table 6.5: Sign Up Unsuccessfull(wrong password)

No.	Test Case	Test Data	Expected Result	Actual Result	Pass / Fail /NA
1.	Verify that user does not provide a correct password	Email: Any invalid password provided by not following the length and not including a special character provided e.g. (abcd)	Inform the user that the password provided is not valid as it not matches the required length or does not contain a special character or confirm password and password fields do not match.	As expected	Pass

6.7.2 Unit Test 2

Unit Testing 2: Login to the Application
Testing Objective: To ensure that a user can login to the application
Test Case Id: TC-002
Test Case Description: Test the login functionality
Test Scenario: Verify that on entering a registered email and correct password, the user can login

Table 6.6: Login Successful

No.	Test Case	Test Data	Expected Result	Actual Result	Pass / Fail /NA
1.	Verify that user has provided correct input credentials for login process.	Email: Any valid email that is already registered (abc@gmail.com) Password: valid password (abc@defg)	Login Successful user is redirected to main page	As expected	Pass

Table 6.7: login Unsuccessful(wrong email/password)

No.	Test Case	Test Data	Expected Result	Actual Result	Pass / Fail /NA
1.	Verify that user has provided incorrect input credentials for login process.	Email: Any invalid email (abgmail.com) Password: invalid password (abdefg)	Login Unsuccessful. User is redirected to login page and user is informed that the email or password provided are incorrect and user will be asked a security question for successful login.	As expected	Pass

6.7.3 Unit Test 3

Unit Testing 3: Upload an Image and post a Query

Testing Objective: To ensure that a user uploads an Image and Questions the Image by uploading/selecting an image and by writing a Question

Test Case Id: TC-003

Test Case Description: Test the Main Functionality

Test Scenario: Verify that the user uploads/selects an image and Questions Image

Table 6.8: Successful Upload Image And Post Query

No.	Test Case	Test Data	Expected Result	Actual Result	Pass / Fail /NA
1.	Verify that the user can upload/select an image and can question it before clicking on predict result button.	Image: any valid image Question: any valid question	The users selects /uploads an image and user writes queries about the image in the question box.	As expected	Pass

Table 6.9: Unsuccessful image uploaded or Unsuccessful Post Query

No.	Test Case	Test Data	Expected Result	Actual Result	Pass / Fail /NA
1.	Verify that the user does not uploads/selects an image	Image: no image uploaded	The user will be requested to upload an image before clicking the predict result button.	As expected	Pass
2.	Verify that the user does not questions an image.	Question: no question asked	The user will be requested to write a question before clicking the predict result button.	As expected	Pass
3.	Verify that user neither uploads an image nor writes a question.	Image: no image uploaded Question: no question asked	The user will be requested to upload and image and write a question before clicking the predict result button.	As expected	Pass

Chapter 7

Conclusions

Conclusion

Our project served in Desktop Application domain in which we created different modules. By using this application , users can give his image to the system and system will give a different information to the user which the user want . Our goal was to create an application that could serve it purpose and to overcome those problems that exists. Our system can helps a lot for blinded people. Our system can solve different problems for blinded people as well as for society in different domains. We can make other systems on similar pattern in different data. The other systems can consider any dataset like docvqa, medical imaging or traffic related data to re-design a VQA model accordingly.

7.0.1 Future Work

The VQA is new and emerging problem in the filed of computer vision. There is a lot of potential such type of systems.Our project has a lot of research potential in different areas of computer vision, deep learning and Natural language processing.By using this dataset and future work of this project could be to improve text areas of image, recognizing a text in a better way. Different proposed models and architectures can be used to increase accuracy on this dataset.There can be also worked on combing scene and text of the image and it will be more challenging for VQA community.

References

[1] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 947–952, 2019. Cited on p. 2.

[2] Wei Han, Hantao Huang, and Tao Han. Finding the evidence: Localization-aware answer prediction for text visual question answering. *CoRR*, abs/2010.02582, 2020. Cited on p. 8.

[3] Riddhi Nisar, Devangi Bhuva, and Pramila Chawan. Visual question answering using combination of lstm and cnn: A survey. pages 2395–0056, 10 2019. Cited on p. 8.

[4] Aiwen Jiang, Fang Wang, Fatih Porikli, and Yi Li. Compositional memory for visual question answering. 11 2015. Cited on p. 8.

[5] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. Cited on p. 8.

[6] Wei Han, Hantao Huang, and Tao Han. Finding the evidence: Localization-aware answer prediction for text visual question answering. *CoRR*, abs/2010.02582, 2020. Cited on p. 8.

[7] Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu. Relation-aware graph attention network for visual question answering. *CoRR*, abs/1903.12314, 2019. Cited on p. 9.

[8] Asma Ben Abacha, Mourad Sarrouti, Dina Demner-Fushman, Sadid A. Hasan, and Henning Müller. Overview of the vqa-med task at imageclef 2021: Visual question answering and generation in the medical domain. In *CLEF 2021 Working Notes*, CEUR Workshop Proceedings, Bucharest, Romania, September 21-24 2021. CEUR-WS.org. Cited on p. 9.

[9] Minesh Mathew, Dimosthenis Karatzas, R. Manmatha, and C. V. Jawahar. Docvqa: A dataset for VQA on document images. *CoRR*, abs/2007.00398, 2020. Cited on p. 9.

[10] Zhengyuan Yang, Yijuan Lu, Jianfeng Wang, Xi Yin, Dinei A. F. Florêncio, Lijuan Wang, Cha Zhang, Lei Zhang, and Jiebo Luo. TAP: text-aware pre-training for text-vqa and text-caption. *CoRR*, abs/2012.04638, 2020. Cited on p. 10.