JSS MAHAVIDYAPEETHA

JSS Science and Technology University

"AI enabled Vision Assistance system"

A technical project report submitted in partial fulfillment of the award of the degree of

BACHELOR OF ENGINEERING

IN

ELECTRONICS AND COMMUNICATION ENGINEERING

BY

Ashwin Vijayakumar Bhandari	**01JST20EC013**
B R Manoj	**01JST20EC015**
K Preksha Bhat	**01JST20EC042**
Krishnakumar Balachandra Bhat	**01JST20EC045**

Under the guidance of
Prof. Halesh M R
Assistant Professor
Department of Electronics and Communication Engineering
SJCE, JSS S&TU, Mysuru.

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING
2023-2024

JSS MAHAVIDYAPEETHA

JSS Science and Technology University

CERTIFICATE

This is to certify that the work entitled "AI enabled Vision Assistance system" is a Bonafide work carried out by Ashwin Vijayakumar Bhandari, B R Manoj, K Preksha Bhat, Krishnakumar Balachandra Bhat in partial fulfilment of the award of the degree of Bachelor of Engineering in Electronics and Communication Engineering for the award of Bachelor of Engineering by JSS Science and Technology University, Mysuru, during the year 2023-2024. The project report has been approved as it satisfies the academic requirements in respect to project work prescribed for the Bachelor of Engineering degree in Electronics and Communication Engineering.

Under the guidance of

Prof. Halesh M R
Assistant Professor,
Department of ECE, SJCE,
JSS STU, Mysuru.

Head of the Department

Dr. U.B. Mahadevaswamy
Professor and Head,
Department of ECE, SJCE,
JSS STU, Mysuru.

Examiners:

1. .....

2. .....

Visual impairment

ORIGINALITY REPORT

10%	5%	5%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to JSS Science & Technology University, Mysore Student Paper	1%
2	Viktar Atliha. "Improving image captioning methods using machine learning approaches", Vilnius Gediminas Technical University, 2023 Publication	1%
3	Nilesh Deotale, Shubham Raut, Nirmit Patil, Vedashree Patil, Priyal Bari. "Smart Assistive Stick for Visually Impaired People using YOLOv8 Algorithm", Research Square Platform LLC, 2024 Publication	<1%
4	www.atlantikelektronik.com Internet Source	<1%
5	Submitted to Manchester Metropolitan University Student Paper	<1%
6	"Computer Vision – ECCV 2022", Springer Science and Business Media LLC, 2022	<1%

	Publication	
7	researchonline.federation.edu.au Internet Source	<1 %
8	Submitted to Manipal University Jaipur Online Student Paper	<1 %
9	Submitted to University at Buffalo Student Paper	<1 %
10	www.freepatentsonline.com Internet Source	<1 %
11	www.politesi.polimi.it Internet Source	<1 %
12	Shreyansh Chordia, Yogini Pawar, Saurabh Kulkarni, Utkarsha Toradmal, Shraddha Suratkar. "Chapter 52 Attention Is All You Need toTell: Transformer-Based Image Captioning", Springer Science and Business Media LLC, 2022 Publication	<1 %
13	Submitted to University of Glasgow Student Paper	<1 %
14	Submitted to The Robert Gordon University Student Paper	<1 %
15	Yassine Bouteraa. "Design and Development of a Wearable Assistive Device Integrating a Fuzzy Decision Support System for Blind and	<1 %

DECLARATION

We do hereby declare that the project titled “AI enabled Vision Assistance system” is carried out by the project group, under the guidance of Guide Prof. Halesh M R, Assistant Professor, Department of Electronics and Communication Engineering, JSS Science and Technology University, Mysuru, in partial fulfilment of requirement for the award of Bachelor of Engineering by JSS Science and Technology University, Mysore, during the year 2023-2024.

We also declare that we have not submitted this dissertation to any other university for the award of any degree or diploma course.

Date:May 27, 2024

Place: Mysore

Ashwin Vijayakumar Bhandari

B R Manoj

K Preksha Bhat

Krishnakumar Balachandra Bhat

ACKNOWLEDGEMENT

We would like to extend our sincere and heartfelt obligation towards all the personages who have helped us in this endeavour.

Firstly we would like to express my sincere gratitude to JSS Science and Technology University, Mysuru which provided us a great platform and opportunity to carry out this project work.

We are very much obliged to Dr. C Nataraju, Principal and Dean, JSS Science and Technology University, Mysuru. We would like to profoundly thank Dr. U B Mahadevaswamy, Professor and Head of the Department, Electronics and Communication Engineering, JSS Science and Technology University, Mysuru for his guidance and assistance.

We owe our special thanks to Prof. Halesh M R, Assistant Professor, Department of Electronics and Communication Engineering, JSS Science and Technology University, Mysuru for his timely guidance, support and encouragement. He has made learning highly engaging with his practical advice, unceasing ideas and technical inputs which has helped us carry out the project seamlessly.

Finally, we would like to express our heartfelt gratitude to all the staff members of Electronics and Communication Engineering, JSS Science and Technology University, Mysuru.

ABSTRACT

This work explores the potential of artificial intelligence (AI) for vision assistance systems. These systems leverage computer vision and machine learning techniques to analyze visual data captured through cameras and provide assistance to users using large language model. The primary focus is on empowering individuals with visual impairments by enabling environment recognition and information delivery. This can be achieved through voice prompts, audio cues.Key features of these systems include the provision of voice prompts and audio cues, which serve as indispensable aids in navigating various environments. Through the judicious utilization of AI technologies, individuals with visual impairments can effortlessly traverse both indoor and outdoor spaces, overcoming obstacles with unparalleled ease. this work underscores the transformative potential of artificial intelligence in the realm of vision assistance systems. By harnessing the capabilities of AI, these systems have the capacity to revolutionize the lives of individuals with visual impairments, offering them unprecedented levels of autonomy and empowerment. heralding a new era of innovation and accessibility.

Keywords : Q-Former Encoder¡BLIP model, Bootstrapping Language-Image Pre-training, Image captioning, Reinforcement learning,Large language models (LLMs), Cross-modal encoder,Vision Encoder, Text Encoder, Decoders, Fine-tune,Gradient descent, Image-Text Contrastive Learning (ITC), Image-Text Matching (ITM), Masked Language Modeling (MLM), Cross-Entropy Loss.

Contents

1 Introduction **1**

1.1 Preamble 1

1.2 Motivation 2

1.3 Problem Statement 3

1.4 Block Diagram 3

1.5 Objectives 4

1.6 Organization of the report 4

2 Literature Review with Gap in the Literature **5**

2.1 Previous research 5

2.2 Summary of Literature Review 8

2.3 Summary of the chapter 9

3 Present Work Carried Out **10**

3.1 BLIP Model 10

3.1.1 Overview of BLIP Model 11

3.1.2 Image-Text Matching (ITM) 12

3.1.3 Fine-Tuning BLIP Model 12

3.1.4 Combining Image Caption and OCR Text 14

3.1.5 Processing Text with LLama Model 14

3.1.6 Text-to-Speech Conversion 14

3.2 Hardware Requirements 15

3.2.1 Raspberry pi Zero 2W 15

3.2.2 Adafruit I2S 3W Class D Amplifier 16

3.2.3 Speaker 17
3.2.4 Rechargeable Battery 17
3.2.5 Push Buttons 19
3.2.6 xcluma 5V Lithium Battery Charger Step Up Module . 19
3.3 Software Requirements 20
3.3.1 Python 20
3.3.2 Integrated Development IDE 21
3.3.3 Putty 22
3.3.4 Python Libraries 23
3.3.5 Fine tuned model 25
3.4 Novelty of the work 26
3.5 Summary of the chapter 26

4 Results and Discussions **28**
4.1 Image captioning and Large Language Model 29
4.2 Conversion of text to audio file 31
4.3 Audio Playback 32
4.4 Summary of the chapter 32

5 Conclusion **34**
5.1 Conclusion 34
5.2 Advantages and Disadvantages 35
5.2.1 Advantages 35
5.2.2 Disadvantages 36
5.3 Future scope 36

References **38**

List of Figures

1.1 Generalized Block Diagram 3

3.1 Blip Model Block Diagram 11
3.2 Learning framework of BLIP Model 12
3.3 Raspberry pi Zero 2W 16
3.4 Adafruit I2S 3W Module 17
3.5 Speaker 18
3.6 Lithium Rechargeable battery 18
3.7 Tactile Push Buttons 19
3.8 Charging Module 20
3.9 Python Logo 20
3.10 VScode Logo 21
3.11 Thonny Logo 22
3.12 Putty Logo 22

4.1 Prototype of our work 28
4.2 Person wearing the Vision Assistant Glasses 29
4.3 Ocr-generated text as input 30
4.4 Output of BLIP Model 30
4.5 Output of LLama Model of OCR 30
4.6 Output of LLama Model of Image Description 31
4.7 Text to Audio Conversion Output 31
4.8 Audio File Recived and Playback 32

List of Tables

2.1 Summary of Papers 8

Chapter 1

Introduction

1.1 Preamble

Visual impairment significantly affects an individual's ability to interact with their environment, posing daily challenges in navigation and understanding visual information. This work aims to develop an AI-powered vision assistant system designed to provide real-time descriptive audio feedback from visual inputs, thereby enhancing the independence and quality of life for visually impaired individuals.

The system captures images using a compact and efficient device, processes these images to generate descriptive captions and extract textual information, and then synthesizes this data into audible descriptions. By converting visual data into comprehensible audio feedback, the system enables users to better understand and navigate their surroundings.

The core functionality of the model involves capturing an image, processing the visual and textual data, and delivering an audio response. Upon capturing an image, the data is sent to a server where advanced AI models process the image to generate a detailed description and identify any text within the image. This combined information is then refined and converted into speech, which is transmitted back to the user for playback. This vision assistant system demonstrates the potential of artificial intelligence to

address real-world challenges faced by visually impaired individuals. By providing detailed and accurate audio descriptions, the system empowers users to navigate and interact with their environment more effectively, thereby promoting greater autonomy and confidence.

The development of this work highlights the transformative impact of AI in creating accessible solutions that improve the quality of life for individuals with disabilities. Future enhancements may include improving the accuracy and responsiveness of the system, as well as expanding its capabilities to provide more complex scene understanding and interaction features. This AI-powered vision assistant represents a significant step towards leveraging technology to create inclusive and supportive tools for the visually impaired community.

1.2 Motivation

There are 39 million blind people, 246 million of whom are visually impaired and 285 million around the world are visually impaired and have approx 4.95 million blind, 35 million blind, and 0.24 million blind children in India.only environment for BRAILLE literacy for the blind that the sighted can read and cannot interpret and limited to those who learn. This limits freedom for a person with low vision. AI-powered vision boost aid systems based on the goal of making more inclusive and Accessible environments allow people with visual impairments to live a more independent and prosperous life by using technology. By addressing the specific needs and challenges faced by blind individuals, our work aims to enhance their independence, autonomy, and overall quality of life. We believe that by harnessing the power of technology, we can create a more inclusive and accessible world for individuals with visual impairments, ensuring that they can participate fully in society and pursue their goals and aspirations without limitations.

1.3 Problem Statement

This work aims to develop a model that can generate natural language descriptions from images. The model has the potential to be used in a variety of applications, such as helping blind people get around or providing real-time traffic updates.

1.4 Block Diagram

The figure 1.1 shows an AI-powered vision assistance system using a Raspberry Pi Zero 2W. It starts with a camera module that captures visual information from the user's surroundings. This data is then processed by Raspberry Pi Zero 2W, which is the core of the model.

BLIP, an AI model optimized for use with microcontrollers, analyzes captured images to gain meaningful insights about the environment. Finally, the Google Text-to-Speech (gTTS) library converts this data into audio format. Synthesized audio feedback is sent through the speaker, providing the user with a sense of surround sound. Essentially, this model translates visual information into auditory units, allowing blind people to navigate their environment independently.

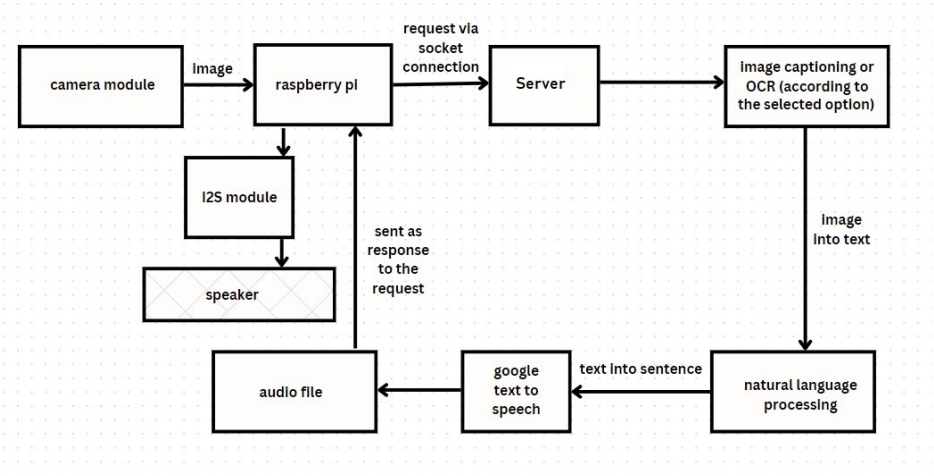

Figure 1.1: Generalized Block Diagram

1.5 Objectives

1. To generate efficient and coherent natural language descriptions of captured images.

2. To integrate the NLP system with the image recognition component to provide a unified, coherent experience for users.

3. To create a user-friendly interface that allows visually impaired users to prompt the system for image descriptions

1.6 Organization of the report

The report is organized into several key sections to comprehensively document our research and findings. It begins a brief introduction about the vission assistant for blind peoples. The motivation section discusses the rationale for this choice, while the problem statement provides an overview of the current problem. A generalized block diagram is then provided along with several objectives and an overview of the report's organization. Following this, a literature review section presents previous research, identifies gaps in the literature, and summarizes the findings. The subsequent section details the present work carried out, including Hardware and software requirements and a summary of our methodology. Results and discussions follow, where we present all the output that are obtained in different models. The report concludes with a section summa- rizing our conclusions, discussing the advantages and disadvantages of our work, outlining future research opportunities, and providing a list of references for further reading.

Chapter 2

Literature Review with Gap in the Literature

2.1 Previous research

This paper proposes an intelligent system designed for blind people to operate independently. The model includes various functions to improve mobility and safety, including speech recognition for intuitive control, traffic light recognition for safe crossing, laser to avoid obstacles to navigate in mountainous conditions. [1].

This paper proposes the concept of smart glasses designed to improve the lives and productivity of people with disabilities. This model uses a Raspberry Pi 2 camera to capture visual data and translate it into a format that can be used by audio description or text display. This innovative approach has the potential to empower blind people by improving their ability to navigate and interact with the world around them. [2].

This study explores the use of CNNs for speech recognition in industry. Although effective, the accuracy of the recognition of frequently spoken command words in noisy environments with varying tones needs improvement. This is critical for reliable human-machine interaction in industrial settings. [3].

This study proposes a new vision model using deep learning for real-time image registration. Powered by a combination of Convolutional Neural Networks (CNN) and Short-Term Memory (LSTM), this model aims to create natural sound recordings for live images stored on the device. By training on a large database, the model strives to achieve high accuracy in describing visual content, providing a valuable tool for users who may need a description of the surrounding sound [4].

This paper proposes an intelligent navigation device for the disabled. It uses deep image sensors and multi-sensor fusion algorithms to navigate in complex indoor environments. The device appeals to blind and partially sighted users, providing voice guidance for augmented and augmented reality (AR), allowing them to navigate with greater independence. [5].

The paper introduces an end-to-end trainable neural network with a recurrent attention model over an external memory, suitable for tasks like question answering and language modeling. It demonstrates competitive results and suggests extending to handle multiple layers for improved training efficiency [6].

The paper proposes Transformer, a new neural network architecture for linguistic problems. This model achieves advanced results by relying only on "noise" mechanisms, iterative or convolutional layer mining. This greatly simplifies the architecture, allowing faster computation times and better parallelization. [7].

The paper introduces a neural conversational model that predicts next sentences in conversations, enabling end-to-end training with minimal handcrafted rules [8].

The paper introduces a simple, efficient, and real-world solution for multilingual Neural Machine Translation, enhancing its effectiveness and enabling zero-shot translation between language pairs [9].

The paper presents a novel approach to neural machine translation, enhancing its performance by learning to align and translate jointly, overcoming the limitations of traditional methods [10].

The work utilizes AI, Machine Learning, and text recognition to assist visually impaired individuals with an Android mobile app, enhancing their engagement with the world [11] [12].

The paper proposes a wearable navigation system for blind and visually impaired people, using sensors, real-time processing boards, fuzzy logic, and a mixed voice-haptic interface for safety orientation [13].

The proposed work uses automated voice navigation for Visually Impaired Persons (VIP) to visualize their surroundings and share their location with family members. It uses MobileNet architecture and has shown satisfactory results in six pilot studies [14]. This paper presents You Only Look Once (YOLO), a CNN representative that innovates object detection with a simple and efficient method [15].

This article introduces a new assistive device for the disabled. Using deep learning and stereo cameras, the model goes beyond providing location information to users. It empowers them by detecting and classifying real-world obstacles, providing critical information about their environment for safer and more autonomous navigation. [16].

This paper introduces YOLO, a deep learning object detection algorithm built on a convolutional neural network (CNN). Trained on large databases like COCO, YOLO excels in real-time detection. But the authors learned how to customize YOLO for unique visual problems using their own data. This opens the door to applying YOLO to specific needs and environments. [17] [18].

2.2 Summary of Literature Review

Paper Name	Author(s)	Technology Used	Advantages	Disadvantages
Ibgs: A wearable smart system to assist visually challenged	K. Xia, X. Li, H. Liu, M. Zhou, and K. Zhu	Transformers	It has a good Accuracy	It has too much delay
Building a voice-based image caption generator with deep learning	M. Anu, S. Divya, et al.	Deep Learning	Image caption generation, Voice-based	It does't have NLP in Voice Generation
End-to-end memory networks	S. Sukhbaatar, J. Weston, R. Fergus, et al.	End-to-end memory networks	Memory integration, Neural Information Processing	It does not have basic attention mechanism
Google's multilingual neural machine translation system	M. Johnson, M. Schuster, Q. V. Le, et al.	Neural Machine Translation	Zero-shot translation, Multilingual support	It does not have greater accuracy in Individual Languages
Attention is all you need to tell: Transformer-based image captioning	S. Chordia, Y. Pawar, S. Kulkarni, U. Toradmal, and S. Suratkar	Transformer-based Image Captioning	Attention mechanism	It does not have greater accuracy in memory
A neural conversational model	O. Vinyals and Q. Le	Neural Network	Conversational model	It does not have Time specifief sequence to sequence
AI enabled Vision Assistance system	Ashwin Vijayakumar Bhandari, Krishnakumar Balachandra Bhat, k Preksha Bhat, BR Manoj	NLP, Cloud Computing	Human understandable audio output	Not specified right now

Table 2.1: Summary of Papers

The table 2.1 summarizes the main findings of various studies on assistive technologies for the blind. Each entry details the paper's title, author, underlying technology (such as Transformers or deep learning), strengths (such as high accuracy or multilingual support), and weaknesses (such as processing

delays or limitations). It offers a quick comparison, allowing you to see how different technologies solve the challenges faced by blind users.

2.3 Summary of the chapter

This chapter reviews existing research on assistive technology for the visually impaired. Here is a summary:

Previous Research Efforts: This chapter reviews a variety of research on smart glasses, voice recognition, recording systems, navigation devices, and object detection for the blind. This work uses technologies such as convolutional neural networks (CNN), transformers, and deep learning.

Established Strengths: This chapter highlights advances in areas such as image captioning, obstacle detection, and navigation for users with disabilities. Some studies achieve high accuracy and include real-time processing.

Limitations and Exclusions: This chapter sets out limitations in certain areas. For example, some voice recognition systems lack accuracy, and some recording systems may not integrate Natural Language Processing (NLP) to produce natural sounds. In addition, some studies do not determine functions such as temporal discourse or memory integration.

Chapter Takeaway: Although significant progress has been made, there is still room to develop comprehensive assistive technologies that combine different functions to create a safer and more user-friendly experience for seen disability

Chapter 3

Present Work Carried Out

This work explores the development of an AI-enabled vision assistance system using the Raspberry Pi Zero 2W. The model aims to empower users, especially those with low vision, by providing real-time information about their environment.

The main components include the Raspberry Pi Zero 2W, a single board computer that acts as a processing unit, and an I2S Digital-to-Analog Converter (DAC) to convert digital audio signals into audio output from speakers. This work adopts the BLIP machine learning model designed specifically for microcontrollers such as the Raspberry Pi. This model is designed to interpret visual information and translate it into meaningful information.

3.1 BLIP Model

The BLIP model, short for Bootstrapping Language-Image Pre-training, is a cutting-edge model used for unified vision-language understanding and generation. It aims to enhance image captioning by providing detailed and comprehensive descriptions of images. The model leverages a three-stage training procedure involving pre-training on unsupervised data, fine-tuning on supervised data, and reinforcement learning techniques to maximize performance. BLIP is known for its ability to generate high-quality captions by

incorporating large language models and vision-language models like CLIP and BLIP2-ITM.

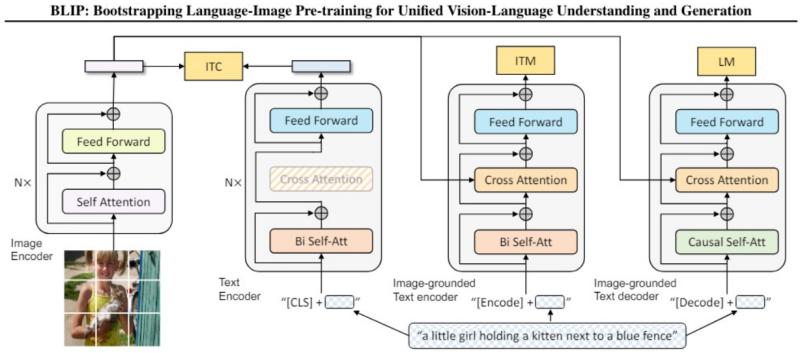

Figure 3.1: Blip Model Block Diagram

On the other hand, the LLama 3 model, which is not extensively detailed in the provided sources, is another model used for image captioning. It is likely designed to complement existing models like BLIP in generating enriched visual descriptions. Fine-tuning the output of BLIP and LLama 3 models involves adjusting the models' parameters and training processes to optimize their performance for specific tasks or datasets.

3.1.1 Overview of BLIP Model

The BLIP model is built upon the following key components:

1. Image Encoder: Typically a convolutional neural network (e.g., ResNet, Vision Transformer) that extracts visual features from images.

2. Text Encoder: A transformer-based model (e.g., BERT) that encodes text descriptions into contextual embeddings.

3. Cross-modal Encoder: Another transformer that fuses visual and textual features.

4. Pre-training Objectives: The model is trained with multiple objectives like image-text contrastive loss, image-grounded text generation loss, and masked language modeling.

3.1.2 Image-Text Matching (ITM)

This objective involves a binary classification task where the model predicts whether a given image-text pair matches. The ITM loss can be defined as:

$$\mathcal{L}\text{ITM} = -\frac{1}{N}\sum i = 1^{N} \left(y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)\right)$$

where:

- y_i is the ground truth label indicating if the image-text pair (i) matches.
- $\hat{y}_i$ is the predicted probability of the pair matching.

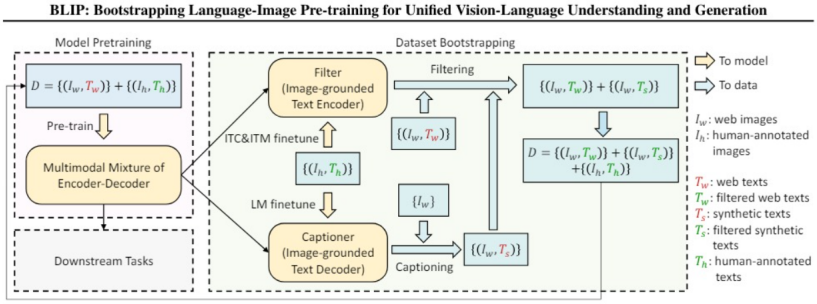

Figure 3.2: Learning framework of BLIP Model

3.1.3 Fine-Tuning BLIP Model

1. **Dataset Preparation**: Prepare a dataset of images with corresponding captions.
2. **Model Initialization**: Load the pre-trained BLIP model.
3. **Training Configuration**: Define training parameters, including the learning rate, batch size, and number of epochs.

4. **Loss Function**: Use Cross-Entropy Loss for caption generation.

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_{\text{caption}}(\hat{y}_i, y_i) \tag{3.1}$$

Fine-tuning a pre-trained BLIP model involves updating its parameters using a custom dataset. The loss function used in fine-tuning can be represented as:

where:

- N is the number of training samples.
- $\mathcal{L}_{\text{caption}}$ is the loss function for caption generation (e.g., Cross-Entropy Loss).
- $\hat{y}_i$ is the predicted caption.
- y_i is the ground truth caption.

5. **Optimization**: Update the model parameters using gradient descent. The fine-tuning process can be expressed mathematically as:

$$\theta = \theta - \eta \nabla_\theta \mathcal{L} \tag{3.2}$$

where:

- θ are the model parameters.
- η is the learning rate.
- $\nabla_\theta \mathcal{L}$ is the gradient of the loss function with respect to the parameters.

The fine-tuning process involves training the model specifically for the task of generating image captions, leveraging a dataset of images and their corresponding captions. The loss function used in this phase is typically Cross-Entropy Loss, which measures the difference between the predicted caption and the ground truth caption. The equation for Cross-Entropy Loss is:

$$\mathcal{L}\text{caption} = -\frac{1}{N} \sum i = 1^N \sum_{t=1}^{T} \log P(y_{i,t} | \mathbf{v}i, \mathbf{y}i, < t)$$

where:

- N is the number of training samples.
- T is the length of the caption.
- $y_{i,t}$ is the t-th token of the ground truth caption for the i-th image.
- $\mathbf{v}_i$ is the visual representation of the i-th image.
- $\mathbf{y}_{i,<t}$ are the tokens generated by the model up to the t-th position.

3.1.4 Combining Image Caption and OCR Text

To combine the image caption and OCR text, we concatenate the two text sequences:

$$\text{Combined Text} = \text{Caption} \oplus \text{OCR Text} \tag{3.3}$$

where $\oplus$ denotes concatenation.

3.1.5 Processing Text with LLama Model

The refined text is generated using the LLama model. Given an input prompt x, the model generates an output y:

$$y = \text{LLama}(x) \tag{3.4}$$

The prompt combines the image caption and OCR text:

$$x = \text{"This is image captioning output with OCR in it. Describe it for a blind person within 19 v} \tag{3.5}$$

3.1.6 Text-to-Speech Conversion

The gTTS library converts the refined text y into speech:

$$\text{Audio} = \text{gTTS}(y) \tag{3.6}$$

Image-Text Generation:

$$p(T|I) = \prod_{t=1}^{T} p(T_t|I, T_{<t}; \theta) \tag{3.7}$$

.

The text generated from blip model is further improved by using LLama Model which was developed by Meta. The work uses the Google Text-to-Speech (gTTS) library to provide information to users. This library converts text generated by the LLama model into audible speech, enabling user-to-system interaction through voice.

3.2 Hardware Requirements

In bringing our work to fruition, we carefully selected a range of hardware components to serve as the foundation for our model. These components were chosen based on their performance, compatibility, and suitability for the intended application.

3.2.1 Raspberry pi Zero 2W

The Raspberry Pi Zero 2 W packs a punch in a small credit card-sized package. This single board computer is the predecessor of the Raspberry Pi Zero W. It features a quad-core 64-bit ARM Cortex-A53 CPU running at 1GHz, delivering a 40% improvement in single-thread performance and five times the multi-core performance compared to the original Zero.

Despite its compact size, the Raspberry Pi Zero 2 W does not compromise on performance. Equipped with 512MB LPDDR2 SDRAM, it allows you to run various operating systems and applications smoothly. Built-in Wi-Fi and Bluetooth 4.2 connectivity enable wireless communication, opening the door to projects that interact with the Internet or connect to Bluetooth devices. To connect peripherals and sensors, the Raspberry Pi Zero 2 W offers a 40-pin GPIO (General Purpose Input / Output) header. This allows you to

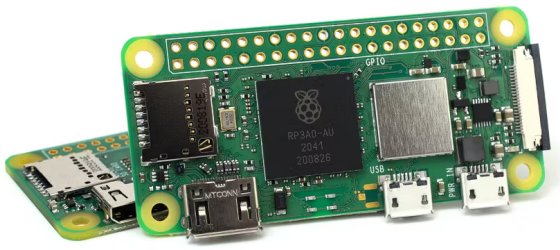

Figure 3.3: Raspberry pi Zero 2W

interface with various electronic components, making it ideal for building robots, automating tasks, or creating custom electronic projects. It also has a micro HDMI port for connecting to a monitor and a micro USB port for power supply.

The Raspberry Pi Zero 2 W's low power consumption makes it perfect for battery-powered projects. In addition, its microSD card allows you to load your operating system and save data. With its capabilities, impressive processing power, and rich feature set, the Raspberry Pi Zero 2 W is an attractive choice for hobbyists, educators, and makers looking for a versatile and portable platform for their projects.

3.2.2 Adafruit I2S 3W Class D Amplifier

Adafruit I2S 3W Amplifier Breakout - MAX98357A is a one-stop shop for adding amplified sound to your projects. This fifty little board packs a surprising punch, delivering 3.2 watts of power to a 4 ohm speaker (with a 5V power supply). But its secret weapon is the use of I2S (IC-IC Audio), a digital audio protocol. Unlike conventional amplifiers that require a separate DAC (Digital-to-Analog Converter) chip, the onboard MAX98357A chip manages digital-to-analog conversion and amplification. This amplifier is perfect for applications where space is limited. It operates over a wide voltage range (2.7V to 5.5V) and is ideal for battery-powered projects. In addition, ther-

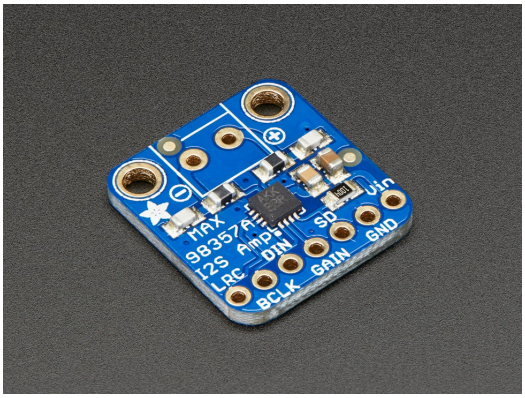

Figure 3.4: Adafruit I2S 3W Module

mal and overcurrent protection features protect items from overheating or overloading. Although the output is mono, you can choose from five levels of gain to fine-tune the volume. Whether you're building a portable speaker, adding sound effects to a robot, or building a mini music player, the Adafruit I2S 3W Class D Amplifier Breakout offers a friendly and efficient solution for powering your audio projects.

3.2.3 Speaker

Speakers act as audio for your electronic devices by converting electrical signals into audible sound waves. These electromagnetic transmitters are available in different shapes and sizes to serve different applications. From the small speakers in your smartphone to the giant speakers in the concert hall, they all work on the same principle. A dynamic driver, usually a combination coil and diaphragm, vibrates when an electric current passes through it. This vibration is then translated into sound waves, which we think of as music, words, or other sounds.

3.2.4 Rechargeable Battery

The Rechargeable lithium batteries in our case 18650, have become the main power to power our portable electronics that is we can carry it anywhere. Li-

Figure 3.5: Speaker

ion batteries can be recharged and reused hundreds of times till there charging cycles are over. The some advantages are high energy density means they pack a lot of power into a small package and slow down the self-discharge, so they hold their charge for a long time. If we talk about the limitation Li-ion

Figure 3.6: Lithium Rechargeable battery

batteries require special care for safety . They operate in a certain voltage range, and if we use outside this range then it can damage the battery or create a fire hazard. For this reason, Device like smarthphone, earphone etc have a built-in charging circuit to monitor the voltage. In addition, Li-ion

batteries gradually lose their capacity over time, and extreme temperatures can accelerate this degradation. if proper care and use is taken into consideration then they can last long, also this batteries provide an easy and reliable power source for everyday devices.

3.2.5 Push Buttons

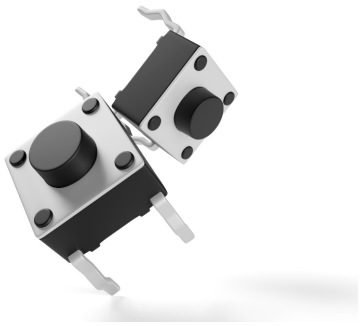

Figure 3.7: Tactile Push Buttons

A button, a fundamental electronic component, features a plastic or metal surface for clicking. When pressed, it completes an electrical circuit, activating the connected device based on pull-up or pull-down configurations. Buttons vary in shape, size, and configuration, enabling versatile control over electronic devices.

3.2.6 xcluma 5V Lithium Battery Charger Step Up Module

The Xcluma 5V Lithium Battery Charger Step-Up Protection Board is a compact module for DIY power bank projects. With a micro USB input, it's compatible with Li-Po and Li-ion 18650 batteries. It boosts battery voltage to a stable 5V output and includes protection against overcharging, overdischarging, and short circuits. This ensures safety and battery longevity for powering portable electronic devices in DIY applications.

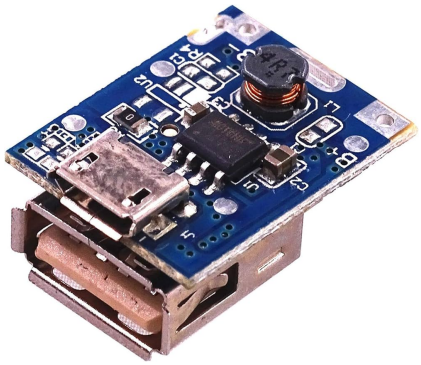

Figure 3.8: Charging Module

3.3 Software Requirements

3.3.1 Python

Python is a widely-used high-level programming language known for its readability, versatility, and extensive libraries. Unlike other languages with complex syntax, Python's code resembles plain English, making it easy to learn and write for beginners and experienced programmers alike. Its focus on readability reduces errors and facilitates maintenance.

Figure 3.9: Python Logo

Beyond its beginner-friendly nature, Python's power lies in its rich standard library, offering built-in tools for various tasks and reducing external dependencies. With a vast ecosystem of third-party libraries, Python enables a wide range of projects, from building dynamic websites with Django and Flask to data science tasks using NumPy, pandas, and Matplotlib. Python

facilitates complex data analysis and machine learning training with libraries like Scikit-learn, providing a robust set of algorithms.

3.3.2 Integrated Development IDE

3.3.2.1 Visual Studio Code

Visual Studio Code (VS Code), a beloved IDE among programmers, is a free and open-source code editor by Microsoft, available on Windows, macOS, and Linux. Its lightweight nature ensures smooth performance, even on older machines.

VS Code excels with features like Syntax highlighting for clarity and IntelliSense for intelligent coding assistance. Its extensive extension library offers themes tailored to specific languages, debuggers, Git integration, and workspace personalization options.

Notably, Git integration in VS Code facilitates collaboration, change tracking, and seamless integration with other projects or teams, empowering effortless version control and collaboration.

Figure 3.10: VScode Logo

3.3.2.2 Thonny

comes pre-installed, simplifying setup. Its clean interface focuses on Python essentials, aiding learning without distractions.

Thonny includes a recent Python version, eliminating environment setup hassles. Its debugger allows line-by-line code execution, facilitating error

identification. Visualizations for variables and call stacks make abstract concepts tangible, especially for visual learners.

For Raspberry Pi Python beginners, Thonny offers a supportive, beginner-friendly platform to learn, experiment, and gain coding confidence.

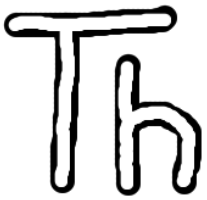

Figure 3.11: Thonny Logo

3.3.3 Putty

The Putty is a free and open-source application which functions as a terminal emulator, allowing us to connect the device to remote computers over a network. It is just like a window which opens of another machine. Putty supports various protocols, including the widely used SSH (Secure Shell) for encrypted communication for secure communication. This ensures our login credentials and data transmitted between the computer and the remote machine remain confidential. The Putty can also be used for serial console

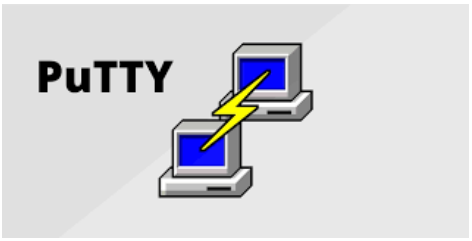

Figure 3.12: Putty Logo

connections, enabling us to interact with devices that lack a graphical user

interface (GUI) through a Command line Interface(cmd). Additionally, we can transfer files between local machine using SCP protocol(Secure Copy).

Putty offers a powerful and user-friendly solution. Its lightweight nature and flexibility make it a popular choice for various remote access needs.

3.3.4 Python Libraries

Python libraries are pre-written code bundles that you can import into your programs. They save you time and effort by providing functions and tools for various tasks, like data analysis (NumPy, Pandas), machine learning (TensorFlow, PyTorch), web development (Django, Flask), data visualization (Matplotlib, Seaborn), and more. With a vast amount of options, there's likely a library to simplify almost any programming need. We have used those libraries which are essential for products which are listed

3.3.4.1 Socket

he socket library in Python is crucial for network applications, offering a low-level interface for creating communication endpoints between programs on different machines. Sockets act as specialized ports for data exchange.

Two main socket types exist: stream sockets (e.g., TCP) ensure reliable, ordered data transfer, ideal for continuous flow apps like file transfers. Datagram sockets (e.g., UDP) prioritize speed over reliability, perfect for real-time applications like gaming or messaging.

Though socket programming can be complex, offering detailed network control, it's invaluable for developers needing custom protocols or precise networking.

3.3.4.2 Transformers

The Transformers library, based on PyTorch, revolutionizes NLP tasks with pre-trained models like BERT. These models cover various applications like

translation and summarization, trained on vast text data to understand intricate language relationships.

Transformers simplifies complex model usage. No need to build from scratch; it offers a straightforward interface to load and fine-tune pre-trained models, saving time and enabling advanced NLP without deep expertise. Whether you're a researcher, developer, or curious about NLP, Transformers empowers exploration of language processing with ease.

3.3.4.3 Google Text to Speech

GTTS, the Google Text-to-Speech library, is a free and open-source tool that effortlessly converts text into natural-sounding speech, perfect for beginners.

With a user-friendly interface, GTTS lets you select languages and voices, and adjust speaking pace. It offers offline functionality and saves speech as WAV files for playback compatibility. Whether for interactive apps, presentations, or personal projects, GTTS makes integrating text-to-speech into Python projects easily and accessible.

3.3.4.4 Pygame

Pygame, a beginner-friendly Python library for creating multimedia games, offers an accessible approach to game development. Its core features manage graphics, sound effects, and user input, providing essential building blocks for crafting engaging games.

With a clean API and abundant tutorials, Pygame facilitates quick learning and understanding for newcomers to game programming. While not ideal for high-end graphics, Pygame fosters creativity and experimentation, allowing beginners to grasp fundamental game development concepts like game loops and collision detection.

3.3.4.5 Pytorch

PyTorch, a dynamic deep learning framework, offers rapid prototyping with on-the-fly model architecture definition. Its ecosystem includes Torchvision

for computer vision and TorchAudio for audio tasks. PyTorch integrates smoothly with NumPy for efficient data manipulation. This ease, flexibility, and rich ecosystem make PyTorch ideal for researchers and developers diving into deep learning.

3.3.5 Fine tuned model

Transformer Architecture: BLIP uses a transformer-based architecture, which is the backbone of many modern natural language processing and computer vision models. This includes both vision transformers (ViT) for image processing and text transformers for language understanding and generation.

Vision-Language Pre-training (VLP): BLIP employs vision-language pre-training, which involves training on large datasets containing paired images and texts to learn the associations between visual content and language. This helps in creating a model that can effectively generate captions for images.

Dual-Encoder Structure: It features a dual-encoder structure where separate encoders are used for processing images and text. The image encoder processes the visual features of an image, while the text encoder processes the textual information.

Cross-Modal Attention: BLIP uses cross-modal attention mechanisms to integrate information from both the visual and textual domains. This allows the model to generate coherent and contextually relevant captions based on the visual content of the images.

Contrastive Learning: During pre-training, BLIP employs contrastive learning techniques. This involves learning to distinguish between correct image-text pairs and incorrect ones, which enhances the model's ability to understand and generate accurate captions.

Large Pre-training Datasets: The model is trained on extensive datasets comprising millions of image-text pairs. These datasets include images with diverse and rich descriptions, which help in improving the generalization capability of the model.

Fine-Tuning: After pre-training, the model undergoes fine-tuning on specific image captioning tasks or datasets to refine its ability to generate high-quality captions for new images.

Modality Bridging: BLIP integrates the representations from both image and text encoders to create a unified understanding of the image and its potential descriptions.

3.4 Novelty of the work

The novelty of our work are

Multimodal Integration: The work innovatively combines object detection and NLP, offering a multimodal approach that provides users with a rich and dynamic interaction with their environment.

Dynamic Adaptability to Technologies: The project's continuous monitoring and adaptation to updates in technologies like BLIP showcase its dynamic nature, ensuring ongoing compatibility with the latest advancements.

3.5 Summary of the chapter

This chapter covers work done and then the software and hardware requirements for the work.

BLIP Model: The chapter introduces the BLIP model for vision-language tasks, emphasizing its three-stage training process and components like image and text encoders. It outlines fine-tuning steps and describes how BLIP integrates with LLama for text refinement, followed by text-to-speech conversion using the gTTS library for user interaction. textbfHardware: The chapter discusses about the Micro-processor Board Raspberry Pi Zero 2W, which is small in size yet powerful. This chapter also discusses about push buttons, speakers, and rechargeable batteries. This chapter also discusses about the I2S DAC, which is used to convert digital signals from a Raspberry pi to analog signals followed by amplification for driving speaker. The chap-

ter discusses about lithium-ion batterie 18650, the TP4056 charging module. It highlights its features, like built-in overcharge protection and adjustable charging current.

Software: The chapter discusses on Python as the programming language of choice due to its readability, versatility, and extensive libraries. It also discusses two Integrated Development Environments (IDEs) for development: Visual Studio Code (VS Code) and Thonny. While VS Code is a powerful option with vast customizability, Thonny is specifically designed for beginners on the Raspberry Pi.For the remote access needs, the chapter mentions about the Putty, a free and open-source terminal emulator that supports various protocols for secure communication. This chapter also discuss into several Python libraries similar to different functionalities The chapter also briefly mentions a fine-tuned model, BLIP, which is a transformer-based architecture used for image captioning. It elaborates on BLIP's architecture, pre-training process, and key functionalities

Chapter 4

Results and Discussions

The prototype of our AI-enabled Vision Assistance model for blind individuals has been successfully developed. Both Figure 4.1 depict the prototype, showcasing the culmination of our efforts in creating a solution aimed at enhancing the lives of visually impaired individuals.

Figure 4.1: Prototype of our work

In Figure 4.2, our team member is shown wearing the glasses we've developed, providing a visual demonstration of our innovative wearable technology solution.

Figure 4.2: Person wearing the Vision Assistant Glasses

4.1 Image captioning and Large Language Model

The text extracted through optical character recognition (OCR), exemplified in Figure 4.3, serves as the primary input data for the subsequent Blip model. This sophisticated algorithm generates a descriptive analysis of the image based on the provided OCR text.

The output generated from BLIP model is shown in figure 4.4 which is the Caption of the image

However, owing to potential limitations in its output's human comprehensibility, it undergoes further refinement through the subsequent stage of processing: the LLama model. This advanced model operates by ingesting both the OCR-generated text and the image description produced by the Blip model. Through its intricate mechanisms, the LLama model refines the content, ensuring that the resulting text is not only coherent but also significantly more reader-friendly and conducive to human understanding. This iterative refinement process underscores a commitment to enhancing the accessibility and clarity of image-derived information, contributing to a more enriched and intelligible user experience.

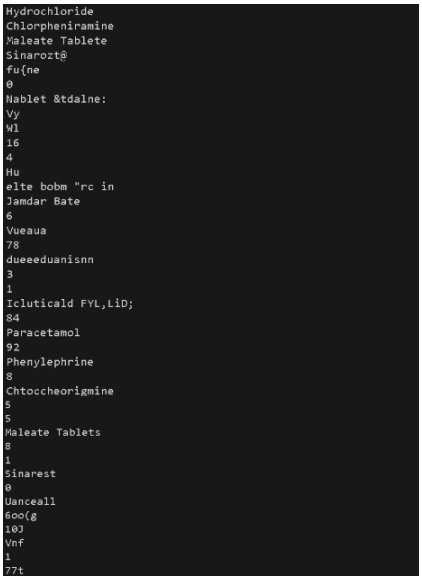

Figure 4.3: Ocr-generated text as input

Figure 4.4: Output of BLIP Model

The image described in Figure 4.5 showcases the refined output of the LLama model, notable for its heightened human readability. This exemplifies a significant advancement in data processing, where intricate algorithms refine OCR data and image descriptions, presenting them in a coherent and easily understandable format, enhancing accessibility and comprehension.

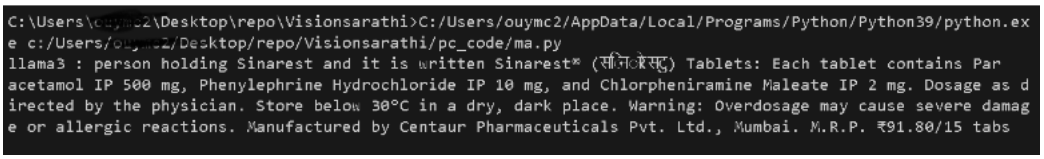

Figure 4.5: Output of LLama Model of OCR

In addition to OCR, Figure 4.6 illustrates how the Blip Model provides image captions. These captions, generated alongside OCR data, represent a holistic approach to image analysis. By incorporating both OCR and cap-

tioning, the Blip Model offers comprehensive insights into image content, facilitating clearer understanding and interpretation of visual information.

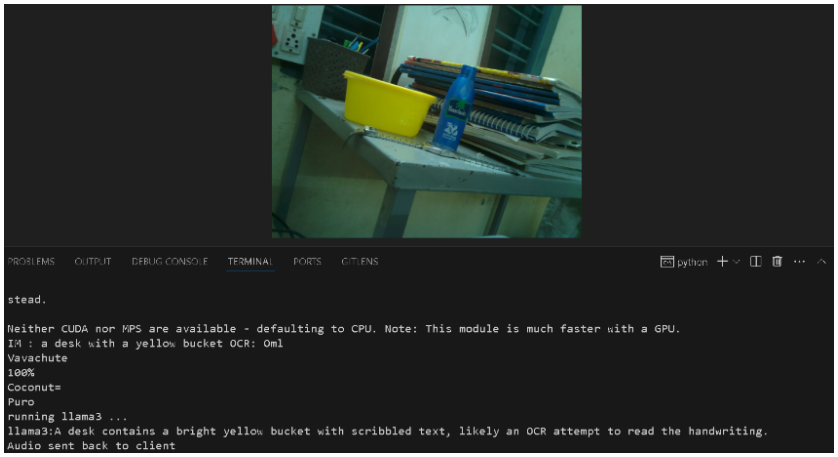

Figure 4.6: Output of LLama Model of Image Description

4.2 Conversion of text to audio file

We have generate human-readable text uing the LLama model, which is then converted into an audio file format such as .mp3 using the Google Text-to-Speech (gtts) API. This process involves transforming the textual output from the model into spoken words, which are synthesized into speech by the TTS engine provided by Google.

```
a desk with a yellow Cup.
The text is being converted to Audio.mp3
Audio sent back to client
```

Figure 4.7: Text to Audio Conversion Output

The resulting .mp3 file contains the spoken rendition of the text, allowing users to listen to the content rather than reading it. This conversion enhances accessibility and provides an alternative way to consume information for individuals who may prefer auditory learning or have visual impairments.

Additionally, it expands the reach of generated content to a broader audience by offering a multi-modal experience.

4.3 Audio Playback

The Audio generated by gtts is being sent back to raspberry pi from the server(in our case its laptop) after all the processing is completed, including gtts and the audio file is saved in raspberry pi. Then a pygame library is used to play the audio file. The Audio is converted to analog signal using I2S DAC and its being amplified and the audio of the text is being played in the speaker. The figure 4.8 depicts that the audio file is recived from the server and it has started its playback using pygame.

```
pi@raspberrypi:~/p/Visionsarathi/rasppi $ python main.py
pygame 1.9.6
Hello from the pygame community. https://www.pygame.org/contribute.html
GPIO pin is LOW. Capturing and sending image...
hello
<re.Match object; span=(3, 17), match='192.168.120.55'>
192.168.120.55
Connected to the device at IP: 192.168.120.55
Image sent to server
Received audio data (format: mp3)
audio.mp3
audio recived.audio is playing...
```

Figure 4.8: Audio File Recived and Playback

4.4 Summary of the chapter

The Summary of result and discussion is as follows

1. Input Processing and Captioning:
    - Text Extraction: Optical Character Recognition (OCR) extracts text from images (Figure 4.1).
    - Image Content: The system takes the image itself as input.

- The BLIP model analyzes the image content and generates descriptive captions based on its analysis.

2. Text Refinement with LLama (Figure 4.3):

   - LLama, a large language model, further refines the combined output from OCR and BLIP.
   - This refinement step ensures the final description is clear, coherent, and easily understandable by humans.

3. Audio Generation (Section 4.2):

   - The refined text is converted into an audio file format (e.g., MP3) using the Google Text-to-Speech (gTTS) API.

4. Audio Playback (Section 4.3):

   - The audio file is sent back to the Raspberry Pi from the server.
   - The Pygame library is used for playing the audio file.
   - An I2S Digital-to-Analog Converter (DAC) converts the digital audio signal to an analog signal.
   - The analog audio signal is then amplified and played through a speaker connected to the Raspberry Pi.

Chapter 5

Conclusion

5.1 Conclusion

The AI-powered vision assistant system, built using Raspberry Pi Zero 2W and integrated with advanced AI models such as BLIP, demonstrates a remarkable achievement in assistive technology for visually impaired individuals. By combining image capture, image captioning, OCR, and text-to-speech conversion, the model translates visual information into detailed auditory descriptions, providing users with enhanced accessibility and autonomy.

Key results from the work, including the output from the provided sample data, reinforce the system's effectiveness:

Effective Image Understanding: The system demonstrates the capability to accurately interpret visual scenes, generating concise image captions that capture the essence of the content.

Refined Auditory Feedback: Through LLama processing and text-to-speech conversion, the model delivers tailored auditory descriptions that combine image captions, OCR outputs, and refined context, providing visually impaired users with comprehensive and accessible information about their surroundings.

Code Implementation: The project's code, consisting of Python scripts for both the server (PC) and client (Raspberry Pi), showcases efficient image

processing, model inference, and communication protocols using sockets. The server-side code receives image data, processes it using AI models, and sends back audio responses, while the client-side code captures images, sends them to the server, and plays back the received audio feedback.

In conclusion, the AI-powered vision assistant model exemplifies the transformative potential of AI and innovative hardware integration in creating inclusive technologies. By empowering visually impaired individuals with detailed auditory descriptions of their environment, the system promotes accessibility, independence, and equality, signaling a promising future for assistive technologies in the digital age. Continued research, development, and collaboration in this field will drive further advancements, ensuring that individuals of all abilities can fully participate in and contribute to society.

5.2 Advantages and Disadvantages

5.2.1 Advantages

- Cost-effective and Portable: The Raspberry Pi Zero 2W is a low-cost, single-board computer, making the system affordable and accessible. Its compact size allows for easy portability, ideal for everyday use.

- Open-source Software: The work utilizes open-source libraries like BLIP, gTTS and LLama model. This makes the system replicable and adaptable for further development by the community.

- User-Friendly Interface: Voice prompts delivered through the speaker provide clear and concise information for users, promoting ease of use and also push button for the occurance of event.

- Energy Efficiency: Raspberry Pi Zero 2W is energy-efficient, reducing power consumption for cost savings and longer battery life, ideal for off-grid or portable applications.

- Expandability and Customization: With GPIO pins, Raspberry Pi Zero 2W offers limitless customization options, allowing users to connect various sensors and peripherals for tailored projects and applications.

5.2.2 Disadvantages

- Limited Processing Power: The Raspberry Pi Zero 2W has a less processing power compared to more powerful computers. This might decrease the performance of the process.

- Accuracy of BLIP Model and oOCR: The accuracy of object recognition and information delivery relies on the BLIP model's training data and capabilities. Further refinement might be needed to improve the system's reliability.

5.3 Future scope

The future of this AI-enabled vision assistance system using Raspberry Pi holds immense potential. Here are some exciting possibilities:

Model Enhancements: Advanced AI models with deeper learning capabilities could be explored to improve object recognition accuracy and provide more nuanced descriptions of the environment. This could include identifying specific objects (e.g., types of furniture) or even hazards like oncoming traffic.

Sensory Integration: Incorporating additional sensors like LiDAR or depth cameras could enable the system to create a 3D understanding of the surroundings. This would be especially beneficial for tasks like obstacle detection and path planning, further enhancing assistive capabilities.

Advanced User Interface: Beyond voice prompts, exploring haptic feedback through wearable devices could provide additional information about objects or surroundings.

Cloud Integration: Leveraging cloud computing could enable the system to access vast amounts of data and real-time information. This could allow for object recognition beyond the model's pre-trained capabilities or even connect users with emergency services if needed.

References

[1] K. Xia, X. Li, H. Liu, M. Zhou, and K. Zhu, "Ibgs: A wearable smart system to assist visually challenged," *IEEE Access*, vol. 10, pp. 77810–77825, 2022.

[2] E. Ali Hassan and T. B. Tang, "Smart glasses for the visually impaired people," in *Computers Helping People with Special Needs: 15th International Conference, ICCHP 2016, Linz, Austria, July 13-15, 2016, Proceedings, Part II 15*, pp. 579–582, Springer, 2016.

[3] X. Yang, H. Yu, and L. Jia, "Speech recognition of command words based on convolutional neural network," in *2020 International Conference on Computer Information and Big Data Applications (CIBDA)*, pp. 465–469, IEEE, 2020.

[4] M. Anu, S. Divya, *et al.*, "Building a voice based image caption generator with deep learning," in *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 943–948, IEEE, 2021.

[5] J. Bai, S. Lian, Z. Liu, K. Wang, and D. Liu, "Smart guiding glasses for visually impaired people in indoor environment," *IEEE Transactions on Consumer Electronics*, vol. 63, no. 3, pp. 258–266, 2017.

[6] S. Sukhbaatar, J. Weston, R. Fergus, *et al.*, "End-to-end memory networks," *Advances in neural information processing systems*, vol. 28, 2015.

[7] S. Chordia, Y. Pawar, S. Kulkarni, U. Toradmal, and S. Suratkar, "Attention is all you need to tell: Transformer-based image captioning," in *Advances in Distributed Computing and Machine Learning: Proceedings of ICADCML 2022*, pp. 607–617, Springer, 2022.

[8] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[9] M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado, *et al.*, "Google's multilingual neural machine translation system: Enabling zero-shot translation," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 339–351, 2017.

[10] O. Vinyals and Q. Le, "A neural conversational model," *arXiv preprint arXiv:1506.05869*, 2015.

[11] S. N. Saud, L. Raya, M. I. Abdullah, and M. Z. A. Isa, "Smart navigation aids for blind and vision impairment people," in *International Conference on Computational Intelligence in Information System*, pp. 54–62, Springer, 2021.

[12] A. V. Yadav, S. S. Verma, and D. D. Singh, "Virtual assistant for blind people," *International Journal*, vol. 6, no. 5, 2021.

[13] Y. Bouteraa, "Design and development of a wearable assistive device integrating a fuzzy decision support system for blind and visually impaired people," *Micromachines*, vol. 12, no. 9, p. 1082, 2021.

[14] F. Ashiq, M. Asif, M. B. Ahmad, S. Zafar, K. Masood, T. Mahmood, M. T. Mahmood, and I. H. Lee, "Cnn-based object recognition and tracking system to assist visually impaired people," *IEEE Access*, vol. 10, pp. 14819–14834, 2022.

[15] J. Du, "Understanding of object detection based on cnn family and yolo," in *Journal of Physics: Conference Series*, vol. 1004, p. 012029, IOP Publishing, 2018.

[16] J.-H. Kim, S.-K. Kim, T.-M. Lee, Y.-J. Lim, and J. Lim, "Smart glasses using deep learning and stereo camera," in *2019 IEEE 8th Global Conference on Consumer Electronics (GCCE)*, pp. 294–295, IEEE, 2019.

[17] M. R. Kadhim and B. K. Oleiwi, "Blind assistive system based on real time object recognition using machine learning," *Engineering and Technology Journal*, vol. 40, no. 1, pp. 159–165, 2022.

[18] H. S. Abdul-Ameer, H. J. Hassan, and S. H. Abdullah, "Development smart eyeglasses for visually impaired people based on you only look once," *Telkomnika (Telecommunication Computing Electronics and Control)*, vol. 20, no. 1, pp. 109–117, 2022.