VQA Voyager: Voice-Based Visual Question Answering for Cultural Heritages in Kathmandu Valley

Team Members

Arnab Manandhar **(THA077BEI008)**
Chandra Mohan Sah **(THA077BEI017)**
Looza Subedy **(THA077BEI024)**
Santosh Acharya **(THA077BEI040)**

Under the Supervision of
Associate Prof. Suramya Sharma Dahal

Department of Electronics and Computer Engineering
Institute of Engineering, Thapathali Campus

July, 2024

Presentation Outline

- Motivation
- Introduction
- Objectives
- Scope of Project
- Methodology
- Dataset Analysis
- Results
- Discussion and Analysis
- Remaining Tasks
- References

Motivation

- Enhance tourists' cultural understanding and appreciation
- Bridge information gaps at heritage sites
- Provide interactive, real-time artifact information
- Utilize AI for enriched tourist experiences
- Foster deeper engagement with cultural heritage
- Make heritage sites more accessible
- Empower tourists with instant historical insights

Introduction

- AI automates mundane tasks, saving time and effort
- Opens new possibilities for cultural understanding
- CV and NLP methods have potential to significantly improve tourists knowledge
- VQA: Promising CV and NLP task
- Most common VQA model answers image-related questions
- Image and question is taken as input based on which accurate predictions is done

Objectives

- To develop a Visual Question Answering (VQA) tool that answers questions based on the context of captured image
- To create a voice-based app that allows the user to ask questions and capture images

Scope of Project

- Develop an app to help tourists identify artifacts
- Integrate Visual Question Answering (VQA) for image processing and natural language processing
- Provide accurate answers to queries about the captured artifacts
- Ensure an intuitive and accessible experience for tourists

Proposed Methodology - [1]
System Block Diagram

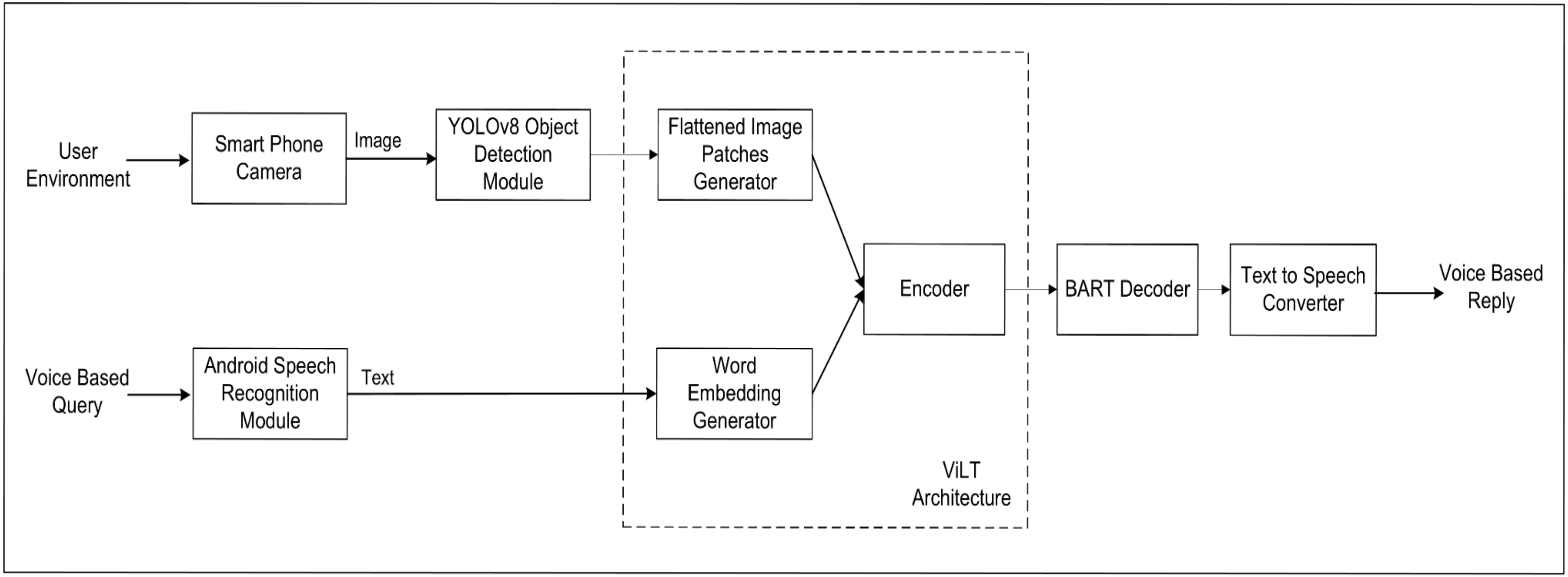


Figure: System Block Diagram

Proposed Methodology - [2] Description of Working Principle

- User can capture a picture and ask a voice based question
- Android speech recognizer converts speech to text
- The objects in the picture are identified by YOLOv8 model
- The bounding box region and question are passed through ViLT model
- The ViLT model provides a joint embedding of image and question
- The joint embedding is passed to BART decoder
- BART decoder provides a descriptive answer
- The textual answer is converted into voice based reply using android text-to-speech

Proposed Methodology - [3]
YOLOv8 - [1]

- YOLOv8 takes image as an input and outputs the bounding box of the objects
- Three components: Backbone, Neck and Head
- Backbone
  - Extracts features from images using multiple layers
  - Uses CSPDarknet backbone for feature extraction
- Neck
  - Merges feature maps from different stages of the backbone to capture information at various scales

Proposed Methodology - [3]
YOLOv8 - [2]

- Head
  - predict bounding boxes, objectness scores, and class probabilities for each grid cell in feature map
- YOLOv8-nano is used for object detection
  - Lightweight and consumes less computational resources
  - Achieves similar accuracy comparable to it’s larger models

Proposed Methodology - [3]
ViLT (Vision and Language Transformer) - [1]

- Processes both visual and textual information directly through transformer layers without convolutional operations
- The image is divided into fixed-size patches (16x16 pixels).
  - Each patch is treated as a separate token
- Each image patch is linearly embedded into a vector representation
- Positional encoding is added to retain spatial information
- Text is tokenized as well and linearly embedded with positional encoding

Proposed Methodology - [3]
ViLT (Vision and Language Transformer) - [2]

- The image embedding and text embedding are combined and fed into the transformer model
- The model outputs a contextual embedding of [number of image-text pairs, sequence length, embedding dimension]
  - Sequence length : number of tokens or patches in the sequence
  - Embedding dimension: size of the feature vectors for each token or patch

Proposed Methodology - [3] BART (Bidirectional Autoregressive Transformer) - [1]

- Utilizes separate encoder and decoder components, enabling sequence-to-sequence learning
- Uses bidirectional encoder and auto regressive decoder
- BART Decoder
  - It predicts the next tokens by taking the previously generated tokens into consideration (Auto-regressive)
- We only use BART decoder, which accepts the context embedding from ViLT Encoder
- Generates the answer based on given embeddings

Dataset Analysis -[1]

- Created custom dataset which contains features such as:
  - Images depicts cultural objects and sites in Kathmandu Valley
  - Currently contains a total of 597 images
  - Contains over 40 images for each object
- Dataset has major two components
  - Objects
  - Question Answer pairs

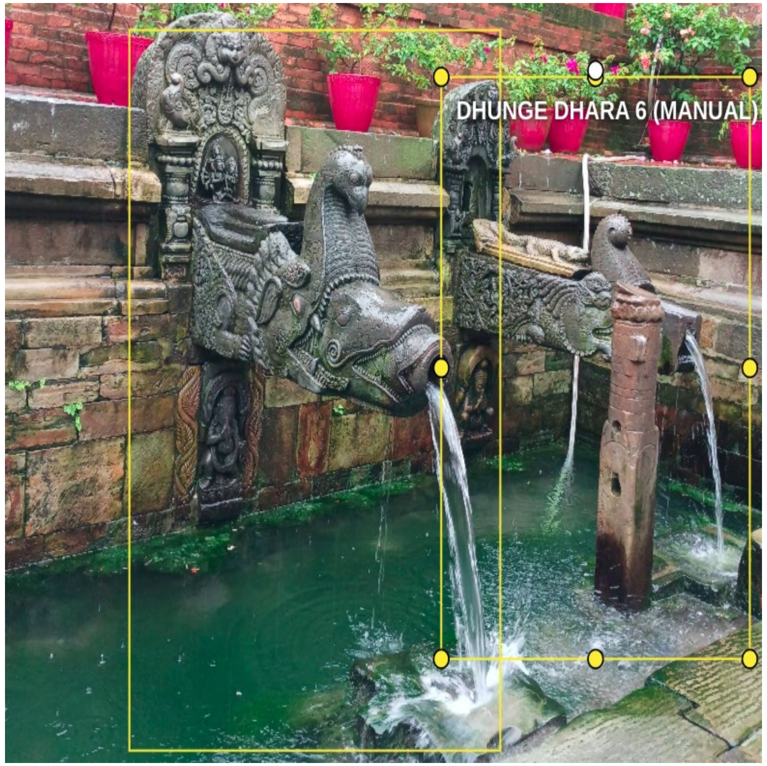
Dataset Analysis -[2]

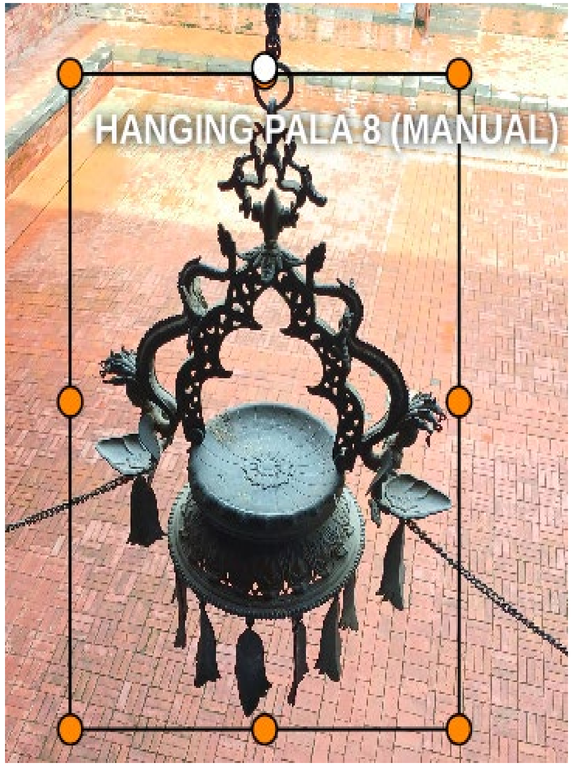
- Image augmentation has been done to increase the size and diversity of training dataset
- Common techniques used are rotation, brightness adjustment and saturation adjustment
- Classes of images prepared till now are:

    - Class A : Ankhi Jhyal
    - Class B : Taleju Bell
    - Class C : Dhunge Dhara
    - Class D : Krishna Mandir
    - Class E : Hanging Pala
    - Class F : Prayer Wheel
    - Class G : Yali

Dataset Analysis -[3]

- The objects and their bounding boxes are as shown below:





Dataset Analysis -[4]

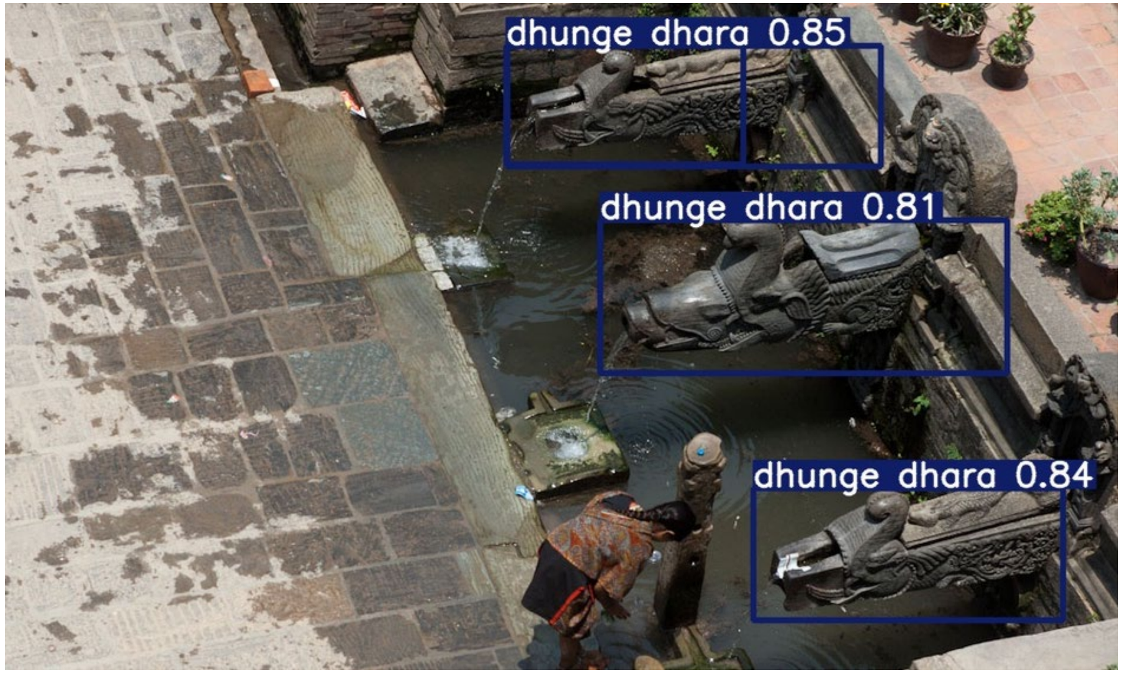
- Question answering pairs has been illustrated along with 20 QA pairs per object
- Around 140 question answer pair has been created till now
- Category of Question:
  - Where, How, When, What, Why?


```
"image_id": 6,
"qa_pairs": [
  {
    "question": "When was Krishna Mandir built?",
    "answer": "Krishna Mandir was built in the 17th century."
  },
  {
    "question": "Who built Krishna Mandir?",
    "answer": "Krishna Mandir was built by King Siddhi Narsing Malla."
  },
  {
    "question": "What style is Krishna Mandir built in?",
    "answer": "Krishna Mandir is built in the Shikhara-style."
  },
```

```
"image_id": 2,
"qa_pairs": [
  {
    "question": "Where is the Taleju Bell located?",
    "answer": "The Taleju Bell is located in Patan Durbar Square, Kathmandu, Nepal."
  },
  {
    "question": "What is the Taleju Bell opposite to?",
    "answer": "The Taleju Bell is situated opposite the Royal Palace."
  },
  {
    "question": "When was the Taleju Bell erected?",
    "answer": "The Taleju Bell was erected in 1736."
  },
```

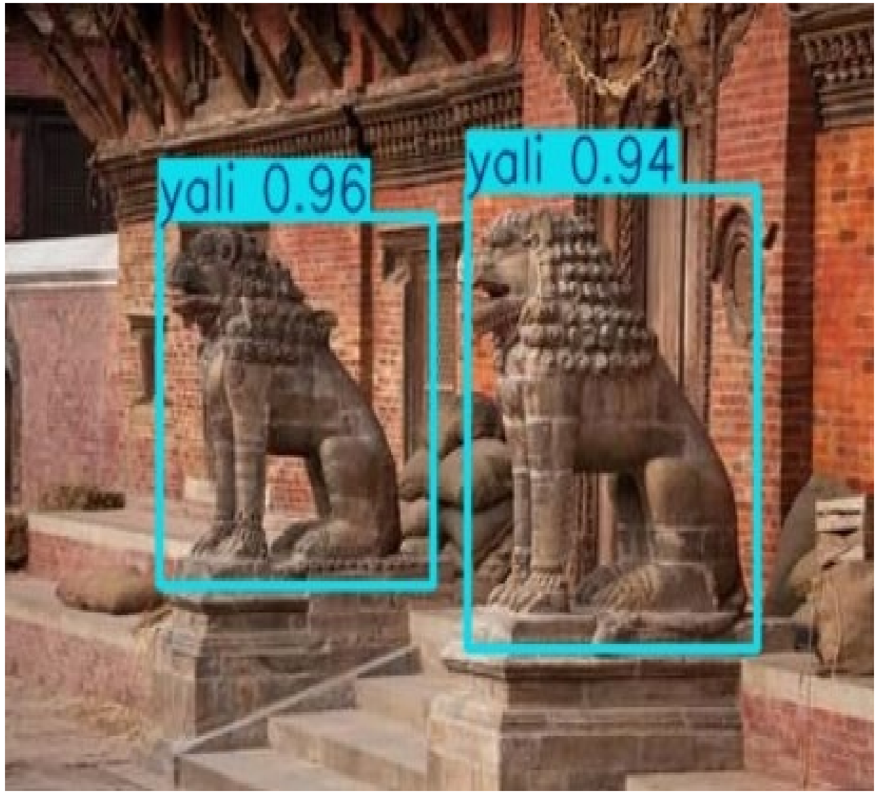
Results [1] - Object Detection Module [1]

- Inference results using YOLOv8

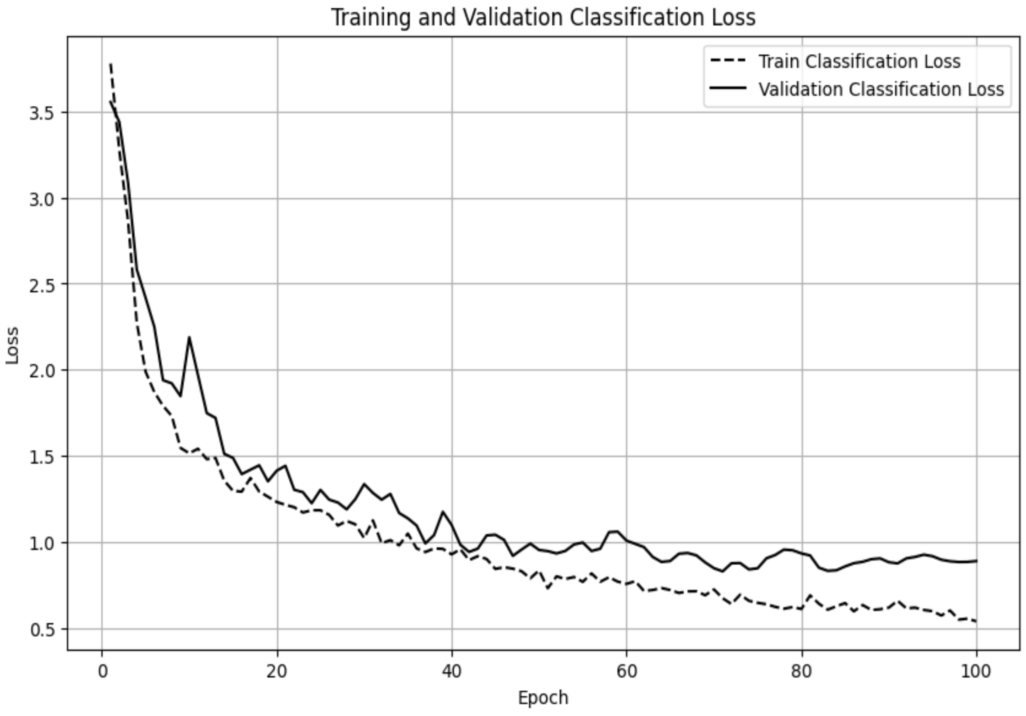


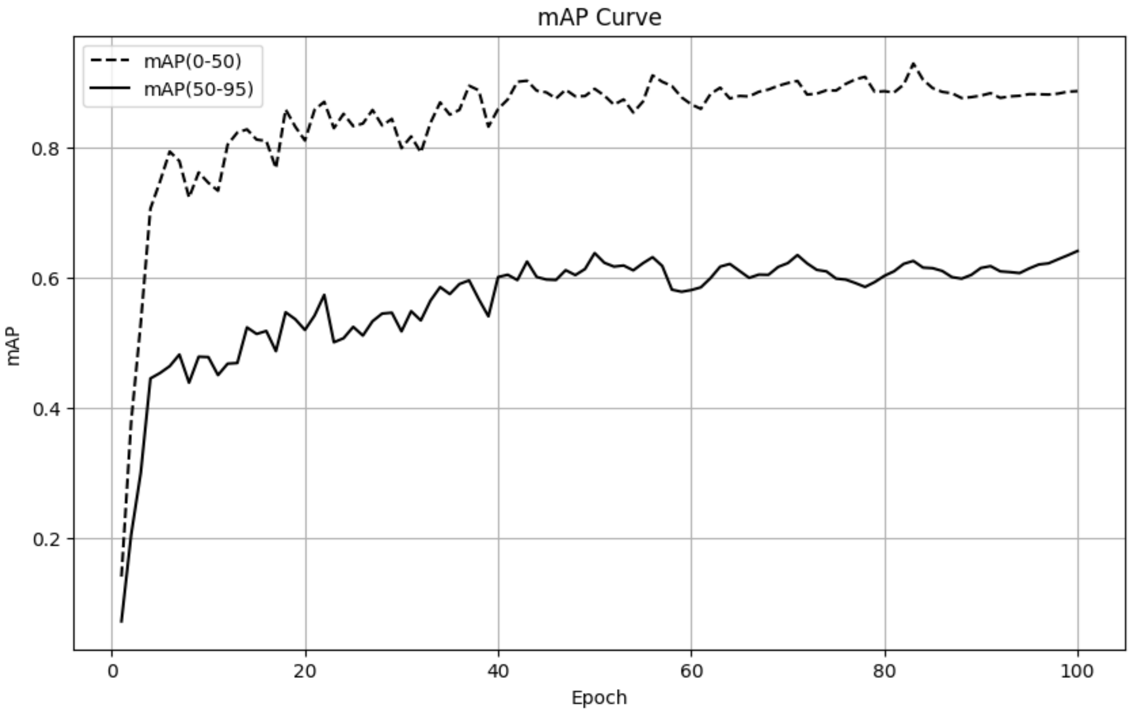
Results [1] - Object Detection Module [2]

- Inference results using YOLOv8

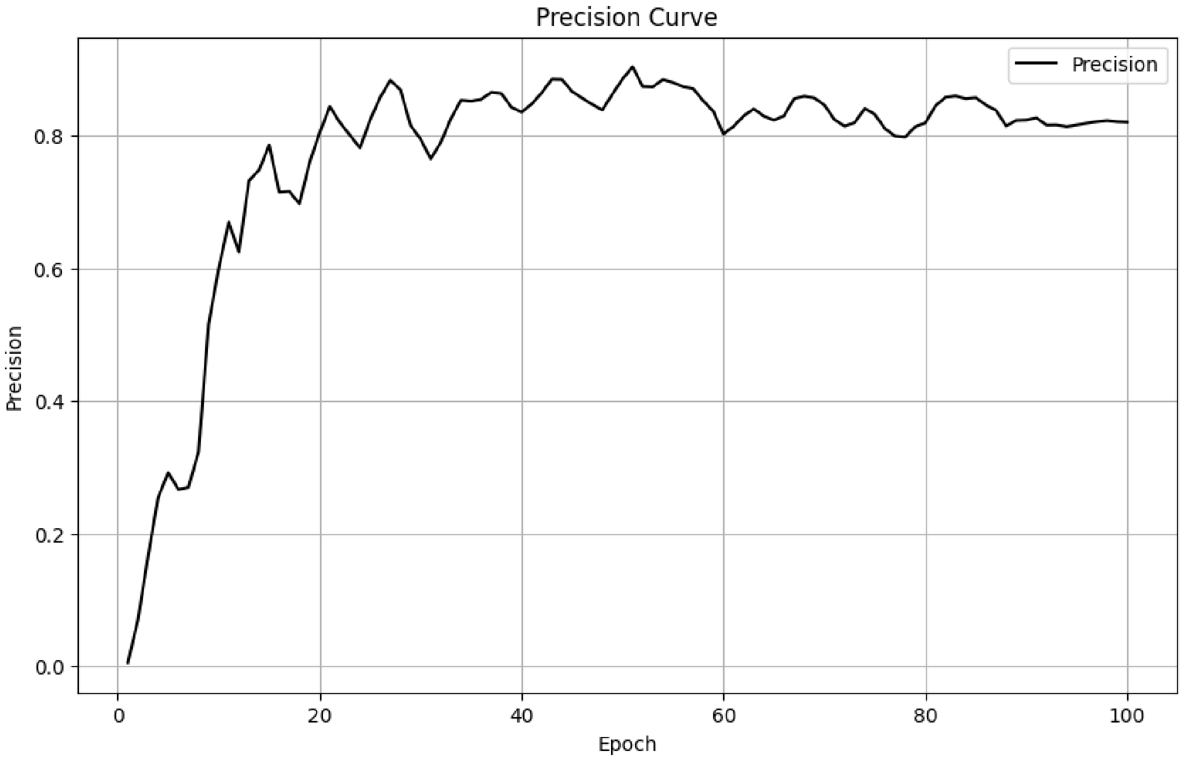


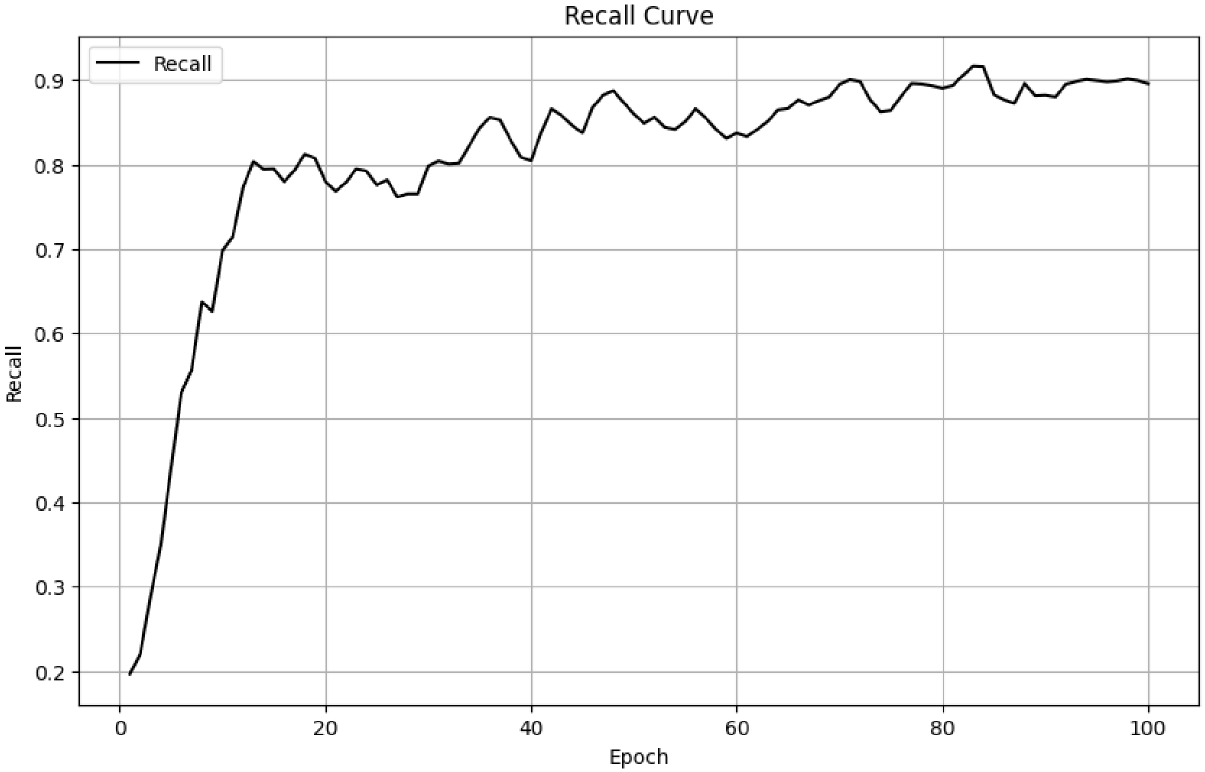
Results [1] - Object Detection Module [3]



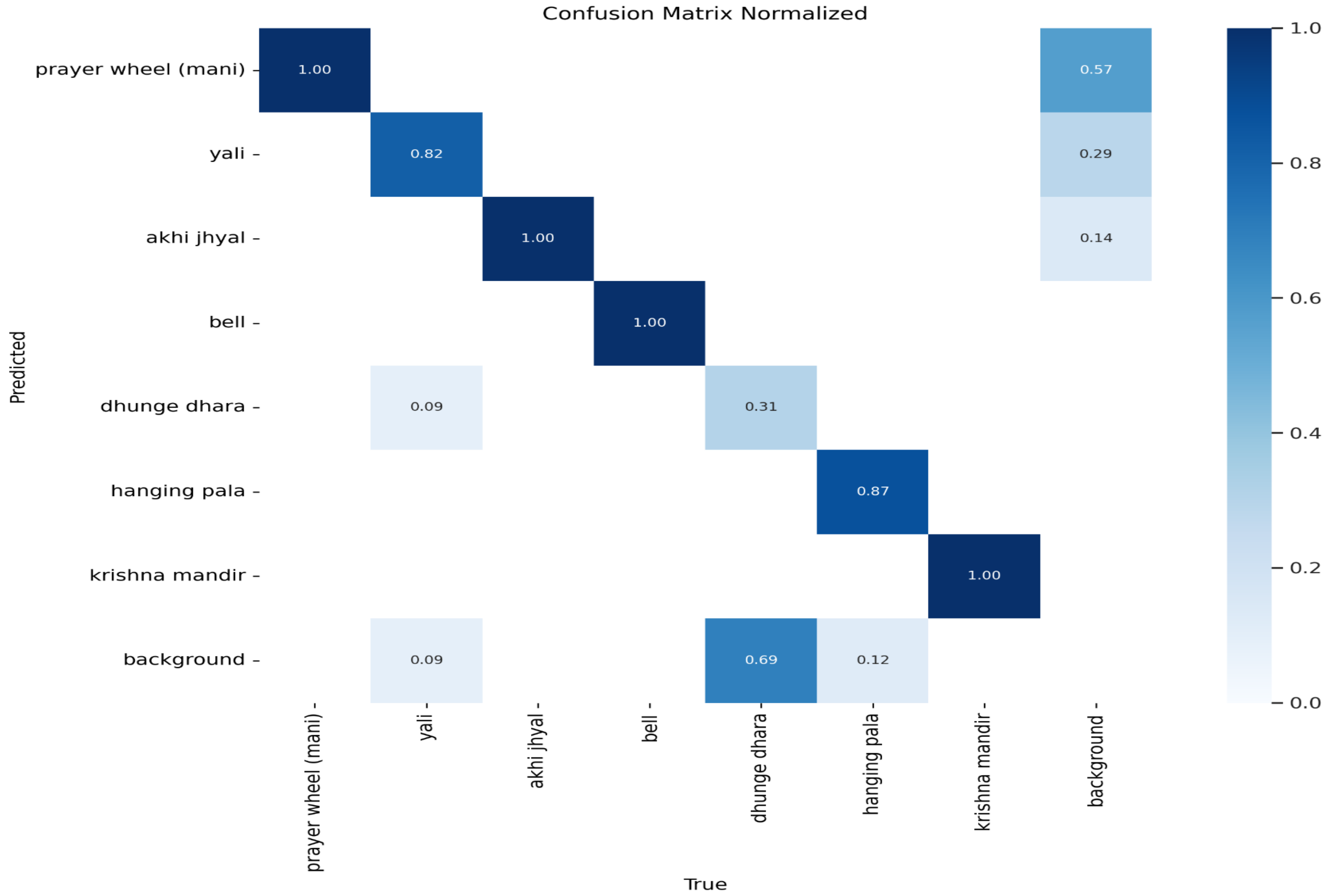


Results [1] - Object Detection Module [4]



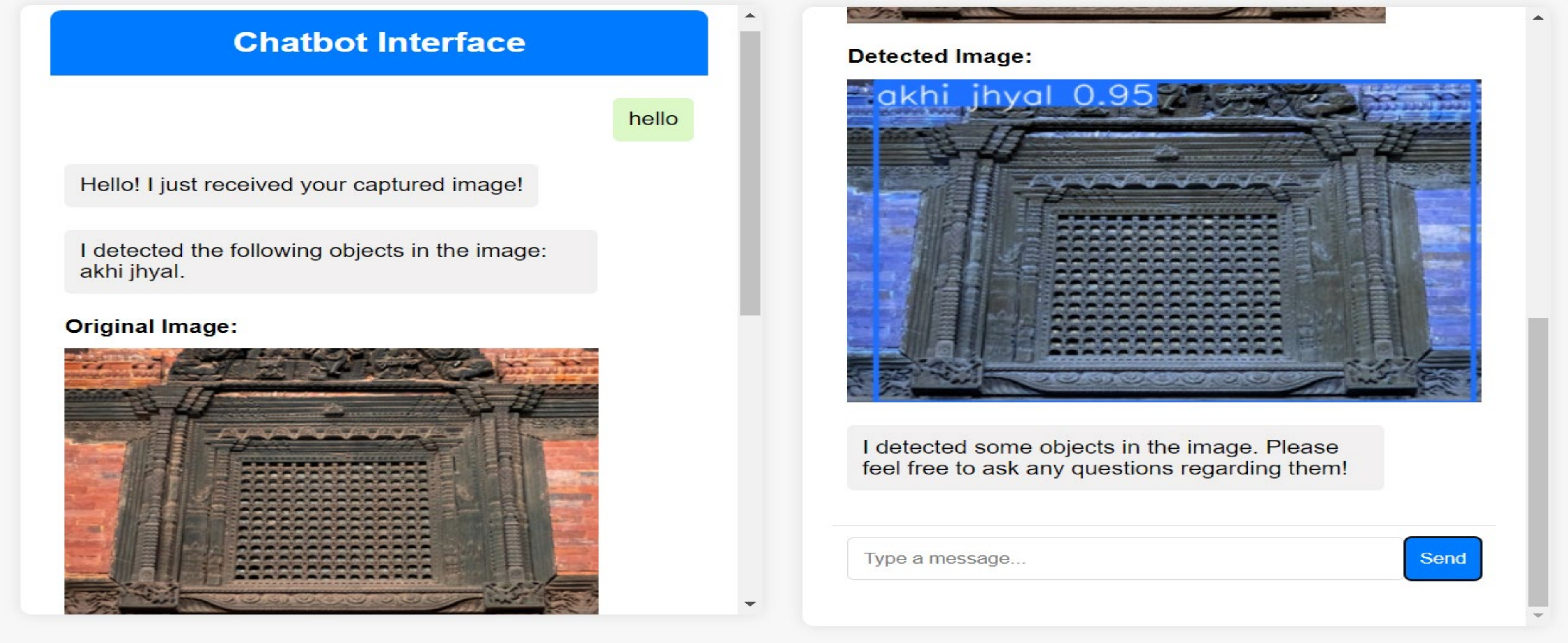


Results [1] - Object Detection Module [5]



Results [2] – Chatbot [1]

- Chatbot Integration



Discussion and Analysis – Object Detection

- The maximum values of evaluation metrics were obtained to be as follows:
    - Precision : 94.97%
    - Recall: 92.59%
    - mAP50: 92.36%
    - mAP90: 64.16%
- mAP90 only achieves a 64.26% and the errors in confusion matrix are found to be due to smaller size of dataset

Discussion and Analysis - Chatbot

- RASA has been used to create the chatbot
- Chatbot has currently been used in a website hosted locally
- Chatbot is able to integrate YOLO detection model for input
- Interface displays original and object detected images in the chat
- Needs to be trained on more probable questions and answers for further interaction
- Deploying the Chatbot on a mobile application aligns more with the project objective

Remaining Tasks

- Part A -
- Training and testing ViLT encoder and BART decoder on current small dataset
- Further training of chatbot to integrate entire VQA model
- Part B -
- Increasing the size of dataset
- Final training and testing on larger dataset
- App development

References -[1]

[1] M. M. a. M. Fritz, "Towards a Visual Turing Challenge," 2015.

[2] S. A. e. al, "VQA: Visual Question Answering,," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 2015.

[3] M. R. Mateusz Malinowski, "Ask Your Neurons: A Neural-based Approach to Answering Questions about Images," in *Conference: International conference on computer vision (ICCV)*, Santiago, 2015.

[4] R. K. a. R. Z. Mengye Ren, "Image Question Answering: A Visual Semantic Embedding Model and a New Dataset," in *Deep Learning Workshop at ICML 2015*, 2015.

[5] N. P. H. S. B. H. Hyeonwoo, "Image Question Answering Using Convolutional Neural Network with Dynamic Parameter Prediction," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, 2016.

References -[2]

[6] X. H. J. G. e. a. Z. Yang, "Stacked Attention Networks for Image Question Answering," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, 2016.

[7] X. H. e. a. Peter Anderson, "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, 2018.

[8] D. Denis, "Using Deep Learning to Answer Visual Questions from Blind People," KTH Royal Institute of Technology, Stockholm, 2019.

[9] A. B. P. A. M. S. a. P. A. P. Patil, "Speech Enabled Visual Question Answering using LSTM and CNN with Real Time Image Capturing for assisting the Visually Impaired," in *TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON)*, Kochi, 2019.