PROJECT REPORT
ON

Visual Question Answering for the Visually Impaired

SUBMITTED TO THE SAVITRIBAI PHULE PUNE UNIVERSITY, PUNE
IN THE PARTIAL FULFILLMENT FOR THE AWARD OF THE DEGREE

OF

BACHELOR OF ENGINEERING IN INFORMATION TECHNOLOGY

BY

Shubham Mahajan	**71829233C**
Omkar Deshpande	**71828664C**
Devesh Chandak	**71828613J**
Sanya Varghese	**71829168K**

UNDER THE GUIDANCE OF

Dr. S.C. Dharmadhikari

DEPARTMENT OF INFORMATION TECHNOLOGY
PUNE INSTITUTE OF COMPUTER TECHNOLOGY, PUNE.

2020-2021

CERTIFICATE

This is to certify that the preliminary project report entitled

VISUAL QUESTION ANSWERING FOR THE VISUALLY IMPAIRED

Submitted by

Shubham Mahajan	**71829233C**
Omkar Deshpande	**71828664C**
Devesh Chandak	**71828613J**
Sanya Varghese	**71829168K**

is a bonafide work carried out by them under the supervision of Dr. S. C. Dharmadhikari is approved for the partial fulfillment of the requirement of Savitribai Phule Pune University for the award of the Degree of Bachelor of Engineering (Information Technology).

This project report has not been earlier submitted to any other Institute or University for the award of any degree or diploma.

Dr. S. C. Dharmadhikari
Internal Guide
Department of Information Technology

Dr. Anant M. Bagade
Head of Department
Department of Information Technology

Prof. J. K. Kamble
External Examiner

Dr. R. Sreemathy
Principal
PICT, Pune

Date :
Place: Pune

ACKNOWLEDGEMENT

It gives us great pleasure and satisfaction in presenting this report on “Visual Question Answering for the Visually Challenged”. The completion of this report required a lot of guidance and assistance from many people and we consider ourselves to be extremely privileged to have received this support throughout the semester.

We would like to thank Mrs. Radhika V Kulkarni and our HoD Dr. A.M. Bagade, for giving us an opportunity to do the BE Project work in PICT, IT Dept, and giving us all support and guidance which helped us to complete the report duly.

We owe our deep gratitude to our project guide Dr. S.C. Dharmadhikari who took a keen interest in our project work and guided us all along and provided us with all the necessary information for developing a good report.

A special thanks to our external reviewers Mr. J. K. Kamble, Ms. Sarika Patil, and Mrs. R. Chhajed for their recommendations and moreover for their timely support and guidance till the completion of our work. We are thankful and fortunate enough to get constant encouragement, support, and guidance from all the staff of the IT Department for helping us successfully complete our project report.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Sr. No.	Abbreviation	Full Form
1.	VQA	Visual Question Answering
2.	MCA	Modular Co-Attention
3.	SA	Self Attention

CONTENTS

ABSTRACT

The lack of access to basic visual information like text labels, icons, and colors can cause exasperation and decrease independence for the visually impaired. We thus propose a visual question answering system [1] to mitigate concerns about undesired consequences from today's status quo for visually impaired people, that is, relying on able-bodied humans to answer visual questions. We wish to assist visually impaired to overcome their daily visual challenges and break down social accessibility barriers.

In this project, we propose the idea of an open ended visual question answering system to assist visually impaired. The task requires an in-depth understanding of visual and language features, finally, we need to assess the relationship between the two modalities and use co-attention. The proposed model is goal-oriented, stressing on the images and questions generated by visually impaired people.

Keywords: Visual Question Answer(VQA),Open ended questions, SelfAttention(SA), Guided-Attention(GA), Long short-term memory(LSTM), Modular Co-Attention(MCA), Optical Character Reader, Bottom-Up and Top-Down

CHAPTER 1: INTRODUCTION

1.1. MOTIVATION

As able-bodied humans, it is easy for us to see an image and answer any question about it using our 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 an image.We wish to assist visually impaired people to overcome their daily visual challenges and break down social accessibility barriers. The purpose of the project is to bring sight to visually impaired and low-vision people.

1.2. OVERVIEW

We want to build an AI system, which takes as input an image and a free-form, open-ended, or natural language question about the image and produces a natural language answer as the output. The system will answer a question similar to humans in the following aspects:

1. It will learn the visual and textual knowledge from the inputs (image and question respectively)
2. Combine the two data streams
3. Use this advanced knowledge to generate the answers to open ended questions

Vision-and-language reasoning requires the understanding of visual contents, language semantics, and cross-modal alignments and relationships. Considerable amount of work has been done in both the fields, vision and language. Despite these distinguished single-modality works, studies for the modality-pair of vision and language, principally, pretraining and fine-tuning are still under developed.

1.3. PROJECT UNDERTAKEN

The proposed model focuses on learning vision-and-language interactions, especially for representations of a single image and its illustrative sentence.

The aim of the model is to address the following two tasks:

1. Predict the answer to a visual question
2. Predict the answerability of visual questions.

Our aim is to learn, design and implement the VQA model which we have proposed, which is a relatively new concept and is under developed.

The goal is to build an app that is :-

1. Easy to handle for visually impaired people
2. The app should be voice activated.
3. Users can use gesture to capture images
4. Users can use a voice assistant to register question
5. The app will respond with an appropriate answer to the question.
6. Instead of just describing the image, our app would answer the questions based on the image.
7. Users can double click the screen anywhere to click a new image.

1.4. ORGANISATION OF REPORT

The report is divided into 4 parts that consist of different aspects of our project.

1. Background and Literature Survey

1.1. Existing Methodology

1.2. Proposed Methodology

2. Requirement Specification and Analysis

2.1. Problem Definition

2.2. Scope

2.3. Objective

2.4. Project Requirements

2.5. Project Plan

3. System Design

3.1. Architecture

3.2. Structural Diagrams

3.3. Behavioral Diagrams

3.4 Algorithms and Methodologies

4. Implementation

4.1. Stages of Implementation

4.2. Implementation Software/Techniques

5. Result and Evaluation

5.1 Experiments

6. Conclusion

6.1 Limitations

6.2 Scope

CHAPTER 2: BACKGROUND AND LITERATURE REVIEW

2.1. EXISTING METHODOLOGIES

1. ENSEMBLING RICH IMAGE FEATURES

Recent approaches have introduced the visual attention mechanism into VQA by adaptively learning the attended image features for a given question, and then performing multimodal feature fusion to obtain the accurate prediction. Beyond understanding the visual contents of the image, VQA also requires to fully understand the semantics of the natural language question. Therefore, it is necessary to learn the textual attention for the question and the visual attention for the image simultaneously.

The focus of this approach is the Modular Co-Attention(MCA) Layer. The MCA layer is a modular composition of the two basic attention units, i.e., the self-attention (SA) unit and the guided-attention (GA) unit The model uses GloVe + LSTM to extract question features and uses Faster RCNN to extract image features. Self Attention and Guided Attention are applied to build encoder and decoder. It uses MLP to output these features, which are fused together. The loss function is called Multi label BCE loss.

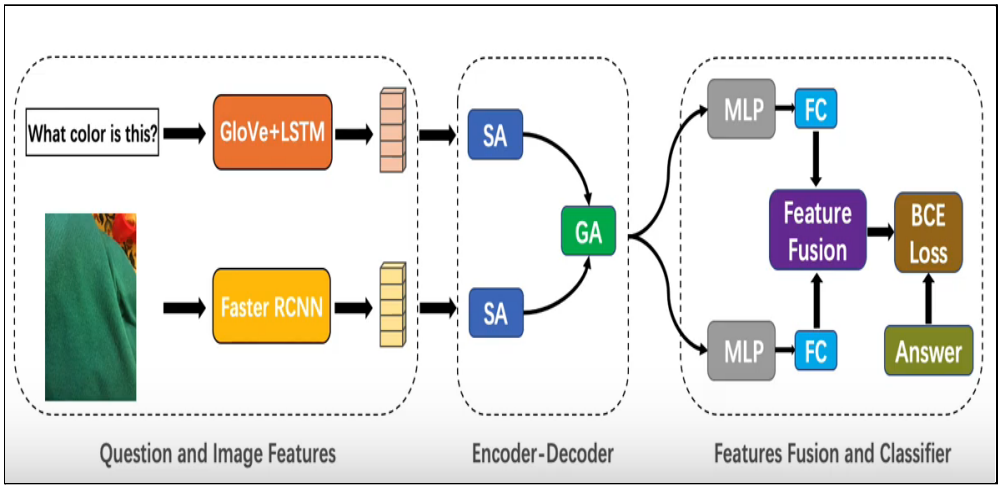


Fig. 1. SUDOKU MODEL [14]

To get precise convolutional features the architecture resizes an image to get FC features from ResNet 152. Bottom up attention to generate image features using ResNet 101. For question representation the paper uses the following steps. In the beginning a question is trimmed to a maximum of 14 words. The vectors from pretrained GloVE which generate word embeddings are fed into an LSTM.

Lastly, by ensembling 14 models the paper achieves an accuracy of 56.20.
This model was the runner up of the VizWiz-VQA 2020 challenge

2. BOTTOM-UP AND TOP-DOWN ATTENTION

To achieve the best results the architecture uses the extensively used Top-Down visual attention mechanism. The architecture is well-accepted for image captioning, image description and answering open-ended visual questions(VQA) combined with Bottom-Up attention to capture more salient features. The Bottom-Up mechanism (object detection) puts forward regions of image, each with an associated feature vector, the top-down method determines feature weightings for the feature vector.

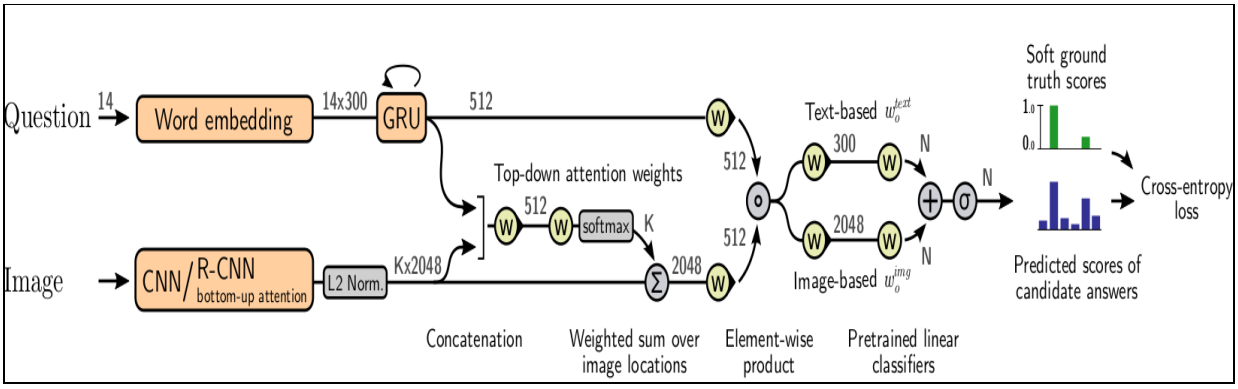


Fig. 2. BOTTOM-UP TOP-DOWN ATTENTION[3]

The highlights of this architecture are:
1. Image Features: The input image is passed through a Convolutional Neural Network (CNN).

Bottom-up attention results in higher performance. The method is based on a ResNet CNN within a Faster R-CNN framework. It is trained to focus on specific elements in the given image.

2. Image Attention: The top-down attention, concatenates image features with the question embedding. Establishing a relation between the input text and image later in the architecture.

3. Fusion: The representations of the question and of the image are passed through non-linear layers and then combined with a simple Hadamard product The model achieves accuracy of 72 on the VQA dataset. The model was the winner of the VQA 2018 challenge.

3. SHOW, ASK, ATTEND AND ANSWER

"Show, Ask, Attend, and Answer" was one of the earliest attempts at Visual Question Answering by Google Research in 2017 [2]. The paper presents a new baseline for answering open ended questions from images. The model is architecturally simple and takes in few parameters, it was the SOTA at the time with an accuracy of 64% on the first VQA dataset.

1. The model uses a pre-trained CNN model to compute features from the image, Normalization is performed in depth to provide better learning dynamics.

2. The questions are tokenized, and the embeddings are fed into an LSTM

3. The model uses stacked attention distributed over the image features. Each image feature is a weighted average of image features. The wights are normalized separately

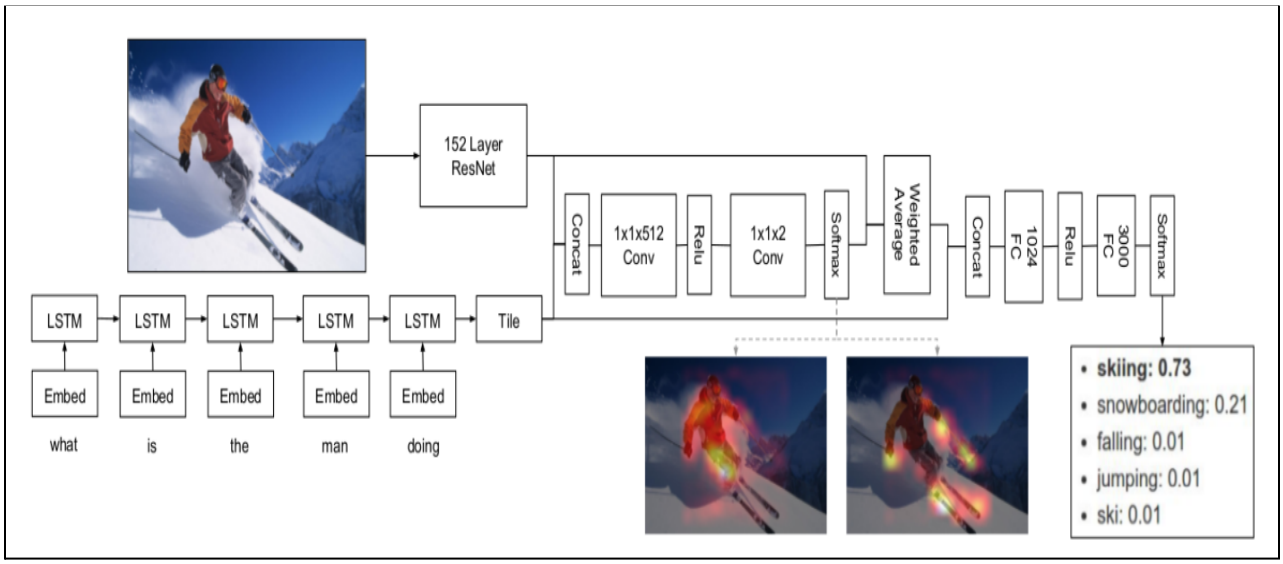


Fig 3: STACKED ATTENTION MODEL [2]

4. DIFFERENTIAL NETWORKS

The contrast between the various forward propagation steps is utilised to bring down the noise and grasp the reaction between the features [16]. The task of Visual Question Answering depends heavily on combining optimised language and visual features. The Image features are extracted using Faster-RCNN .

The differential modules are used to refine the features in both text and images. GRU is used for question feature extraction.

Finally, it is combined with an attention module to classify the answers.

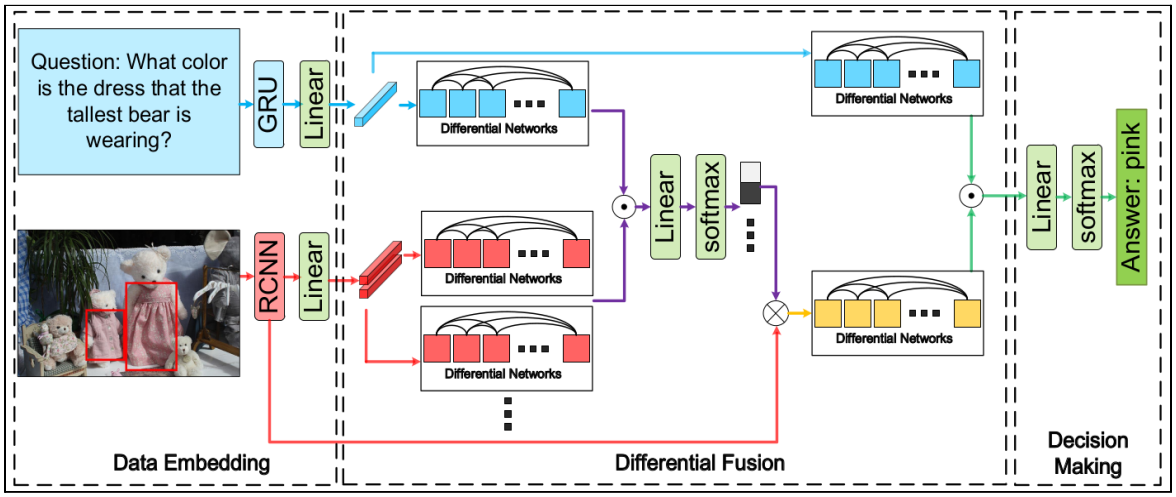


Fig 4: DIFFERENTIAL NETWORK [16]

The method ignores if the question and the image features are in the same space. It also ignores reduction of observation noise from these two features. The paper argues that the differential network is beneficial to reduce the observational noise [16].

5. Pythia

Another architecture that we would like to include in our literature review is Pythia [5]. The architecture is similar to Bottom-Up Top-Down. Fine-tuning, data augmentation and learning rate schedule significantly improve the performance from 65.67% to 70.24% on the VQA dataset.

This modular re-implementation was the winning entry of VQA challenge 2018.

The model include these key changes to improve training speed and accuracy:

1. Feature concatenation was replaced by element wise multiplication for the text and image features
2. Normalised weights followed by RELU
3. The model also included optimised learning rate and grid features.

2.2 PROPOSED METHODOLOGY

We propose a model based on Transformer encoders and novel cross-modality encoders proposed in the paper. The highlight of this architecture is its diverse pre-training tasks and exhaustive dataset. The model firmly adheres to three ideas: Bi-Directional Attention, Transformer, and BUTD.

The paper refers to the idea of LXMERT [14] and Oscar [10] inspired by the addition of semantic words to the language modality. The semantic tags include object and attribute tags.

Cross-modality models are built with a self-attention and cross-attention layer.

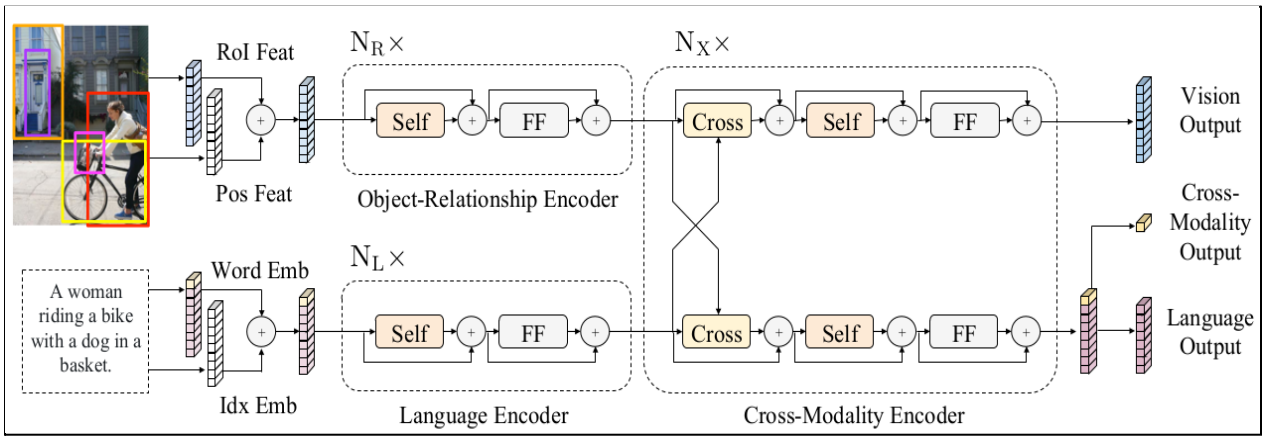


Fig. 5. PROPOSED MODEL [14]

1. Encoders: The architecture includes 3 encoders which work mostly on the basis of two kinds of attention layers: Self-Attention layers and Cross-Attention layers. Attention layers aim to extract

features from the text-question. It tries to revive context from the question/query [14].

2. Single Modality Encoders: The language-encoder and image-encoder focus on single modalities. We first separately apply these modalities of the text and the image feature vector.

3. Cross Modality Encoders: Each layer in the Cross-Modality encoder consists of two self-attention sub-layers, one bi-directional cross-attention sub-layer, and two feed-forward sub-layers.

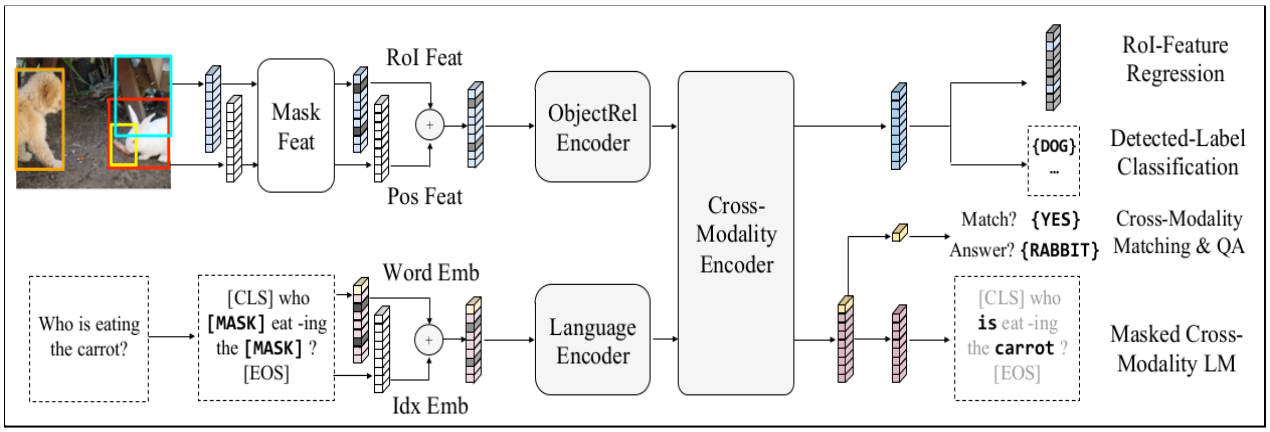


fig. 6. DETAILED PROPOSED MODEL [14]

CHAPTER 3: REQUIREMENT SPECIFICATION AND ANALYSIS

3.1. PROBLEM DEFINITION

To build an interactive and useful system to help visually challenged people to overcome the difficulties they face in day to day activities. The system would consist of an app that enables users to know more about their surroundings. This can be done using VQA technique which primarily is focused on answering free-form open ended questions based on an image.

3.2. SCOPE

Success in developing automated methods would mitigate concerns about undesired consequences from today's status quo for visually impaired people that is, relying on humans to answer visual questions. Example: Humans often must be paid (Cost), can take minutes to provide an answer (Speed), are not always available (Reliability), and pose privacy issues (Privacy).Also, keeping in mind, visually impaired people often are early adopters of computer vision tools to support their real daily needs.

Being a completely user-friendly, free-of-cost, and an absolute necessity for most visually impaired, this product has immense scope in the future to expand further and burgeon in this field as no application exists for this purpose to date.

3.3. OBJECTIVE

The main purpose of this project is to create an android based application that is capable of assisting the visually impaired. At the end of this project we should be able to create an application that can

1. Click a picture using phone camera
2. Record a verbal question
3. Process this image and question using ML model that we have prepared
4. Convert Text based answer to voice output
5. Answer the question using phone's microphone

3.4. PROJECT REQUIREMENT

3.4.1. DATASETS

Visual Question Answering (VQA) is a recent problem in computer vision and natural language processing that has garnered a large amount of interest from the deep learning, computer vision, and natural language processing communities. In VQA, an algorithm needs to answer text-based questions about images. Since the release of the first VQA dataset in 2014[1], additional datasets have been released and many algorithms have been proposed.

In this project we will be focusing on the VizWiz-VQA [4] dataset. It originates from a natural visual question answering setting where visually impaired people each took an image and recorded a spoken question about it, together with ten crowdsourced answers per visual question. Compared to the previous version of this dataset this new version contains more train-validation images with local obfuscation which prevents privacy disclosure.

The evaluation metric followed:

$$Accuracy = min(humans\ that\ answered/3,1)$$

EQ (1)**[1]**

If minimum 3 humans answered the same answer then the prediction is successful otherwise a partial score is offered. Text, color, counting and object identification parameters were used to evaluate the model.

3.4.2. FUNCTIONAL REQUIREMENT

1. EASY TO USE AND INTERACTIVE: Keeping in mind that, specially abled people are the main targeted audience of this product, it is very important that the product should be interactive and easy to use. In this particular case, we have built this app to allow targeted users to get less bothered about starting and interacting with the app. Example, in Fig. 8(Activity Diagram), you can see single clicks anywhere on screen are used to capture user activity events. Also, double tap is used to restart the app activity.
2. VOICE INPUT: Instead of typing in the question, the user can just record the question about the image. which will later be converted to text using some speech to text engine. Example Google Speech API. Using “Voice” as a method of inputting your query saves the user from

going through trouble of typing the question (considering the inadequate ability of the user to see and type on phone).

3. IMAGE INPUT: Users should be able to navigate through the app and click images with a minimum number of interactions (clicks). The clicks should be such that they don't make the user focus on a particular part of the surface of the phone i.e. the user should be able to capture an image by clicking anywhere on screen.

4. ML MODEL PRE-PROCESSING: The Machine Learning model should be able to take input in the form of an image+audio and process it (transform it) to the correct resolution/ dimensions/ length. This should be done in order to improvise algorithm functioning and inference time.

5. GET OUTPUT: Once the input is taken, the system should process the input and give the appropriate answer to the given question. This involves following sub-functions :

   a. GETTING INFERENCE FROM THE MODEL :The Machine Learning model should be able to take input in the form of an image+audio and process it to give an output that is an answer to the question in the input and based on the context of the image. The model should be able to handle real life scenarios significantly.
   b. OUTPUT AS VOICE : Keeping in mind the targeted audience of the app, the app should be able to deliver the output of the ML model in the form of audio. It will help the user to understand their surroundings loud and clear.

6. VOICE FEEDBACK: Voice feedback of every action user does while using the app is an integral requirement of the app. Also, voice commands that suggest the user to take next options and guide him/her through the process should be an added advantage.

3.4.3. NON-FUNCTIONAL REQUIREMENT

1. FAST PROCESSING: As specially abled people are the main targeted audience of this product we should take in consideration that the task given to our product might be an emergency situation. Therefore the system should process the output as fast as possible. The total latency of the system depends on the following factors.

   a. NETWORK LATENCY: Network latency for the app is very much dependent upon the connectivity and speed of the network provider. This is a kind of latency that

cannot be eliminated.

b. SERVER LATENCY: In typical case of 3 tier architecture, the processing speed of the server is a major contributor towards the total latency we can experience. This should be eliminated by hosting the server on a reliable infrastructure. Another way to reduce this latency is to increase computing power of underlying structure by opting for a premium version of the infrastructure provider.

c. INFERENCE LATENCY: Inference latency is the time taken by the machine learning model to predict the answer to the question based on the image. This time should be minimum to get the best user experience.

2. PRIVACY : The privacy and security of the user should be priority. To accomplish this, we are not planning to use any authentication mechanism that asks for user credentials. This app would be free for everyone to use.

3.4.4. HARDWARE REQUIREMENT

1. PC with at least 4 GB RAM and GPU (for development)
2. Android phone with android version 6.0 and above (for user)

3.4.5. SOFTWARE REQUIREMENT

1. Pytorch
2. Tensorflow
3. Google Colab
4. React Native/ Espo client
5. Django Framework
6. Google Speech

3.5. EXISTING SIMILAR SYSTEMS

Existing cross-modality models are limited in the field of application. One existing application named “taptapsee” is the closest existing system to our project. But It does not use the cross-modality principle to answer the diverse set of questions. Instead it just describes the image using image processing.

The key differences:

1. Our app would be voice activated.
2. It will have voice input as well as output.
3. It'll be more user friendly and easy to use.
4. Instead of just describing the image, our app would answer the questions based on the image.

Example:

A visually impaired person is waiting to cross the road. He will open his phone and start the app by voice command. Then he'll click the picture of whatever is in front of him by tapping anywhere on the screen and then they will be prompted to speak out their question about the image that they clicked, they can start recording the question by one tap and similarly stop the recording by again tapping once anywhere on the screen. After this the app will answer the question in natural language.

3.6. PROJECT PLAN

3.6.1 Module Split-UP

The project consists of the following main modules:

1. Image Input from React Native app
2. Audio Input and Output from React Native App
3. Visual Question Answering Module

    3.1 Neural Network Design and Implementation

    3.2 Training Loop Design and Implementation

3.3 Hyperparameter Tuning

3.4 Evaluation Module

These modules, when working together, will be responsible for generating the entire Visual Question Answering System for Visually impaired. All of the modules can be developed independently however the entire system is dependent on all of the modules working in synchronization to carry out the task.

3.6.2 Functional decomposition

3.6.3 PERT TABLE

Sr.	Task Executed	Time Allocated	Team Member
1.	Literature Survey	20 days	Omkar Deshpande, Sanya Varghese
2.	Dataset Collection	15 days	Devesh Chandak
3.	Setup Training Environment Azure	7 days	Shubham Mahajan
4.	Data Preprocessing	20 days	Sanya Varghese
5.	Evaluate different architecture	20 days	Sanya Varghese
6.	Training Models	15 days	Shubham Mahajan
7.	React Native Application	20 days	Omkar Deshpande
9.	TensorRT to optimize model for inference	7 days	Shubham Mahajan
10.	Web Server using Django	15 days	Omkar Deshpande
11.	Testing Debugging and Optimization	20 days	Sanya Varghese
12.	Real World Testing	15 days	--

CHAPTER 4: SYSTEM DESIGN

4.1. ARCHITECTURE

The user should be able to click an image with a button, and record their question with a button. To make the application user-friendly, the buttons would be replaced by taps on the screen. The image and the spoken question once fed into the model would return an answer to the user. The answer would be supported by a speech assistant. The system has been designed to serve as an assisting technology for the visually impaired and blind.

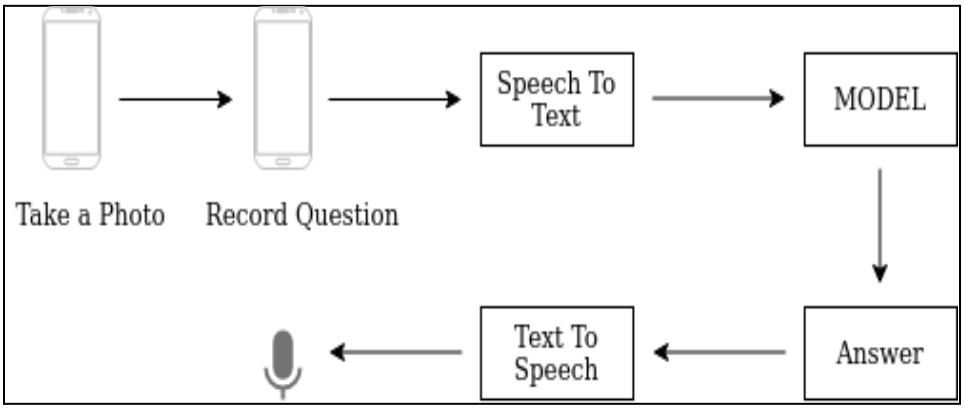
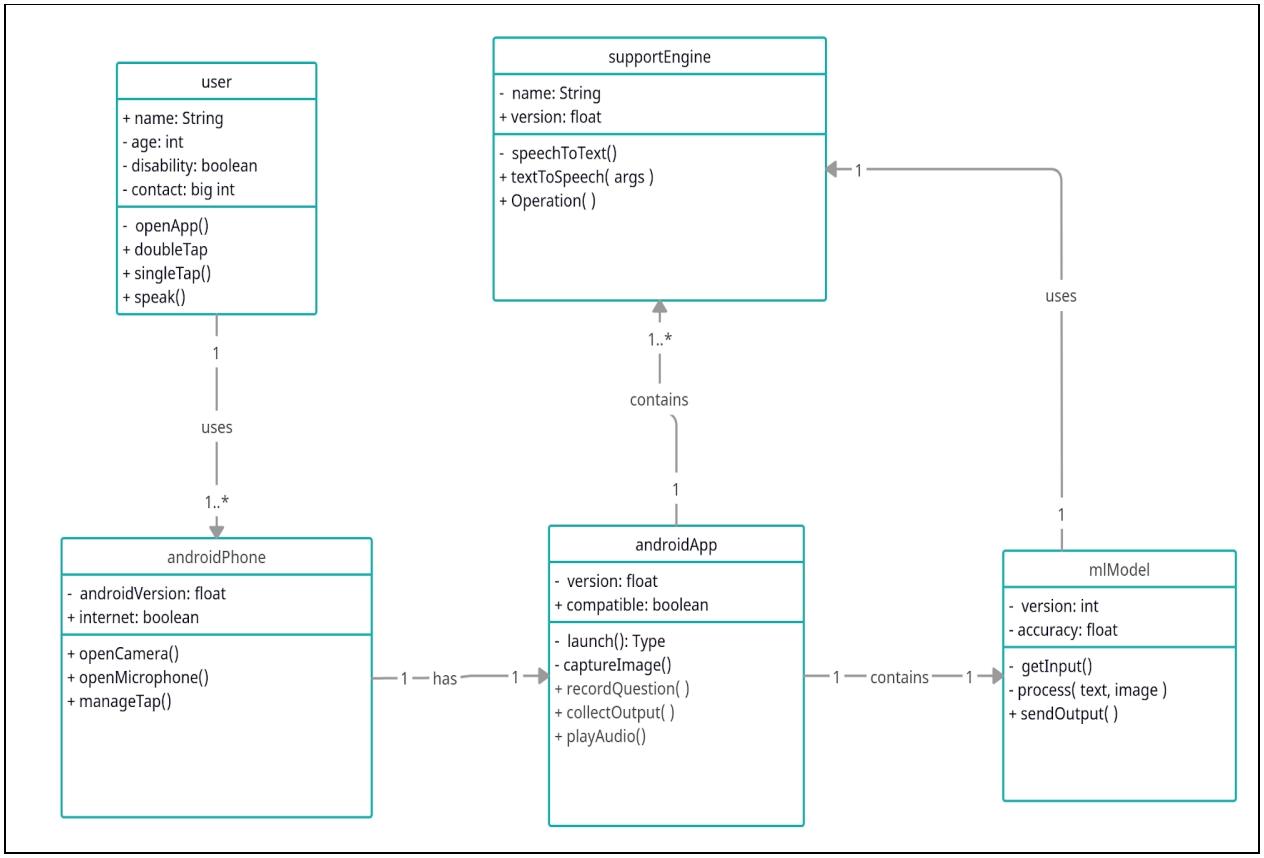


Fig. 5. System Design

Although, we are not planning to implement Speech-To-Text and Text-To-Speech engines by ourselves. There are readily available systems that are trained on very big datasets and have higher accuracies (ex. Google). These systems are more reliable and accurate.

4.2. STRUCTURAL DIAGRAMS



The whole system can be divided into 5 different classes. The user class represents the end-user who is going to use this app. He may provide his name and contact information. He/she can perform actions like opening apps, closing apps or clicking and speaking. The class “androidPhone” represents the device that the user may be using. This device provides the functionality to support the basic app usages. ex. opening camera, opening microphone. The class “androidApp” represents the android application that users will use. This app provides functionalities that are discussed in the previous chapters with the help of other two classes “mlModel” and “supportEngine”. The class supportEngine facilitates text and speech conversions on the other hand mlModel is the core logic of our system.

4.3. BEHAVIOURAL DIAGRAMS

4.3.1 USE CASE DIAGRAM

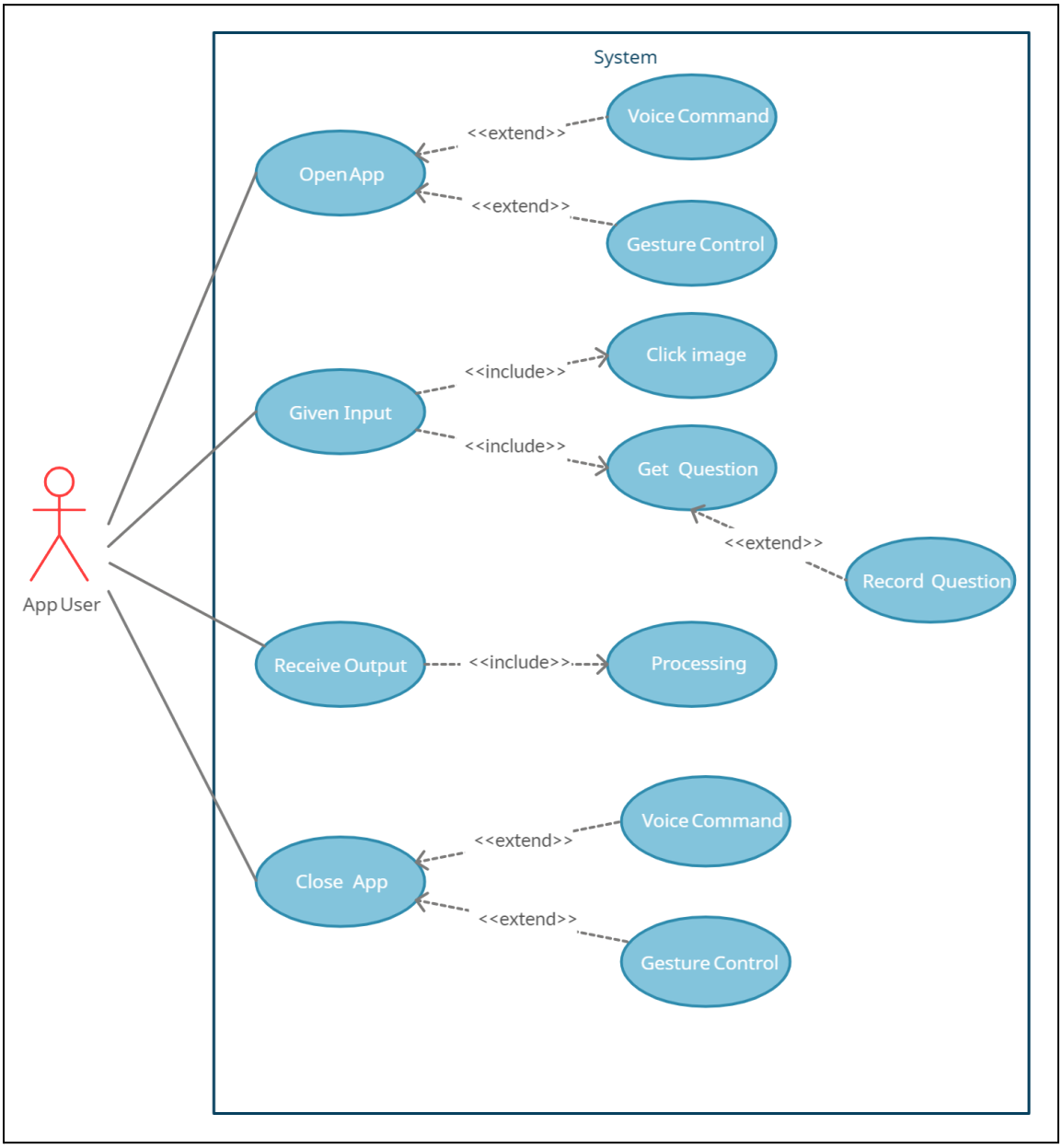


Fig.7. Use Case Diagram

There is only one actor (the end user) and there are four main use cases. The opening of the application may include the use of voice command or gesture control. Users can give input to the app. In order to do that, he must click an image and input the question about it. To input the question he/she may record the question directly or type it manually (if it is

possible). Receiving output includes the processing. The app closing action may include voice command or gesture control.

4.3.2. ACTIVITY DIAGRAM

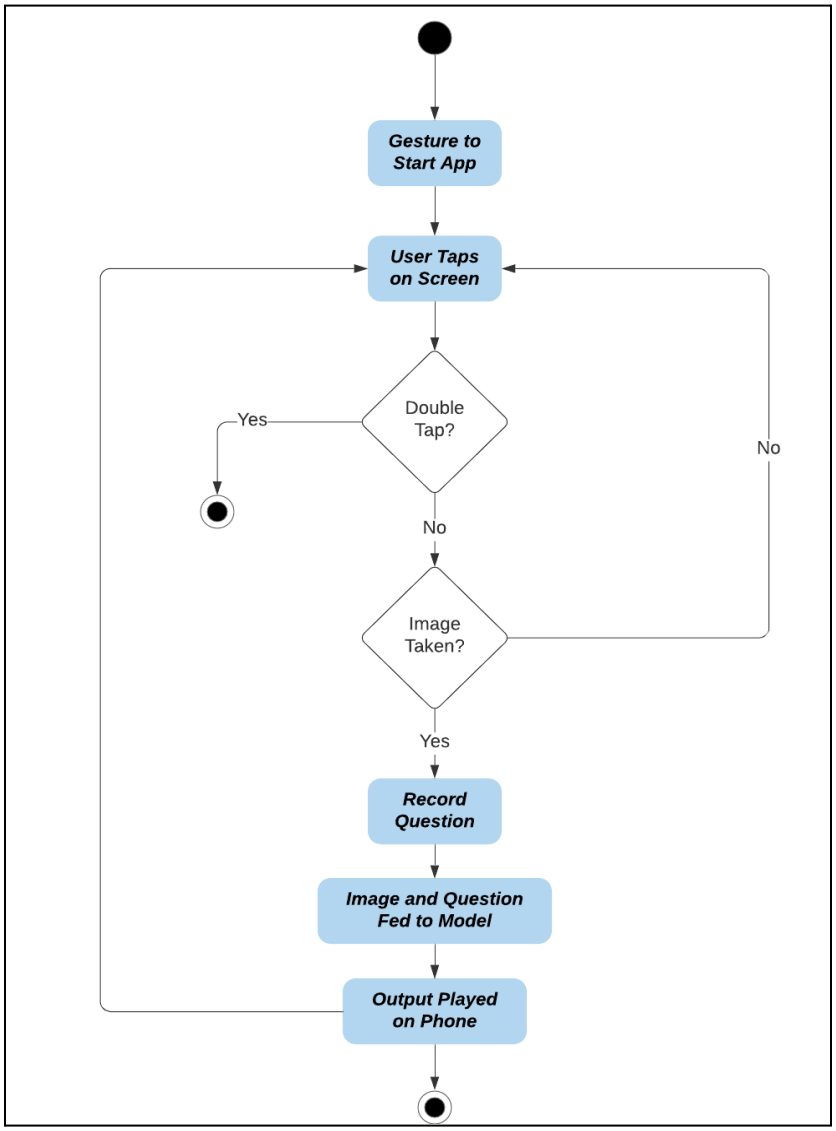


Fig. 8. Activity Diagram

The activity sequence can be states as:

1. User opens the app with voice command or gesture control.
2. User taps on the screen. If it is a double tap, it will exit the app. Otherwise, an image will get clicked.
3. Then a beep will indicate the user to speak and the question will get recorded.

4. The image and question will be fed to the model and the model will generate the output.
5. The output will be played through the speaker and the user will again land on the home page.

4.4 ALGORITHM AND METHODOLOGY

ANDROID APP: Our application is designed with React Native as our primary client-side framework and Django as the backend server-side framework. An API call is made to our ML Model.

ML MODEL: The proposed model is an extension of the BUTD architecture with improved training speed and accuracy. To achieve the best results the BUTD architecture uses the extensively used Top-Down visual attention mechanism. The architecture is well-accepted for image captioning, image description and answering open-ended visual questions(VQA) combined with Bottom-Up attention to capture more salient features.
The top-down method determines feature weightings while the bottom-up method (object recognition) pushes forward sections of the image, each with an associated feature vector.

Pythia(the proposed model) improves results using experiments:

1. VQAv2.0: The model considerably improves the performance of the up-down model on the new dataset.
2. Learning Rate: By continuously incrementing the learning rate from 0.002 to 0.01 in 1000 iterations. This method boosts the performance significantly upto 68.05%.
3. Detectron: The model uses SOTA object detectors which use ResNeXt. This boosts the performance upto 68.49
4. Ensemble: By using a diverse ensemble of models trained with different features and on different datasets, Pythia is able to significantly improve over the 'standard' way of ensembling (i.e. same model with different random seeds) by 1.31 %. Overall, we achieve 72.27% on the test-std split of the VQA v2.0 dataset.

CHAPTER 5: IMPLEMENTATION

5.1. STAGES OF IMPLEMENTATION

5.1.1. DATA PREPARATION

We would be dealing with two different kinds of data within the same dataset. The arrangement of data would be such that our algorithm is able to answer a question with respect to a particular image in context [10]. Thus, each instance of the dataset would have an image with correctly labelled question(s) and answer(s).

5.1.2. PROCESSING

The two different parts of our dataset would require separate kinds of pre-processing. Text, as in any NLP based pipeline would need stop-word removal, tokenization etc. to convert it into a format suitable for input into a neural network. Images on the other hand will be cleaned and made into a uniform size for convolutional operations to run on it properly.

5.2. IMPLEMENTATION SOFTWARE/ TECHNIQUES

We plan on building an application that has minimum reliance on external factors such as network connectivity, so as to ensure excellent availability even in the areas of low or poor network connectivity. For this we plan on not using the 3-tier architecture. The implementation can be divided into 2 Parts.

1. Building an App:

There are two approaches for creating an App viz. using React Native and using Flutter.

The first approach is more preferred because of the following reasons.

a. Compatibility with other npm modules
b. Can be used for cross-platform development
c. Easy to use and efficient
d. Large online community and documentation

The first phase in implementation of the app is allowing the app to click an image. This can be

achieved by providing the whole screen as a single button. Now, if a user clicks anywhere on the screen it will capture the image. This is more interactive and helpful for the blind.

The second phase will include recording a question and converting it into text. This can be done in two ways: using APIs from Google Engine or using python's predefined module “speech_recognition”. The second method is more preferred because we’ll have less reliability on the network and internet.

In the next phase, the inputs will be provided to the ML model. The ML model will be in object serialized format. Now, the model will answer the given question based on the context of the image. Again, convert this answer to speech format using React Native “Voice” module.

5.3 PSEUDO CODES

1. Input capture module

```
# print app opened
mainscreen = getCurrentActivity()
mainscreen.show(system.get(camera))
button capture, listen;
capture.onclick(call captureImage)

function captureImage(){
   Image = system.get(camera).capture
   Image.save()
}

listen.onclick(call captureAudio)

function captureAudio(){
   Audio = system.get(microphone).listen
   Audio.save()
}
# send request
```

2. Send request module

```
request = new Request()
request.url = url
request.headers = headers
formdata = new Formdata()
formdata.add(image:image)
formdata.add(audio:audio)
request.body = formdata
request.send()
# response = request.response
```

3. Get response and output module

```
# response = request.response
String answer;
HashMap<string, string>data
data = response.json()
answer = data.answer
speaker = system.get(speaker)
speaker.speak(answer)
```

4. ML Model

```
Func init():
   Target_audio = TARGET_AUDIO_OPTIONS
   Target_image = TARGET_IMAGE_SIZE
   dataset = load(dataset)
   dataset[audio].setAudioOptions(Target_audio)
   dataset[image].setImageOptions(Target_image)
   transformQuestions(dataset[questions])
   trasnformAnswers(datset[answers])

Func buildModel():
   model  = load(Pythia)
   mode.build()
   model.ApplyCustomConfigurations()
   model.build()

Func predict(image, audio):
   image.transform(Target_image)
   audio.transform(Target_audio)
   model.getPrediction(image, audio)
```

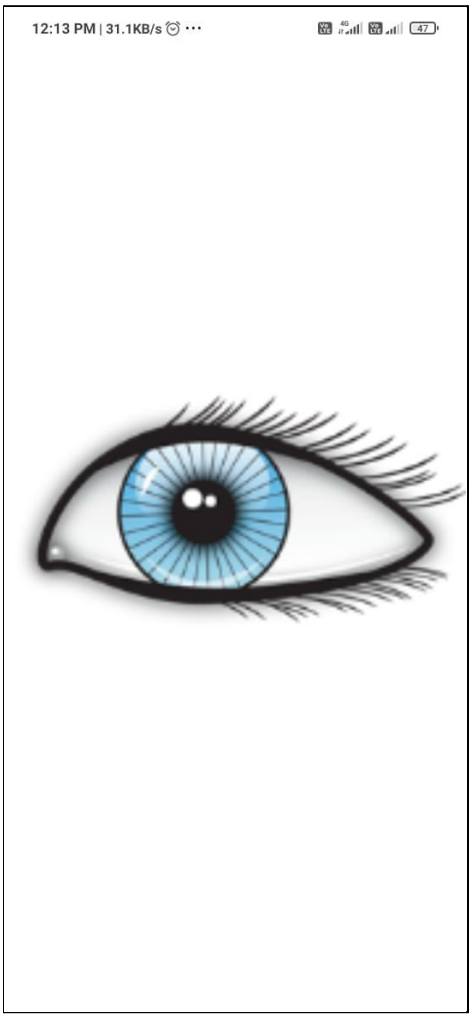
```
    model.eval()
    Return result

Func endpoint(/predict):
    Image = request.body[image]
    Audio = request.body[audio]
    Answer = model.predict(image, audio)
    Return response(answer)
```

5.4 IMPLEMENTATION SNAPSHOTS :

STOP RECORD SCREEN

START RECORD SCREEN



LOADING SCREEN

IMAGE PREVIEW SCREEN

SHOW RESULT SCREEN

CHAPTER 6: RESULTS AND EVALUATION

6.1. Experiments(algorithmic)

6.1.1. Detailed discussion on experiments carried out

6.1.1.1 ML MODEL EXPERIMENT

1. Learning Schedule

We used a warmup strategy to continuously increment the learning rate from 0.002 to 0.01 in 1000 iterations. Further, it is reduced by a factor of 0.1 and training is stopped at 12,000 iterations. This method boosts the performance significantly.

2. Data Augmentation

By adding datasets namely VisualGenome and VisualDialog. Additionally, we also mirror the images on the dataset

3. Object Detection(Detectron)

We use the state-of-the-art detectors which use ResNet based on the feature pyramid net and has two fully connected layers for region classification.

6.1.2. Results of experiments

The objective of the project was to discuss the system design and methodology adopted to design an assisting technology for the blind and visually impaired. We compared different Deep Learning architectures and implemented a real-time application with end-to-end implementation which is user friendly and protects privacy.

Analysing all the test cases, following conclusions can be made -

1. App is user friendly and adheres to the proposed flow.
2. Speech to text module works fine provided clear input voice.
3. The VQA model also performs well in real world scenarios and can be used by real people.

Model	test-dev	test-std
up-down [1]	65.32	65.67
up-down Model Adaptation (§2.1)	66.91	
+ Learning Schedule (§2.2)	68.05	
+ Detectron & Fine-tuning (§2.3)	68.49	
+ Data Augmentation* (§2.4)	69.24	
+ Grid Feature* (§2.5)	69.81	
+ 100 bboxes* (§2.5)	70.01	70.24
Ensemble, 30× same model (§2.6)	70.96	
Ensemble, 30× diverse model (§2.6)	72.18	72.27

6.2. Testing (Test cases)

6.2.1 Unit testing

Project Name	Drishti							
Modeule Name	CLICK_IMAGE							
Test Scenario ID	TS_VQA_01							
Test Scenario Description	Verify Image Capture on single tap							
Creation Date	24-05-2021							
Reviewed Date	26-05-2021							
Test Case ID	Test Case Description	Preconditions	Postcondition	Test Data	Expected Results	Actual Results	Status	Comments (if any)
TC_CLICK_IMAGE_01	Click Image when App is opened for first time.	App installed	Image is captured and you hear the feedback.	N/A	Image Captured	Image Captured	Passed	
TC_CLICK_IMAGE_02	Click Image when App is **NOT** opened for first time.	App installed	Image is captured and you hear the feedback.	N/A	Image Captured	Image Captured	Passed	
Project Name	Drishti							
Modeule Name	START_RECORD							
Test Scenario ID	TS_VQA_02							
Test Scenario Description	Verify Start sound recording on single tap							
Creation Date	25-05-2021							
Reviewed Date	26-05-2021							
Test Case ID	Test Case Description	Preconditions	Postcondition	Test Data	Expected Results	Actual Results	Status	Comments (if any)
TC_START_RECORD_01	Start record question about newly captured image.	Image already captured	"Recording is Started" message seen in logs	N/A	"Recording started" message seen in logs	"Recording started" message seen in logs	Passed	
TC_START_RECORD_02	Start record question about previously captured image. (i.e. new question on same image)	Image already captured	"Recording is Started" message seen in logs	N/A	"Recording started" message seen in logs	"Recording started" message seen in logs	Passed	

Project Name	Drishti
Modeule Name	STOP_RECORD
Test Scenario ID	TS_VQA_03
Test Scenario Description	Verify Stop sound recording on single tap
Creation Date	25-05-2021
Reviewed Date	26-05-2021

Test Case ID	Test Case Description	Preconditions	Postcondition	Test Data	Expected Results	Actual Results	Status	Comments (if any)
TC_STOP_RECORD_01	Stop recording the question.	Recording was previously started	"Recording is Stopped" message seen and sent to Result Page	N/A	Redirected to Result Page.	Redirected to Result Page.	Passed	Recording can only be stopped if it is started previously. There is no intermediate state.

Project Name	Drishti
Modeule Name	STT_QUESTION
Test Scenario ID	TS_VQA_04
Test Scenario Description	Get Speech to text of recorded Question
Creation Date	26-05-2021
Reviewed Date	26-05-2021

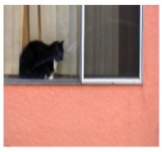
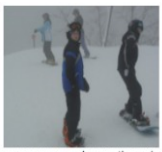
Test Case ID	Test Case Description	Preconditions	Postcondition	Test Data	Expected Results	Actual Results	Status	Comments (if any)
TC_STT_QUESTION_01	Input an audio recording and get the english transcription of the audio	Recorded question already obtained.	Textual transcription of recorded question is sent to the model	Audio data Recording1.m4a	how many people are there in front of me?	how many people are there in front of me?	Pass	Speech to text functionality is outsourced in the form of "google_api"
TC_STT_QUESTION_02	Input an audio recording and get the english transcription of the audio	Recorded question already obtained.	Textual transcription of recorded question is sent to the model	Audio data Recording2.m4a	what animal is in the window?	what animal is in the window?	Pass	Speech to text functionality is outsourced in the form of "google_api"
TC_STT_QUESTION_03	Input an audio recording and get the english transcription of the audio	Recorded question already obtained.	Textual transcription of recorded question is sent to the model	Audio data Recording3.m4a	is there meat in my f ood?	is that meat in my food?	Fail	Substitute "that" for "there"

Project Name	Drishti
Modeule Name	DT_HOME
Test Scenario ID	TS_VQA_06
Test Scenario Description	Double tap to go home page where you can capture a new question.
Creation Date	25-05-2021
Reviewed Date	26-05-2021

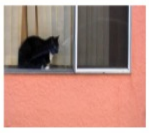
Test Case ID	Test Case Description	Preconditions	Postcondition	Test Data	Expected Results	Actual Results	Status	Comments (if any)
TC_DT_HOME_01	Double tap the screen to go to the home page where you can capture new image	You are on Result Page	Landed on the Home Page	N/A	Redirected to Home Page.	Redirected to Home Page.	Passed	
TC_DT_HOME_02	Double tap the screen to go to the home page where you can capture new image	You are **NOT** on Result	Corresponding Tap functionality executed w.r.t page	N/A	Corresponding Tap functionality executed w.r.t page	Audio record stared and stopped	Passed	The whole process is atomic, you can only if you are on result page

6.2.2. Integration testing

Project Name	Drishti
Modeule Name	VQA_MODEL
Test Scenario ID	TS_VQA_05
Test Scenario Description	Verify VQA model and check the answers status
Creation Date	26-05-2021
Reviewed Date	27-05-2021

Test Case ID	Test Case Description	Preconditions	Postcondition	Test Data	Expected	Actual Res	Status	Comments (if any)
TC_VQA_MODEL_01	Input an image and a text based Question, and get the answer to the question in context of the image.	You have an image and a textual question.	You get a relevant to the question based on the image.	 What animal is in the window ?	Cat	Cat	Passed	
TC_VQA_MODEL_02	Input an image and a text based Question, and get the answer to the question in context of the image.	You have an image and a textual question.	You get a relevant to the question based on the image.	 How many people are there in front of me ?	4	3	Failed	Correct answer should be 4
TC_VQA_MODEL_03	Input an image and a text based Question, and get the answer to the question in context of the image.	You have an image and a textual question.	You get a relevant to the question based on the image.	 Is there meat in my food?	yes	yes	Passed	

Project Name	Drishti
Modeule Name	GET_DATA
Test Scenario ID	TS_VQA_07
Test Scenario Description	Check if the data correctly received correctly to the server.
Creation Date	26-05-2021
Reviewed Date	26-05-2021

Test Case ID	Test Case Description	Preconditions	Postcondition	Test Data	Expected Results	Actual Results	Status	Comments (if any)
TC_GET_DATA_01	Check if the image and audio file correctly received to the server	You have an image and an audio file	Data received and sent to model.	 audio.m4a	Image and sound file received.	Image and sound file received.	Passed	
TC_GET_DATA_02	Check if the image and audio file correctly received to the server	You have an image and an audio file	Data received and sent to model.	 audio.m4a	Image and sound file received.	Image and sound file received.	Passed	The whole process is atomic, you can only if you are on result page

CHAPTER 7: CONCLUSION

7.1 Limitations

As our system includes Deep Learning Models, this makes it computationally heavy and expensive, thus raising issues for deployment. As a result, constant updates and efforts are in progress to try and lighten the model, thus making it easy for deployment and ready to use for any individual.

The model struggles with open-ended questions “Can I cross the road?” and “Is it safe for me to cross the road?” do not give the same result.

Another drawback would be that our model cannot be implemented or executed on any normal machine. A specialized GPU is a must which is required due to the heaviness of the model.

7.2 Future Scope

To extend the functionality of the project an Optical Character Reader(OCR) can be included. It would further increase the viability of the product. With a separate OCR, the app would be able to answer more critical questions based on the image of a textual writing and give better results with text based questions.

7.3 Conclusion

The objective of the project was to discuss the system design and methodology to be adopted to design an assistive technology for the blind and visually impaired. We compared different Deep Learning architectures and proposed a real-time application which is user friendly and protects privacy.

CHAPTER 8 : REFERENCES

[1] Antol, Stanislaw, et al. ”Vqa: Visual question answering.” Proceedings of the IEEE international conference on computer vision. 2015.

[2] Kazemi, Vahid, and Ali Elqursh. ”Show, ask, attend, and answer: A strong baseline for visual question answering.” (2017).

[3] Anderson, Peter, et al. ”Bottom-up and top-down attention for image captioning and visual question answering.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[4] Gurari, Danna, et al. ”Vizwiz grand challenge: Answering visual questions from visually impaired people.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

[5] Jiang, Yu, et al. ”Pythia v0. 1: the winning entry to the vqa challenge 2018.” (2018).

[6] Aafaq, Nayyer, et al. ”Video description: A survey of methods, datasets, and evaluation metrics.” ACM Computing Surveys (CSUR) 52.6 (2019):

[7] Srivastava, Yash, et al. ”Visual Question Answering using Deep Learning: A Survey and Performance Analysis.” (2019).

[8] Teney, Damien, et al. ”Tips and tricks for visual question answering: Learnings from the 2017 challenge.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[9] Bigham, Jeffrey P., et al. ”VizWiz: nearly real-time answers to visual questions.” Proceedings of the 23nd annual ACM symposium on User interface software and technology. 2010.

[10] Li, Xiujun, et al. ”Oscar: Object-semantics aligned pre-training for vision-language tasks.” European Conference on Computer Vision. Springer, Cham, 2020.

[11] Kafle, Kushal, and Christopher Kanan. ”Answer-type prediction for visual question answering.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[12] Shih, Kevin J., Saurabh Singh, and Derek Hoiem. ”Where to look: Focus regions for visual question answering.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[13] Gurari, Danna, et al. ”Captioning Images Taken by People Who Are visually impaired.” arXiv preprint arXiv:2002.08565 (2020)

[14] Tan, Hao, and Mohit Bansal. ”Lxmert: Learning cross-modality encoder representations from transformers.” (2019).

[15] Vaswani, Ashish, et al. ”Attention is all you need.” Advances in neural information processing systems. 2017.

[16] Differential Networks for Visual Question Answering C Wu, J Liu, X Wang, R Li - Proceedings of the AAAI Conference on Artificial …, 2019

Appendices

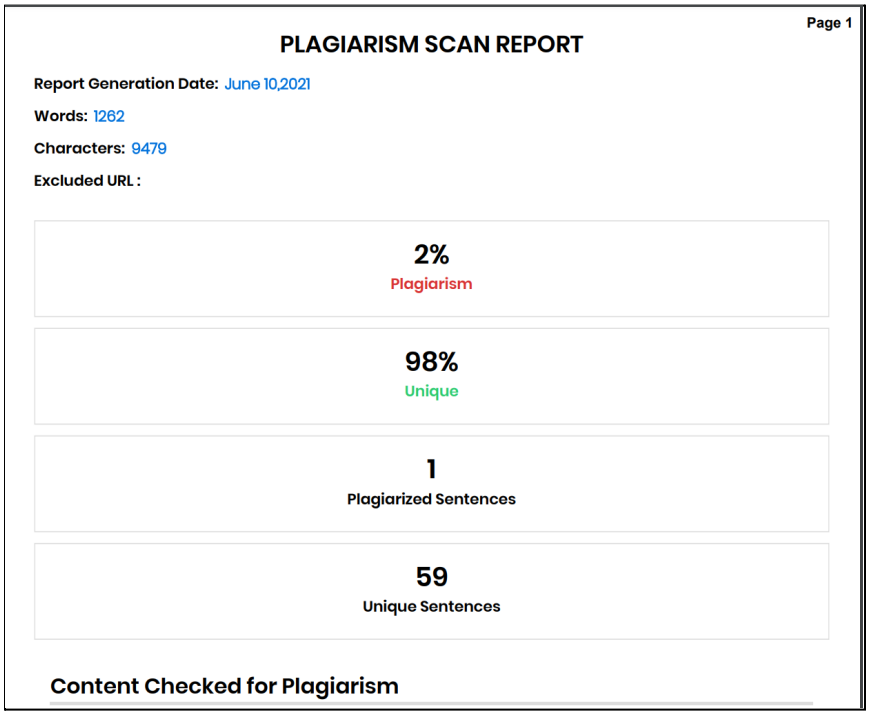
[A]
PLAGIARISM REPORTS
Semester 1

Document Information

Analyzed document	Group 24 - Report for Plag Check.pdf (D90217840)
Submitted	12/19/2020 10:22:00 AM
Submitted by	Shweta Dharmadhikari
Submitter email	scdharmadhikari@pict.edu
Similarity	15%
Analysis address	scdharmadhikari.pict@analysis.urkund.com

Semester 2

Final paper



Page 1

PLAGIARISM SCAN REPORT

Report Generation Date: June 10,2021

Words: 1262

Characters: 9479

Excluded URL :

2%
Plagiarism

98%
Unique

1
Plagiarized Sentences

59
Unique Sentences

Content Checked for Plagiarism

Final Report

Report Title:	report
Report Link: (Use this link to send report to anyone)	https://www.check-plagiarism.com/plag-report/308502ad1ccdc36c2c7df9207bc31ea2391c01623327007
Report Generated Date:	10 June, 2021
Total Words:	4397
Total Characters:	32428
Keywords/Total Words Ratio:	0%
Excluded URL:	No
Unique:	94%

Matched:	6%

[B]
Base paper

Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering

[3] Anderson, Peter, et al. ”Bottom-up and top-down attention for image captioning and visual question answering.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

Link: https://arxiv.org/pdf/1707.07998.pdf

[C]
PLAGIARISM REPORTS
Semester 1
Appendices

PUNE INSTITUTE OF COMPUTER TECHNOLOGY, PUNE.
Department of Information Technology
(Academic Year: 2020-21)

Semester - I

Monthly Planning Sheet

Week No.	Activity Planned	Activity Completed Status	Student Signature	Guide Signature
Week 1	Exploring Topics	Completed		
Week 2	FInalize topic	Completed		
Week 3	Literature survey	Completed		
Week 4	LIterature survey	Completed		Dr. S .C. Dharmadhikari
Week 5	Requirement Analysis	Completed		
Week 6	Requirement Identification	Completed		
Week 7	Existing methods evaluation	Completed		
Week 8	System analysis and design	Completed		Dr. S .C. Dharmadhikari
Week 9	Decide architecture	Completed		
Week 10	Study implementation techniques and tools	Completed		
Week 11	Prototype development	Completed		
Week 12	Prototype development	Completed		Dr. S .C. Dharmadhikari

Project Coordinator

Internal Guide

PUNE INSTITUTE OF COMPUTER TECHNOLOGY, PUNE.
Department of Information Technology

ACHIEVEMENTS

(Academic Year: 2020-21 Semester-I)

Group ID:	24			Date:
Project Title: Visual Question Answering for the Visually Impaired				
Sr.No.	**Roll No.**	**Student Name**	**Contact Details**	**Internal / External Guide Details**
1	43165	Shubham Mahajan	8602775207	Dr. S.C. Dharmadhikari
2	43212	Omkar Deshpande	7709833124	
3	43213	Devesh Chandak	7972663093	
4	43253	Sanya Varghese	9205909374	

Project Competition/ Exhibition

Sr. No.	Name & Place of Project Competition/ Exhibition	Date	Prizes won (if any) or Participation	State / National/ International Level
	NIL	NIL	NIL	NIL

Paper Publication/ Presentation

Sr. No.	Paper Title	Authors	Date	Journal/Conference Name	Indexing	DOI
	NIL	NIL	NIL	NIL	NIL	NIL

Patent Details

Sr. No.	Name of the Patent Holder	Patent No	Title of the Patent	Type of Patent	Status of Patent	Any other information
	NIL	NIL	NIL	NIL	NIL	NIL

Any other achievement:

*** Photocopy of the certificate must be attached to this booklet.**

Name & Signature of Internal Guide
Dr. S. C. Dharmadhikari

PUNE INSTITUTE OF COMPUTER TECHNOLOGY, PUNE.
Department of Information Technology
PROJECT REVIEW – I
(Academic Year:2020-21)

Group ID:	24			Date:
Project Title: Visual Question Answering for the Visually Impaired				
Sr.No.	**Roll No.**	**Student Name**	**Contact Details**	**Internal / External Guide Details**
1	43165	Shubham Mahajan	8602775207	Dr. S.C. Dharmadhikari
2	43212	Omkar Deshpande	7709833124	
3	43213	Devesh Chandak	7972663093	
4	43253	Sanya Varghese	9205909374	

REVIEW – I CHECKLIST : FINALIZATION OF SCOPE **25 Marks**

	PROJECT STATEMENT	
1.	Is the statement short and concise (10-20 words maximum)?	**Y** / N / NA / NC*
2.	Does the statement give clear indication about what your project will accomplish?	**Y** / N / NA / NC*
3.	Can a person who is not familiar with the project understand the scope of the project by reading the problem statement?	**Y** / N / NA / NC*
	REQUIREMENT: SCOPE AND OBJECTIVES	
	Does the Scope and Objectives establish the "context" for the proposed project by referencing to the following elements:	
a.	Are all aspects of the requirements document (i.e., Functional Spec.) addressed in the design	**Y** / N / NA / NC*
b.	Is the architecture / block diagram well defined and understood?	**Y** / N / NA / NC*
c.	The project's objective of study (what product, process, resource etc.) is being addressed?	**Y** / N / NA / NC*

d.	The project's purpose is the purpose of the project addressed properly (why it's being pursued:to evaluate, reduce, increase, etc.)?	**Y** / N / NA / NC*
e.	The project's viewpoint: Is the project's viewpoint is understood? (Who is the project's end user)?	**Y** / N / NA / NC*
f.	Is the project goal statement in alignment with the sponsoring organization's businessgoals and mission?	Y / N / **NA** / NC*
	ANALYSIS	
1.	Is information domain analysis complete, consistent and accurate?	**Y** / N / NA / NC*
2.	Is the problem statement categorized in identified areas and targeted towards specific areas therein?	**Y** / N / NA / NC*
3.	Are external and internal interfaces properly defined?	**Y** / N / NA / NC*
4.	Does the Use Case Model properly reflect the actors and their roles and responsibilities?	**Y** / N / NA / NC*
5.	Are all requirements traceable to system level?	**Y** / N / NA / NC*
6.	Is a similar type of methodology / model used for existing work?	**Y** / N / NA / NC*
7.	Are requirements consistent with schedule, resources, and budget?	**Y** / N / NA / NC*

PUNE INSTITUTE OF COMPUTER TECHNOLOGY, PUNE
Department of Information Technology
PROJECT REVIEW – I
(Academic Year: 2020-21)
STUDENT PERFORMANCE EVALUATION

Students' Contribution and Performance

Particulars		Marks(25M)			
		Group Members			
		1	2	3	4
1.	Background and Topic (4 M)	4	4	4	4
2.	Project Scope and Objectives (4M)	4	4	4	4
3.	Literature Survey (5 M)	5	5	5	5
4.	Project Planning (4 M)	4	4	4	4
5.	Presentation Skills (4 M)	3	3	3	3
6.	Question and Answer (4 M)	3	3	3	3
	Total(25M)	23	23	23	23

Comments if any: Satisfactory work

To be filled by internal guide & reviewer(s) only.

* Whether the presentation / evaluation is as per the schedule. : **YES** / NO (If NO mention the reasons for the same.)

Review – I: Deliverables

• Problem Statement / Title • Purpose, Scope, Objectives • Abstract (System Overview) • Introduction (Architecture and High-Level Design)	• H/W, S/W & other requirement, Test Environment/Tools • Literature Survey • References • Project Plan 1.0 (**Gantt Chart)**

Name & Signature of evaluation committee –

Name of Reviewer Name of Reviewer 2 Name of Internal Guide

PUNE INSTITUTE OF COMPUTER TECHNOLOGY, PUNE.
Department of Information Technology
PROJECT REVIEW – II
(Academic Year:2020-21)

Group ID:	24			Date:
Project Title: Visual Question Answering for the Visually Impaired				
Sr.No.	**Roll No.**	**Student Name**	**Contact Details**	**Internal / External Guide Details**
1	43165	Shubham Mahajan	8602775207	Dr. S.C. Dharmadhikari
2	43212	Omkar Deshpande	7709833124	
3	43213	Devesh Chandak	7972663093	
4	43253	Sanya Varghese	9205909374	

REVIEW – II CHECKLIST : DESIGN	25 Marks

DESIGN		
1.	Are requirements reflected in the system architecture?	**Y** / N / NA / NC*
2.	Does the design support both project (product) and project goals?	**Y** / N / NA / NC*
3.	Does the design address all the issues from the requirements?	**Y** / N / NA / NC*
4.	Is effective modularity achi and modules are functionally independent?	**Y** / N / NA / NC*
5.	Are structural diagrams (Class, Object, etc.) well defined and understood?	**Y** / N / NA / NC*
6.	Are all class associations clearly defined and understood? (Is it clear which classes provide which services)?	**Y** / N / NA / NC*
7.	Are the classes in the class diagram clear? (What do they represent in the architecture design document?	**Y** / N / NA / NC*
8.	Is inheritance appropriately used?	Y / N / **NA** / NC*
9.	Are the multiplicities in the use case diagram depicted in the class diagram?	**Y** / N / NA / NC*
10.	Are behavioral diagrams (use case, sequence, activity, etc.) well defined and understood?	**Y** / N / NA / NC*
11.	Is aggregation/containment (if used) clearly defined and understood?	**Y** / N / NA / NC*
12.	Does each case have clearly defined actors and input/output?	**Y** / N / NA / NC*

13.	Is all concurrent processing (if used) clearly understood and reflected in the sequence diagrams?	**Y** / N / NA / NC*
14.	Are all objects used in sequence diagram?	**Y** / N / NA / NC*
15.	Does the sequence diagram match class diagram?	**Y** / N / NA / NC*
16.	Are the symbols used in all diagrams correspond to UML standards?	**Y** / N / NA / NC*

PUNE INSTITUTE OF COMPUTER TECHNOLOGY, PUNE.

Department of Information Technology

PROJECT REVIEW – II

(Academic Year: 2020-21)

STUDENT PERFORMANCE EVALUATION

Students' Contribution and Performance

	Particulars	Marks(25 M)			
		Group Members			
		1	2	3	4
1.	System Architecture & Literature Survey (Review-I)	**Y**	**Y**	**Y**	**Y**
2.	Project Design (5 M)	5	5	5	5
3.	Methodology /Algorithms and Project Features (5M)	5	5	5	5
4.	Project Planning (2 M)	2	2	2	2
5.	Basic details of Implementation (5 M)	4	4	4	4
6.	Presentation Skills (4 M)	4	4	4	4
7.	Question and Answer (4 M)	3	3	3	3
8.	Summarization of ultimate findings of the Project	**Y**	**Y**	**Y**	**Y**
	Total(25M)	23	23	23	23

Comments if any: Satisfactory Work

To be filled by internal guide & reviewer(s) only.

* Whether the presentation / evaluation is as per the schedule. : **YES** / NO (If NO mention the reasons for the same.)

Review – II: Deliverables

• Problem Statement / Title • Abstract • Introduction • Literature Survey (comparison with existing system) • Methodology • Design / algorithms / techniques used	Modules Split-up Proposed System Software Tools / Technologies to be used Proposed Outcomes Partial Report (Semester – I) Project Plan 2.0 (Gantt Chart)

Name & Signature of evaluation committee

Name of Reviewer 1 Name of Reviewer 2 Name of Internal Guide

PUNE INSTITUTE OF COMPUTER TECHNOLOGY, PUNE.

Department of Information Technology

RESEARCH PUBLICATION REVIEW – I

(Academic Year: 2020-21)

STUDENT PERFORMANCE EVALUATION

Students' Contribution and Performance

	Particulars	Marks(25M)			
		Group Members			
		1	2	3	4
1.	System Architecture & Literature Survey (Review-I)	**Y**	**Y**	**Y**	**Y**
2.	Precise Title, Abstract and Keywords (2 M)	2	2	2	2
3.	Motivation and scope of research work (2 M)	2	2	2	2
4.	Literature Survey and identification of research gap (5 M)	3	3	3	3
5.	Proposed Methodology /Algorithm/System Architecture (5M)	5	5	5	5
6.	Effective Conclusion and Future Scope (2 M)	2	2	2	2
7.	Relevant References (3 M)	3	3	3	3
8.	Effective Technical Writing and Presentation Skills (4 M)	3	3	3	3
9.	Originality (Plagiarism <20%) (2M)	**2**	**2**	**2**	**2**
10.	Identification of quality journals/international conferences	**Y**	**Y**	**Y**	**Y**
	Total(25M)	22	22	22	22

Comments if any: -

To be filled by internal guide & reviewer(s) only.

* Whether the presentation / evaluation is as per the schedule. : **YES** / NO (If NO mention the reasons for the same.)

Research Publication Review – I: Deliverables

• Paper Title, Abstract and keywords • Introduction • Literature Survey • Proposed Methodology/ Algorithm • System Architecture/ Workflow Diagram	• Conclusion and Future Scope • References • Identified WoS /Scopus indexed and /or UGC listed international journals and/or Scopus indexed international conferences.

Name & Signature of evaluation committee –

Name of Reviewer 1 Name of Reviewer 2 Name of Internal Guide

PUNE INSTITUTE OF COMPUTER TECHNOLOGY, PUNE.

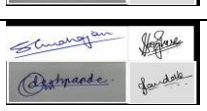
Department of Information Technology

(Academic Year: 2020-21)

Semester - II

Monthly Planning Sheet

Academic Year: 2020-2021

Week No.	Activity Planned	Activity Completed Status	Student Signature	Guide Signature
Week 1	Implement basic App	Completed		
Week 2	Get BUTD pre trained model working.	Completed		
Week 3	Develop the app flow and bind all activities	Completed		
Week 4	Train the model on VQA 2.0 dataset.	Completed		Dr. S. C. Dharmadhikari
Week 5	Fine tuning and improving results	Completed		
Week 6	Fine tuning and improving results.	Completed		
Week 7	Check app and ML model compatibility	Completed		
Week 8	Test different integration techniques	Completed		Dr. S. C. Dharmadhikari
Week 9	implement the best integration method	Completed		
Week 10	host ML model on server and publish app on play store.	Completed		
Week 11	Unit and integration testing (real world testing)	Completed		Dr. S. C. Dharmadhikari
Week 12	Rectify changes from test results	Completed		

Mrs. Radhika V. Kulkarni
Project Coordinator

Dr. S. C. Dharmadhikari
Internal Guide

PUNE INSTITUTE OF COMPUTER TECHNOLOGY, PUNE.

Department of Information Technology

PROJECT REVIEW – III

(Academic Year: 2020-21)

Group ID:	24			Date:30/05/2021
Project Title: Visual Question Answering for the Visually Impaired				
Sr.No.	**Roll No.**	**Student Name**	**Contact Details**	**Internal / External Guide Details**
1	43165	Shubham Mahajan	8602775207	Dr. S.C. Dharmadhikari
2	43212	Omkar Deshpande	7709833124	
3	43213	Devesh Chandak	7972663093	
4	43253	Sanya Varghese	9205909374	

REVIEW – III : IMPLEMENTATION			**25 Marks**
IMPLEMENTATION (SOURCE CODE REVIEW CHECKLIST)			
a.	**Structure**		
	1.	Does the code completely and correctly implement the design?	**Y** / N / NA / NC*
	2.	Does the code comply with the Coding Standards?	**Y** / N / NA / NC*
	3.	Is the code well-structured, consistent in style, and consistently formatted?	**Y** / N / NA / NC*
	4.	Does the implementation match the design?	**Y** / N / NA / NC*
	5.	Are all functions in the design coded?	**Y** / N / NA / NC*
b.	**Documentation**		Y / N / NA / NC*
	1.	Is the code clearly and adequately documented?	**Y** / N / NA / NC*
	2.	Are all comments consistent with the code?	**Y** / N / NA / NC*

PUNE INSTITUTE OF COMPUTER TECHNOLOGY, PUNE.

Department of Information Technology

PROJECT REVIEW – III

(Academic Year: 2020-21)

STUDENT PERFORMANCE EVALUATION

Students' Contribution and Performance

		Marks(25M)			
	Particulars	Group Members			
		1	2	3	4
1.	Architecture / System Design -(if any modification)	N	N	N	N
2.	60 % Implementation (10 M)	10	10	10	10
3.	Partial results obtained (7 M)	6	6	6	6
4.	Presentation skills (4 M)	4	4	4	4
5.	Question and Answer (4 M)	3	3	3	3
6.	Summarize the methodologies / Algorithms	Y	Y	Y	Y
	implemented / to be implemented				
	Total(25M)	24	24	24	24
Comments (if any)					

Name & Signature of evaluation committee –

Mr. Jagdish K. Kamble
Name of Reviewer 1

Dr. S. C. Dharmadhikari
Name of Internal Guide

PUNE INSTITUTE OF COMPUTER TECHNOLOGY, PUNE.

Department of Information Technology

PROJECT REVIEW – IV

(Academic Year: 2020-21)

Group ID:	24			Date:30/05/2021
Project Title: Visual Question Answering for the Visually Impaired				
Sr.No.	**Roll No.**	**Student Name**	**Contact Details**	**Internal / External Guide Details**
1	43165	Shubham Mahajan	8602775207	Dr. S.C. Dharmadhikari
2	43212	Omkar Deshpande	7709833124	
3	43213	Devesh Chandak	7972663093	
4	43253	Sanya Varghese	9205909374	

IMPLEMENTATION AND TESTING

1.	Is every feature tested?		**Y** / N / NA / NC*
2.	Are all functions, user screens and navigation tested? (e.g. module, object,integration, usability, system)		**Y** / N / NA / NC*
3.	Are test cases designed? (manual and automated)		**Y** / N / NA / NC*
4.	Is a testing tool used?		**Y** / N / NA / NC*
5.	Is result analysis done properly and appropriate conclusions drawn?		**Y** / N / NA / NC*
6.	Implementation status (code completion in percentage)		95%
7.	Final thesis status(in percentage)		95%
FILL IN BRIEF			
Final results are known or not? :Known			
Quality of Presentation : Good			
List the chapter numbers of final report : Done			
Project Completion Date :			
Final Report Submission Date :			

General: Is the LOG BOOK of the project up-to-date and signed? - Yes

PUNE INSTITUTE OF COMPUTER TECHNOLOGY, PUNE.

Department of Information Technology

PROJECT REVIEW – IV

(Academic Year: 2020-21)

STUDENT PERFORMANCE EVALUATION

Students' Contribution and Performance

		Marks(25M)			
	Particulars	**Group Members**			
		1	2	3	4
1.	Implementation (100%) (5 M)	4	4	4	4
2.	Testing, Results and Performance Evaluation (5 M)	5	5	5	5
3.	Final Project Report (5 M)	5	5	5	5
4.	Publications (2 M)	2	2	2	2
5.	Presentation skills (4 M)	4	4	4	4
6.	Question and Answer (4 M)	4	4	4	4
	Total(25M)	24	24	24	24

Comments (if any)

Name & Signature of evaluation committee –

Mr. Jagdish K. Kamble
Name of Reviewer 1

Dr. S. C. Dharmadhikari
Name of Internal Guide

PUNE INSTITUTE OF COMPUTER TECHNOLOGY, PUNE.

Department of Information Technology

RESEARCH PUBLICATION REVIEW – II

(Academic Year: 2020-21)

Group ID:	24			Date:30/05/2021
Project Title: Visual Question Answering for the Visually Impaired				
Sr.No.	**Roll No.**	**Student Name**	**Contact Details**	**Internal / External Guide Details**
1	43165	Shubham Mahajan	8602775207	Dr. S.C. Dharmadhikari
2	43212	Omkar Deshpande	7709833124	
3	43213	Devesh Chandak	7972663093	
4	43253	Sanya Varghese	9205909374	

RESEARCH PUBLICATION REVIEW – II CHECKLIST	**25 Marks**
Publication based on the Experimentation Results	
1. Is the Problem Clearly defined and concise? (Which Challenge / issue is addressed by this research?)	**Y** / N / NA / NC*
2. Is Abstract precisely written and are Keywords correctly identified?	**Y** / N / NA / NC*
3. Is the motivation/significance of the research work defined?	**Y** / N / NA / NC*
4. Is Literature Survey comprehensive, systematic?	**Y** / N / NA / NC*
5. Is contribution of the research work is clearly described?	**Y** / N / NA / NC*
6. Is new methodology/algorithm proposed precisely?	**Y** / N / NA / NC*
7. Does the system architecture/ workflow diagram match the proposed methodology?	**Y** / N / NA / NC*
8. Are the experimentation setup and results discussed systematically?	**Y** / N / NA / NC*
9. Is the empirical study compares the results with the state-of-the-art algorithms?	**Y** / N / NA / NC*
10 Is conclusion with future scope communicated effectively? .	**Y** / N / NA / NC*

11 Is plagiarism checked? .	**Y** / N / NA / NC*
12 Are the WoS /Scopus indexed and /or UGC listed international journals . and/or Scopus indexed international conferences identified?	**Y** / N / NA / NC*

PUNE INSTITUTE OF COMPUTER TECHNOLOGY, PUNE.

Department of Information Technology

RESEARCH PUBLICATION REVIEW – II

(Academic Year: 2020-21)

STUDENT PERFORMANCE EVALUATION

Students' Contribution and Performance

		Marks(25M)			
	Particulars	Group Members			
		1	2	3	4
1.	Implementation (Review-III)	Y	Y	Y	Y
2.	Precise Title, Abstract and Keywords (2 M)	2	2	2	2
3.	Motivation and contribution of research work (2 M)	2	2	2	2
4.	Literature Survey and identification of research gap (2M)	2	2	2	2
5.	Proposed Methodology /Algorithm/System Architecture (4M)	4	4	4	4
6.	Experimentation Results and Empirical Analysis (5M)	4	4	4	4
6.	Effective Conclusion and Future Scope (2 M)	2	2	2	2
7.	Relevant References (2 M)	2	2	2	2
8.	Effective Technical Writing and Presentation Skills (4 M)	3	3	3	3
9.	Originality (Plagiarism <20%) (2M)	2	2	2	2
10.	Identification of quality journals/international conferences	Y	Y	Y	Y
	Total(25M)	23	23	23	23
Comments (if any)					

Name & Signature of evaluation committee –

Mr. Jagdish K. Kamble
Name of Reviewer 1

Dr. S. C. Dharmadhikari
Name of Internal Guide

International Journal of Recent Technology and Engineering
ISSN: 2277-3878 (online) | Exploring Innovation | A Key for Dedicated Services
Published by: Blue Eyes Intelligence Engineering and Sciences Publication
G-18, Block A, Tirupati Abhinav Commercial Campus, Tirupati Abhinav Homes, Ayodhya Bypass Road, Damkheda, Bhopal (M.P.)-462037, India
Website: www.ijrte.org Email: submit2@ijrte.org

Acceptance Letter

Lattice Science Bibliographic Database

Dear Author(s): Devesh Chandak, Omkar Deshpande, Sanya Varghese, Shubham Mahajan, Shweta Dharmadhikari

Paper ID:	B60830710221
Paper Title:	Visual Question Answering for Visually Impaired

The above manuscript appraised by the proficient and it is **accepted** by the Board of Referees (BoR) of 'Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)' for publication in the **'International Journal of Recent Technology and Engineering (IJRTE)'** at **Volume-10 Issue-2, July 2021** in Regular Issue on **30 July 2021.** It will be available live at https://www.ijrte.org/download/volume-10-issue-2/

It is advised you to provide us **following supporting documents in a single email** before 16 June 2021 at submit2@ijrte.org

1. **Final Paper | Ms Word .doc | docx. file**
   camera ready paper should be prepared as per journal template which is available at http://ijrte.org/download/ . If you are not able to convert manuscript as per journal template then you can send manuscript in single column format. Concern team will convert your manuscript as per journal template on behalf you
2. **Copyright Transfer Form | Submit Online only | do not send it through email**
   https://www.blueeyesintelligence.org/copyright/
3. **Proof of Registration | Scanned | Online Received Email**
   please visit in this URL. All details available at https://www.blueeyesintelligence.org/registration/

INFORMATION FOR AUTHOR(S)- Please read very carefully.

1. Each author (s) profile of minimum of 100 words along with a photo should be available in the final paper. The final paper should be prepared as per the journal template. The Paper should have a minimum of 03 pages and a maximum of 10 pages. Maximum 05 authors can be seated in a paper. In the case of more than 05 authors, the paper (s) to be rejected. Final paper should not have more than 30% plagiarism including reference section.
2. If the above three supporting documents (Final Paper, Copyright and Registration) does not submit to the journal by the author in the given date (s), then paper will automatically suspend from publication for particular volume/issue. During the final email, you have to attach Final Paper, Copyright and Proof of Registration in a single email. Final paper should be result oriented and should be prepared as per the reviewer (s) comment (s). In the case of failure, it to be suspended for correction. Please read review report carefully. It is compulsory to write the Paper ID of the paper in place of Subject Area in the email during the final paper submission. Header and footer of the paper template will be edited by journal staff only.
3. Author (s) can make rectification/updation in the final paper but after the signing the copyright and final paper submission to the journal, any rectification/updation is not possible. Published paper to be available online from 30 July to 05 August 2021. Paper can not withdraw after submitting the copyright to the journal. Author(s) will receive publication certificate within 01 to 02 weeks after the date of publication of respective volume/issue
4. The DOI can be checked and verified within 02 to 04 weeks after the date of publication of volume/issue: https://www.doi.org/

Jitendra Kumar Sen
(Manager)

Blue Eyes Intelligence Engineering and Sciences Publication BHOPAL

Dr. Shiv Kumar
(Editor-In-Chief)