A PRELIMINARY 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	**43165**
Omkar Deshpande	**43212**
Devesh Chandak	**43213**
Sanya Varghese	**43253**

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	**43165**
Omkar Deshpande	**43212**
Devesh Chandak	**43213**
Sanya Varghese	**43253**

is a bonafide work carried out by them under the supervision of Dr. S. C. Dharmadhikari and it 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 Examin

Dr. R. Sreemathy
Principal
PICT, Pune

Date :
Place:

ACKNOWLEDGEMENT

We would like to express my deepest appreciation to all those who provided me the possibility to complete this report. A special gratitude we give to our final year project guide, Dr. S. C. Dharmadhikari, whose contribution in stimulating suggestions and encouragement, helped us to coordinate our project especially in writing this report.

Furthermore we would also like to acknowledge with much appreciation the crucial role of the HOD IT Dept. Dr. A. M. Bagade, who gave the permission to use all required resources and the necessary materials to complete the task. A special thanks goes to all our team members, who helped with the research and contribution for the work. Last but not least, we thank Mr. J. K. Kamble, Ms. Sarika Patil and Mrs. R. Chajjad for their timely suggestions and inputs without which the project wouldn't reach to this stage. We would appreciate the guidance given by other supervisors as well as the panels especially in our project presentation that has improved our presentation skills thanks to their comment and advice.

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
4.	GA	Guided Attention
5.	LSTM	Long Short Term Memory
6.	BUTD	Bottom-Up Top-Down
7.	OCR	Optical Character Reader

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 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.

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

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 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.

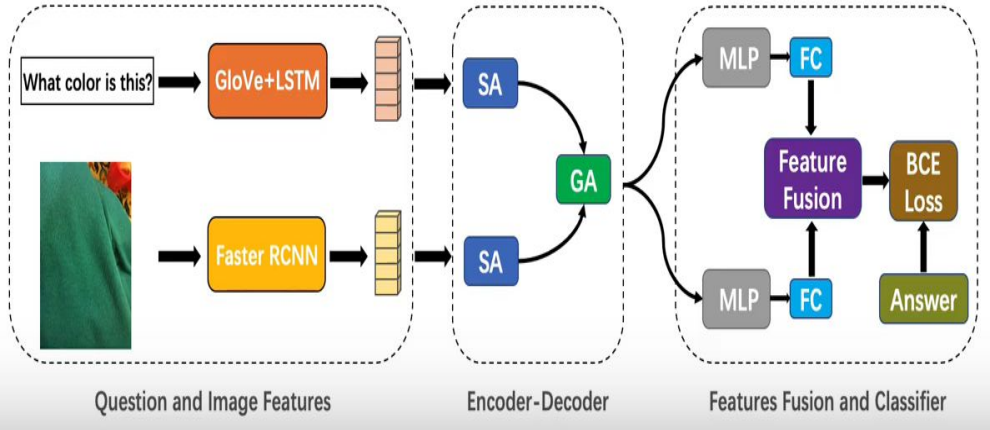


Fig. 1. SUDOKU MODEL

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.

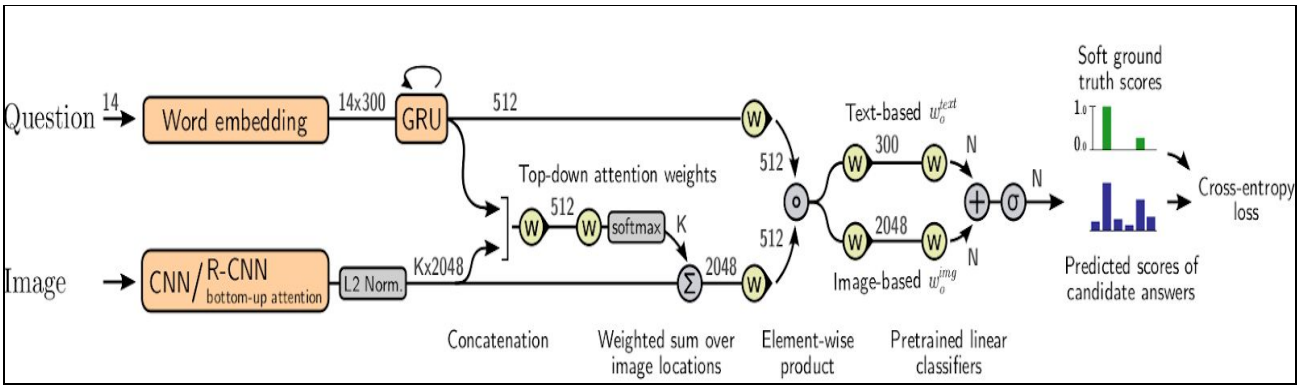


Fig. 2. BOTTOM-UP TOP-DOWN ATTENTION

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 VQA 2018 challenge

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 and Oscar inspired 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.

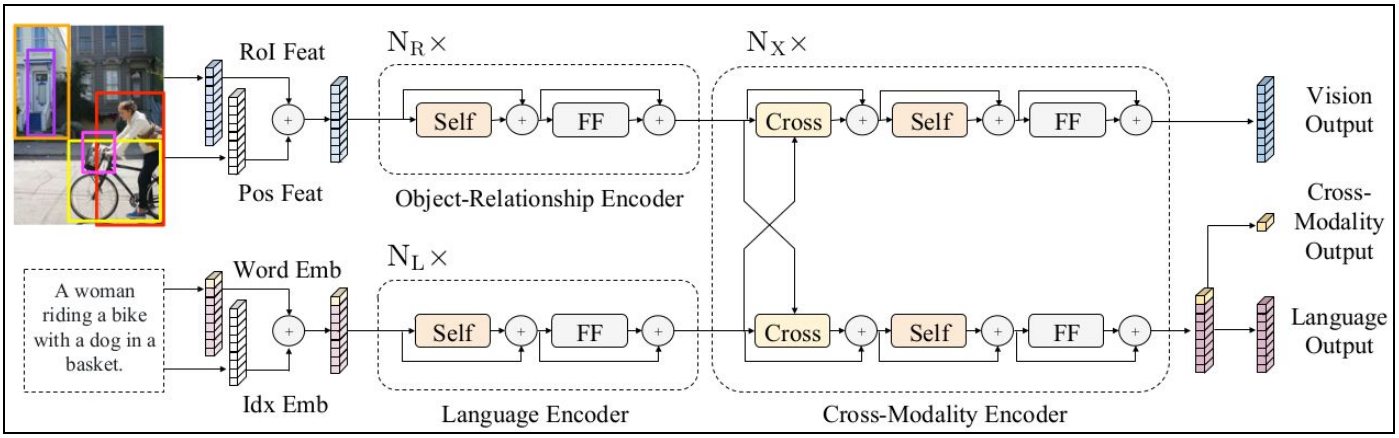


Fig. 3. PROPOSED MODEL

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.

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.

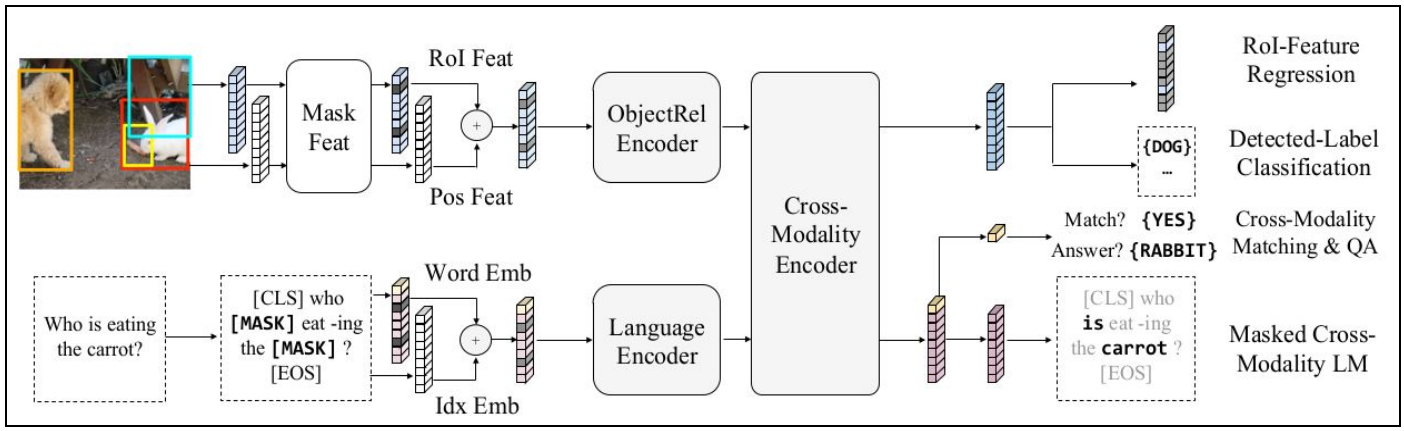


fig. 4. DETAILED PROPOSED MODEL

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 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, additional datasets have been released and many algorithms have been proposed.

In this project we will be focusing on the VizWiz-VQA 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 dataset contains more train-validation images with local obfuscation which prevents privacy disclosure.

The evaluation metric followed:

Accuracy = min(humans that answered/3, 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.

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 like google.

3. IMAGE INPUT: Users should be able to navigate through the app and click images with a minimum number of interactions (clicks).

4. GET OUTPUT: Once the input is taken, the system should process the input and give the appropriate answer to the given question.

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.

2. RELIABILITY ON NETWORK: The product should also work in areas of low or no connectivity. For this, the traditional 3 tier architecture is not appropriate. This type of helpful product should have minimum reliability on external factors.

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/ Kivy
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 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 and ask "Is the signal red?" then our model will evaluate the question and give him the voice output.

3.6. PROJECT PLAN

3.6.1 Project Resources

3.6.2 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

3.6.3 Project Team Role and Responsibilities

3.6.5. PERT Diagram

3.6.6. PERT TABLE

Sr. No.	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
8.	Build a kubernetes cluster in AWS	7 days	Devesh Chandak
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.

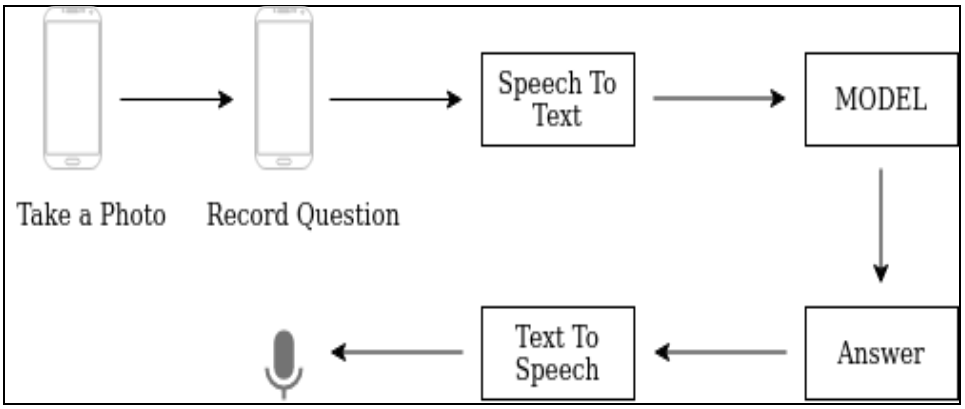
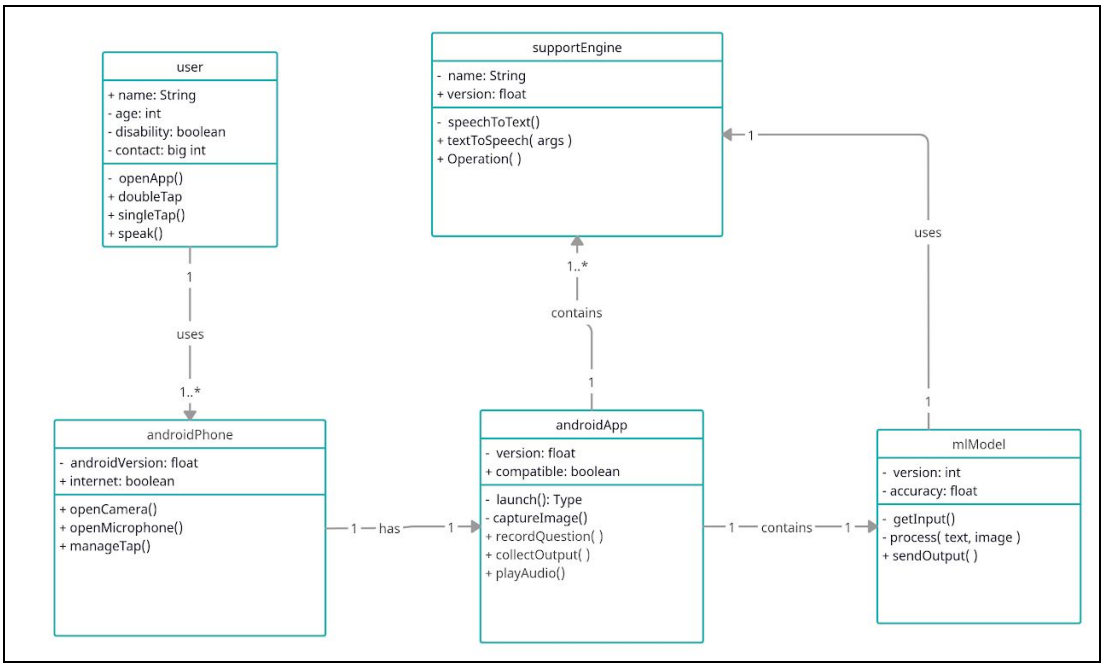


Fig. 5. System Design

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



4.3. BEHAVIOURAL DIAGRAMS

4.3.1 USE CASE DIAGRAM

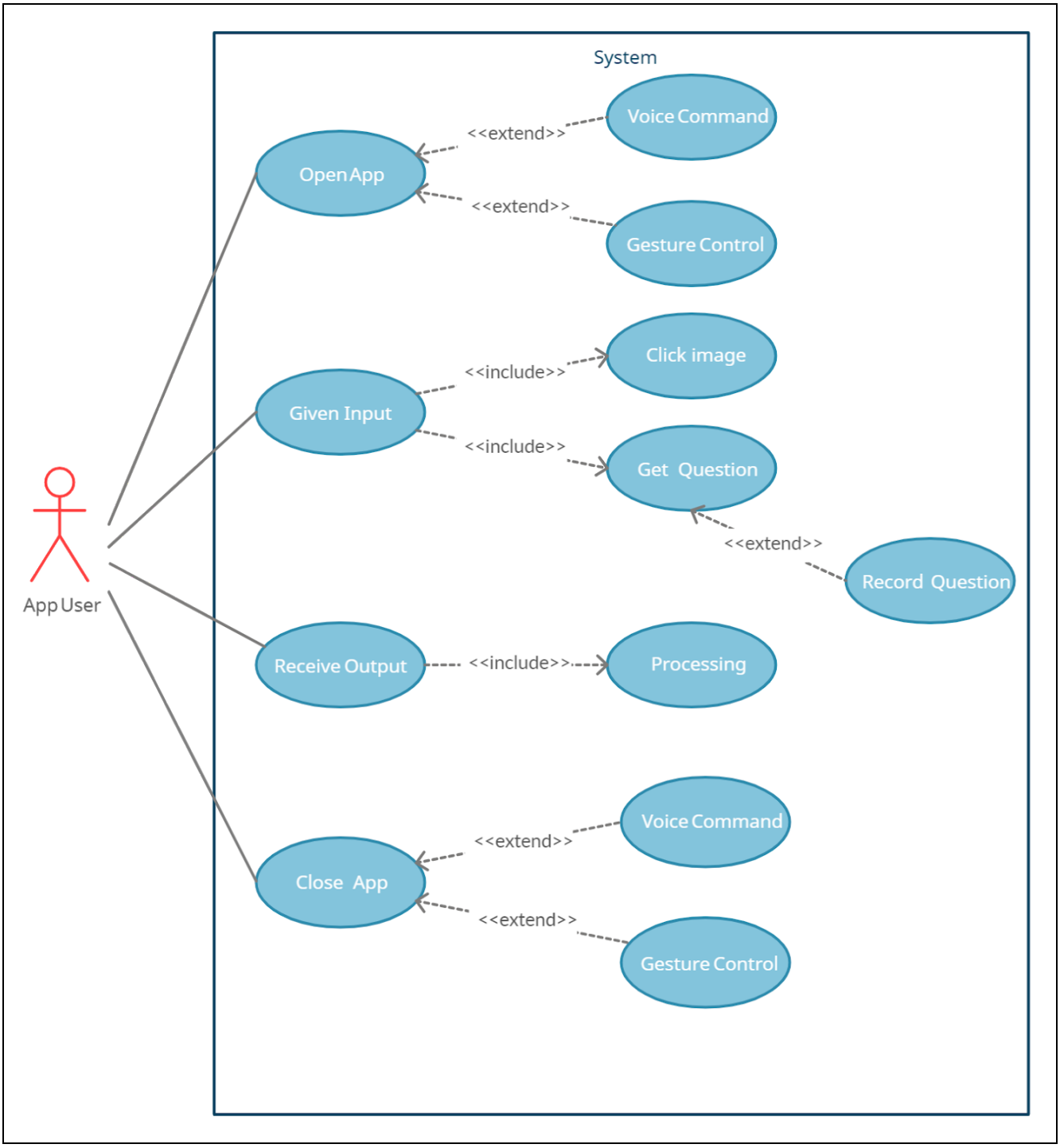


Fig.7. Use Case Diagram

4.3.2. ACTIVITY DIAGRAM

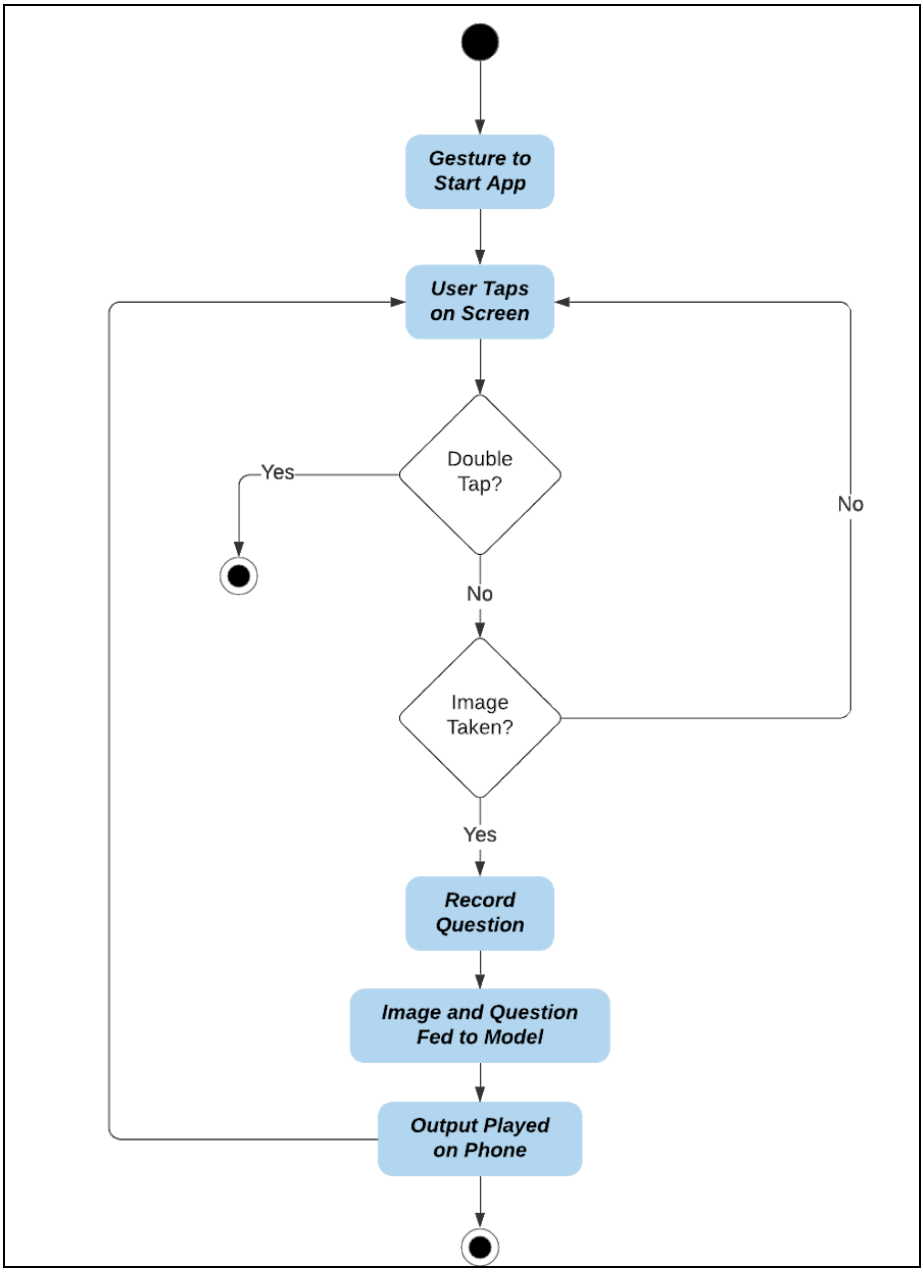


Fig. 8. Activity Diagram

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. 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 second approach is more preferred because of the following reasons.

   a. Compatibility with other python modules
   b. Can be used for cross-platform development

c. Easy to use and efficient

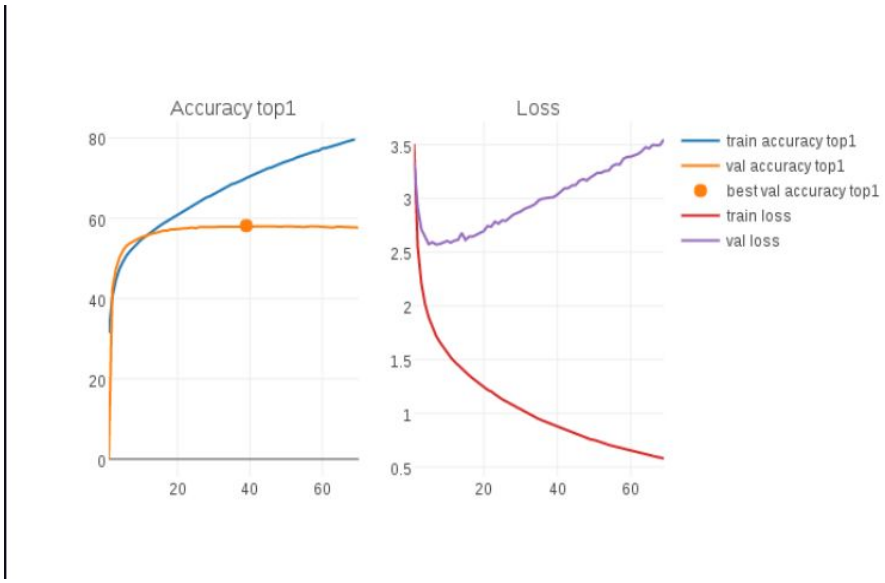
d. Large online community and documentation

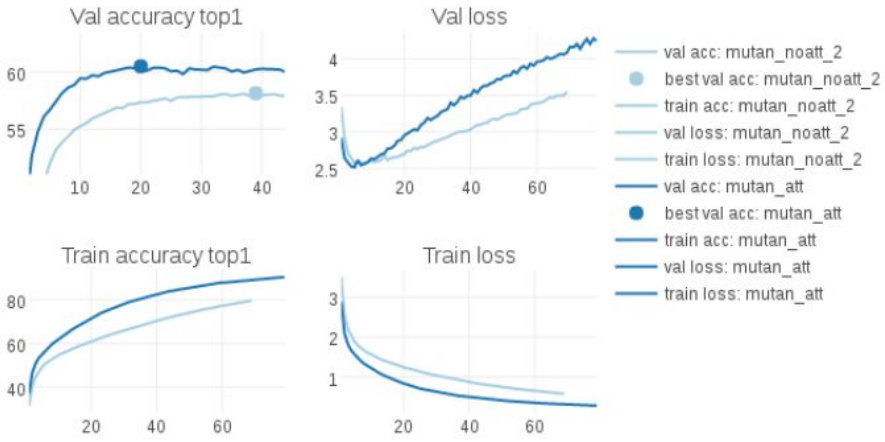
The first phase in implementation of the app is allowing the app to click an image. This can be achieved by providing the whole screen as a single button. Now, if a user clicks anywhere on the screen will capture the image. This is more interactive and helpful for the blind.

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 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 python’s “speech_recognition” module.

CHAPTER 6: RESULTS AND EVALUATION





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 updations 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 improve the architecture 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 and give better results with text based questions.

7.3 Conclusion

The objective of the paper 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.

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

(Academic Year: 2020-21)

Semester - I

Monthly Planning Sheet

Academic Year: 2020-21

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		
Week 5	Requirement Analysis	Completed		
Week 6	Requirement Identification	Completed		
Week 7	Existing methods evaluation	Completed		
Week 8	System analysis and design	Completed		
Week 9	Decide architecture	Completed		
Week 10	Study implementation techniques and tools	Completed		
Week 11	Prototype development	Completed		
Week 12	Prototype development	Completed		

Project Coordinator

Internal Guide