NITTE EDUCATION TRUST | N.M.A.M. INSTITUTE OF TECHNOLOGY
(An Autonomous Institution affiliated to Visvesvaraya Technological University, Belagavi)
Nitte – 574 110, Karnataka, India
(ISO 9001:2015 Certified), Accredited with 'A' Grade by NAAC
☎: 08258 - 281039 – 281263, Fax: 08258 – 281265
Department of Computer Science and Engineering
B.E. CSE Program Accredited by NBA, New Delhi from 1-7-2018 to 30-6-2021

Panel Report

Project Title:

"A Visual assistant chatbot for visually impaired"

Submitted by:

Pramukha R N
4NM16CS101
pramukhrn@outlook.com

Rahul D Shetty
4NM16CS111
35rahuldshetty@gmail.com

Shetty Yashas Shashidhar
4NM16CS137
yashasshetty111@gmail.com

Submitted on: **14th September 2019**

Introduction

A key challenge faced by people who are partially or completely blind is perception and navigation of the environment that they are not accustomed to. Travelling to an unfamiliar place or merely walking down a crowded street or can be a challenge. As a consequence, many people with impaired vision travel with a friendly person or a family member while navigating an alien environment. This proxy person helps them in their navigation, describes the external environment to them and thus helping them in their external cognition. We propose a chatbot assistant which plays the role of this person. Our "Visual assistant chatbot" acts as their artificial eye describing and summarizing the external environment in real-time.

The "Visual assistant chatbot" can be integrated into an android app which can be launched easily with a single click from the phone or can be launched through voice command. The person might ask summary or questions like *"What do I see in front of me?"*, *"What food is on the plate?"*, *"What are the people doing?"* and get answers in real-time. The chatbot also describes any changes in the current scene and also notifies the user about any warnings. This chatbot can be used for a wide variety of uses from reading the price of an item to helping in setting the time of a washing machine, describing the type of beverage in the cup to identifying what the giraffe in a zoo is eating. This chatbot leverages the power of computer vision and natural language processing to liberate any visual burned that visually impaired people may encounter and makes their everyday life easier.

Objectives

1. Design and implement a Visual Question Answering model which takes an image and a text question as input and outputs an answer.
2. Create an interface for speech to text interconversion and prompts for the user.
3. Implement scene summarization to provide a quick summary of the current scene.
4. Implement scene change detection and warning indicators for potential dangers.

Methodology

The photographs captured by visually impaired are not the same as ones captured by visually sound people. The images might be out of focus, might contain partial images etc. Hence, we need a task-specific dataset for this project. The dataset that we use is the VizWiz dataset[1]. This dataset contains visual questions asked by visually impaired people who were seeking an answer to their daily questions. A person asked a visual question by taking a picture and then recording a spoken question. The image and spoken question were collected anonymously. Figure 1 shows the word distribution among all the questions. Most of the questions contain three or fewer words. These questions have to be processed carefully as to avoid any loss of information as these might lead to a drastic decrease in performance. The VizWiz dataset is a difficult dataset to model when compared to other dataset in the literature because of poor lighting, poor framing and image blur. The questions are also directly translated to speech and

the model must also account for the translation errors. The answers to the questions were collected from crowd workers.

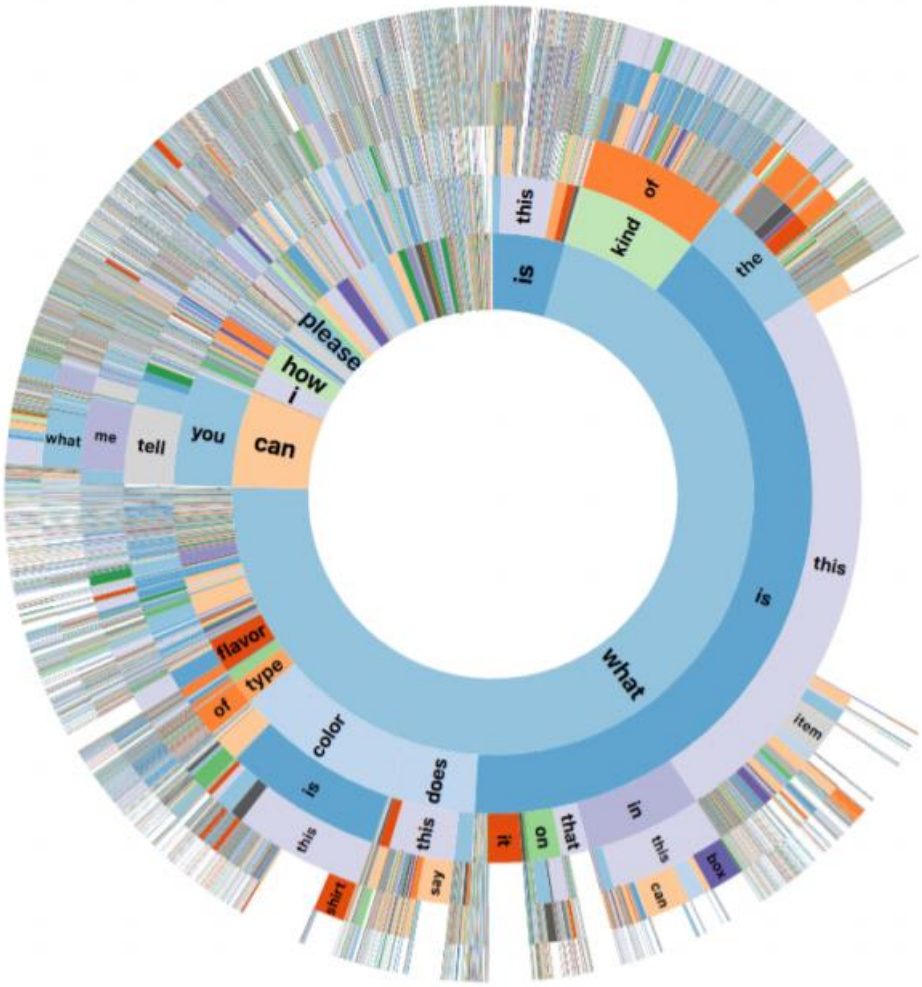

Figure 1. Distribution of the first six words for all questions

A blind person can ask a question to the chatbot through speech. This speech is converted into text using Google's Text-to-Speech API. The natural language question is then fed into model along with an image. The resulting answer is then again converted into speech using the same API and read back to the user. In the case of warning system, the output of VQA model is analyzed to inform the user about any potential dangers. For text summarizer, we implement a scene summarizer model and read out a short description of the scene whenever asked by the user. Scene change model is also implemented to alert user in case of any scene change. The models can also be hosted in cloud and interface with app can be provided using high bandwidth channels for optimal performance.

Expected Output

The evaluation metric which is given in equation 1 is taken from the VQA challenge. It is robust to inter-human variability. Each question is associated with 10 answers. If the answer provided

by the model is equal to at least 3 of those answers. Then the VQA model is said to have 100% accuracy.

$$\text{accuracy} = \min\left(\frac{\#\text{ humans that provided that answer}}{3}, 1\right)$$

Equation 1. Evaluation metric

In the case of text summarization model, an intrinsic, content-based, n-gram evaluation is used. The model is also evaluated extrinsically on the basis of its task-specific summarization capabilities.

With enough bandwidth connection, utilizing the power of the cloud, the delay between query and answer is expected to decrease drastically. All the models are implemented in a single application which works seamlessly. The user can launch the app with just a click of a button or through google assistant. The app also has buttons with large fonts to assist people with partial vision. In the case of extreme blindness, the user can access the whole functionality of the app through their voice.

References

[1] D. Gurari *et al.*, "VizWiz Grand Challenge: Answering Visual Questions from Blind People," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018.