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

August, 2024

Presentation Outline

- Motivation
- Introduction
- Objectives
- Scope of Project
- Project Applications
- 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 the captured image
- To capture images and create a voice-based app that allows the user to ask questions about the image

Scope of Project

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

Project Applications

- Assistance in exploring world heritage sites for tourists
- Assist users in identifying artifacts and understanding their history
- Enhance the cultural experience with detailed information on demand
- Provide educational insights about historical objects and artifacts
- Enhance engagement through interactive and personalized learning
- Promote cultural appreciation and preservation through accessible information

Methodology - [1]
System Block Diagram

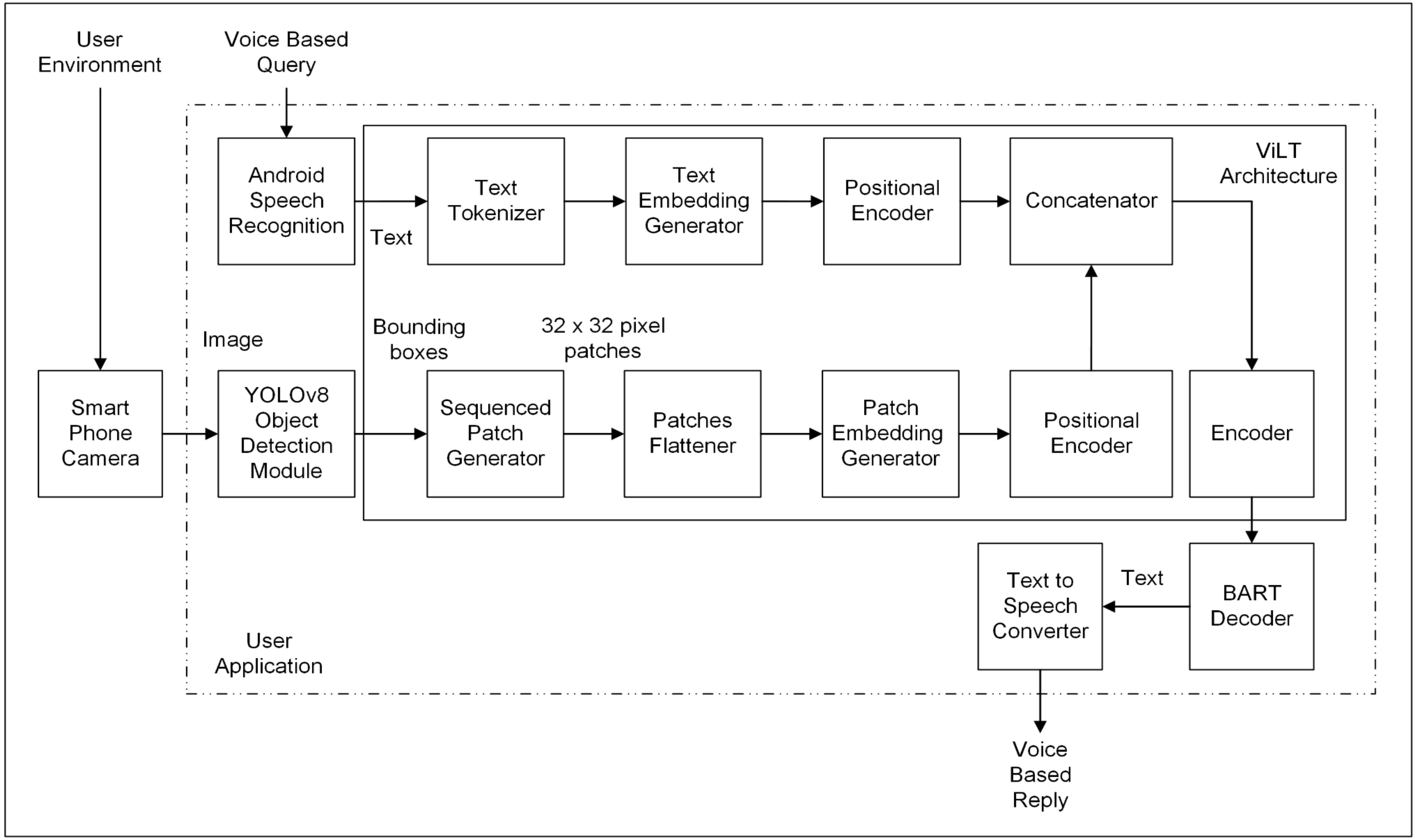


Figure: System Block Diagram

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

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
- Neck
  - Merges feature maps from different stages of the backbone to capture information at various scales
- Head
  - Predict bounding boxes, objectness scores, and class probabilities for each grid cell in feature map

Methodology - [3]
YOLOv8 - [2]

- The training parameters for YOLOv8 are:

Hyperparameters	Values
Epochs	100
Batch size	16
Image size	640
Optimizer	AdamW
Learning rate	0.01
Dropout	0.0

Hyperparameters	Values
Freeze	null
IoU	0.7
Box	7.5
Class	0.5
DFL	1.5

Methodology - [4]
ViLT (Vision and Language Transformer) - [1]

- Processes both visual and textual information directly
- The image is divided into fixed-size patches of (32x32)
- Positional encoding is added to retain spatial information

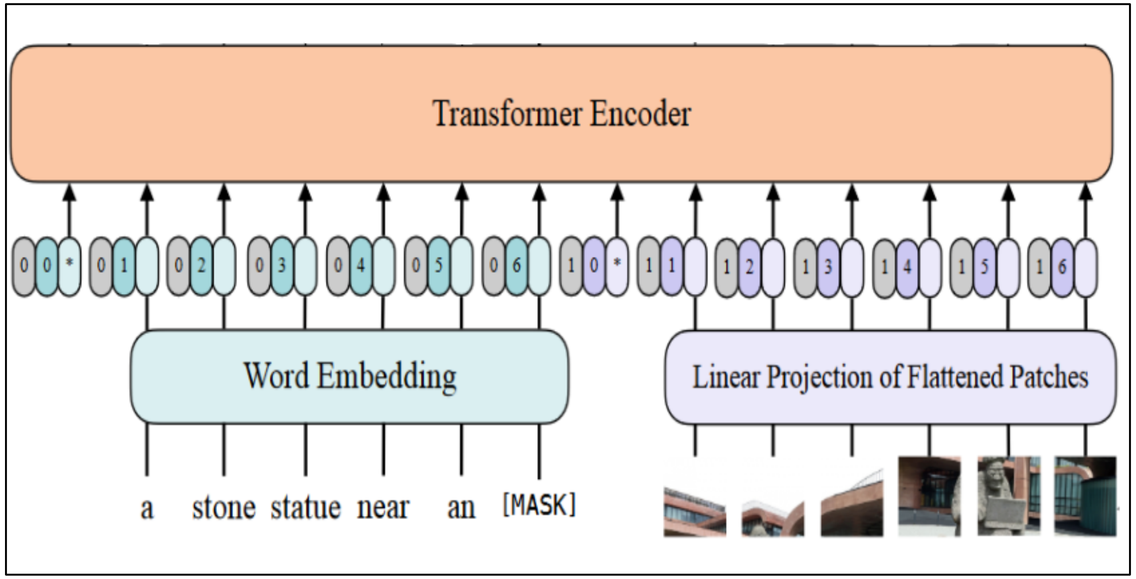


Figure: ViLT encoder

Methodology - [4]
ViLT (Vision and Language Transformer) - [2]

- Text is tokenized as well and embedded with positional encoding
- The image embedding and text embedding are combined and fed into the transformer encoder
- 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 (768 dimensions)

Methodology - [5]
BART (Bidirectional Autoregressive Transformer)

- 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

Methodology - [6]
Android Application - [1]

- What is Flutter?
  - Open-source software development kit (SDK)
  - Uses Dart Programming Language enabling hot reload
- Purpose in project
  - Create attractive and responsive User Interface
  - Send input parameter (image and text) to server
  - Receive the response from server

Methodology - [6]
Android Application - [2]

- Communication Interface
  - Communication Hierarchy between various components can be seen in the diagram

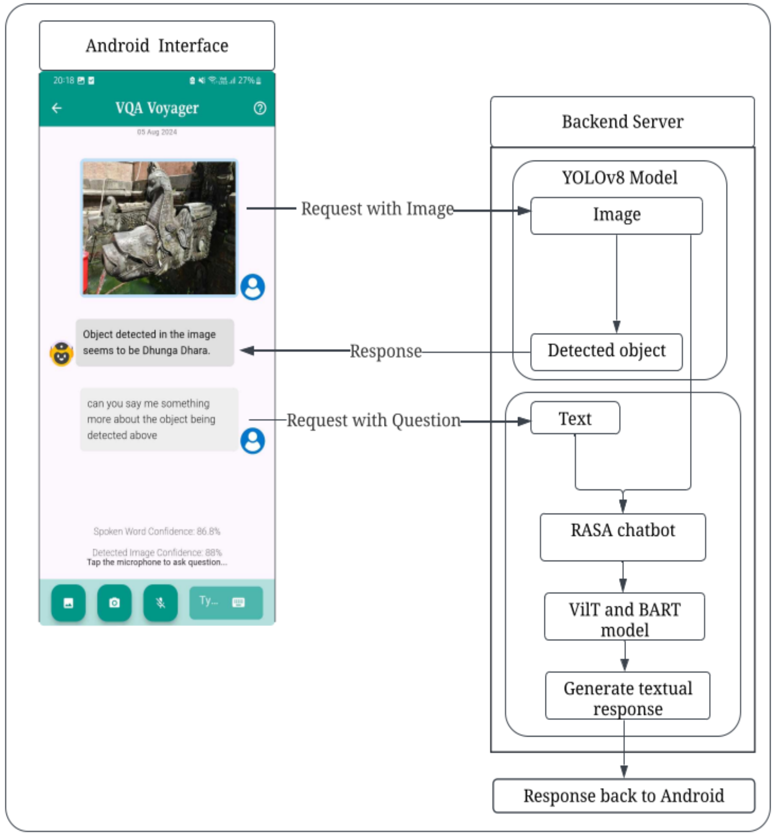


Figure: Application Communication Interface

Methodology - [7]
RASA Chatbot - [1]

- What is Rasa?
  - Open-source framework
  - Conversational AI
- What is Rasa chatbot?
  - AI-driven system that understands user inputs
  - Responds via NLP
- Purpose in project
  - Visualize chatbot interactions during initial development stage
  - Facilitate an engaging conversation for users

Methodology - [7]
RASA Chatbot - [2]

- Rasa chatbot has been used with Web Application
  - Designed for user interaction
  - Facilitates image processing and chat functionality
- Rasa chatbot workflow
  - Users upload images through the web interface
  - Engage in chat to receive responses and processed image
- Domain.yml file sets up the chatbot framework with intents and entities
- Rules.yml file specifies dialogue rules to ensure consistent responses
- Custom actions integrate the YOLO model for object detection

Dataset Analysis - [1]
Data Collection - [1]

- Images were collected through site visits around Kathmandu Valley
- Additional images obtained by web scraping from Shutterstock and existing datasets
- The final dataset contains
  - 6812 images
  - 12 classes

Dataset Analysis - [1]
Data Collection - [2]

- Distribution of images in dataset

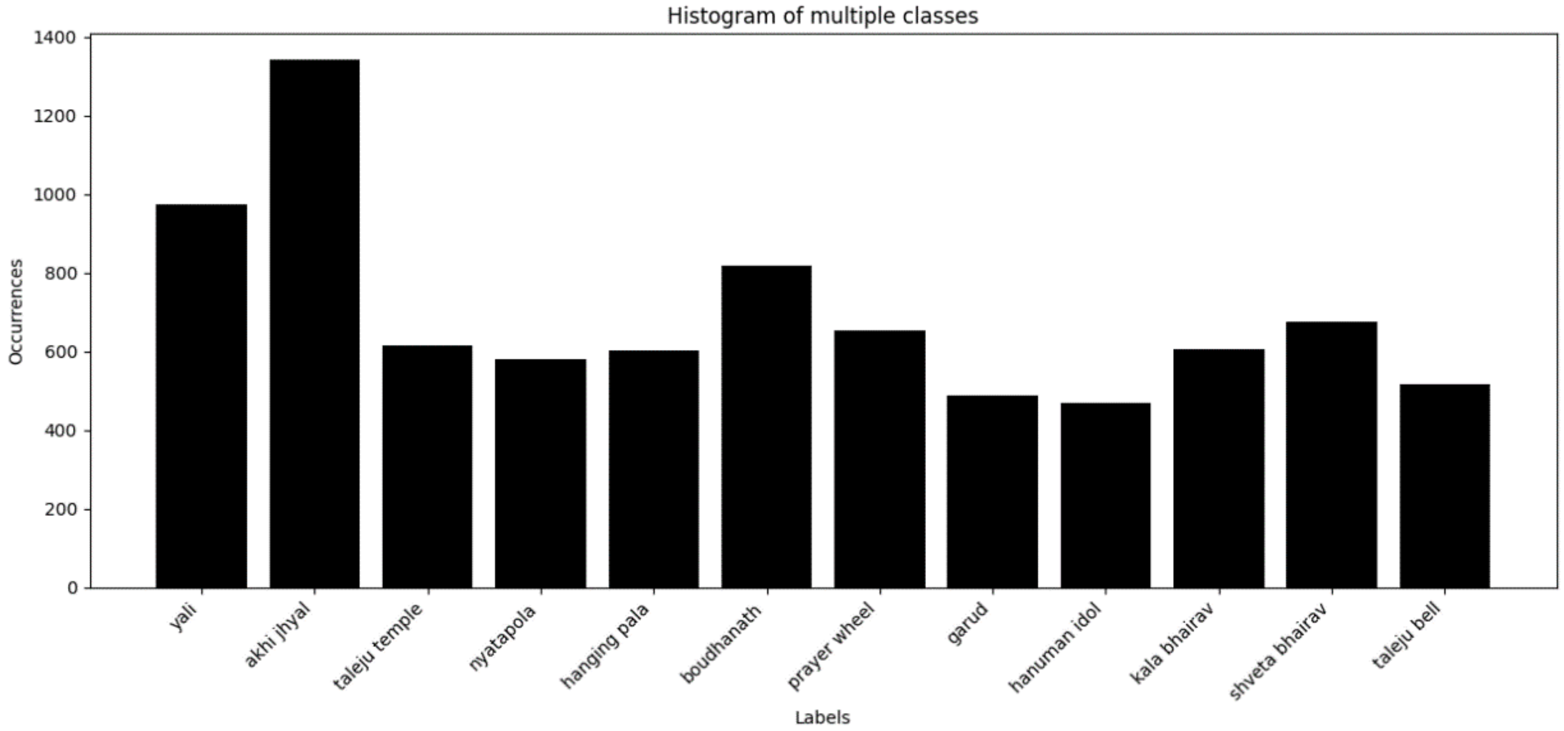


Figure: Dataset image distribution

Dataset Analysis - [2]
Data Annotation - [1]

- CVAT is used for annotation of images
- Bounding box and class labels added
- Exported in YOLO format for training

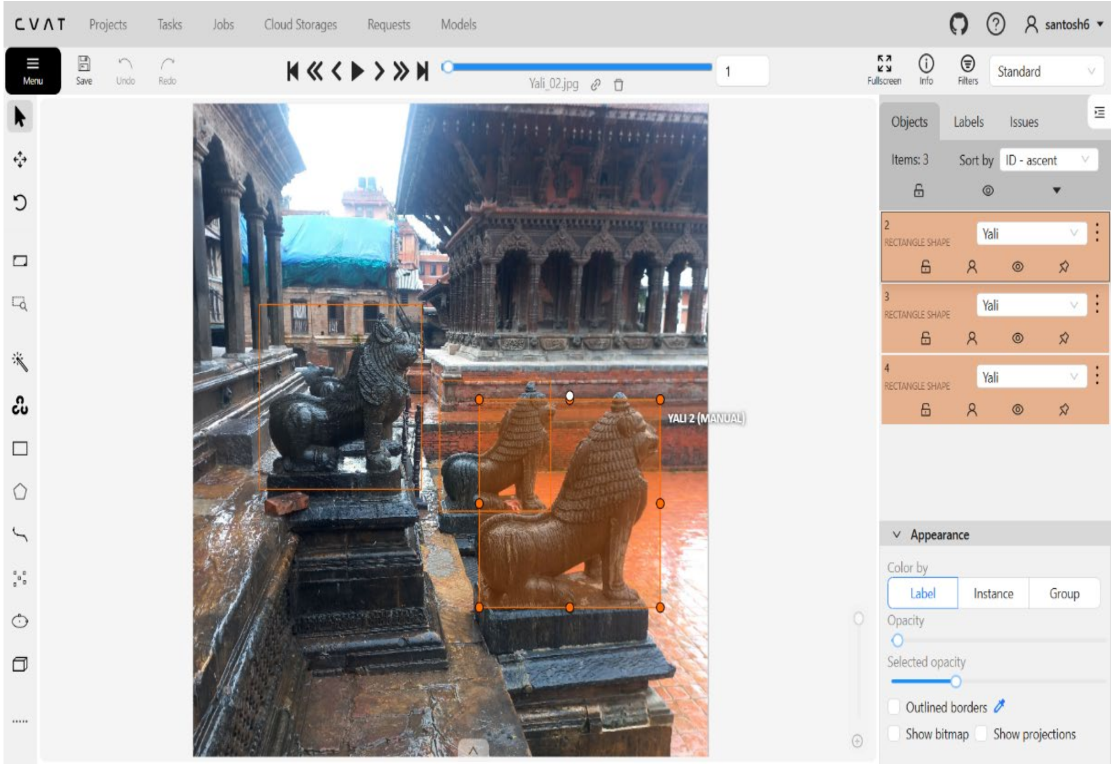


Figure: Annotation using CVAT

Dataset Analysis - [3]
Data Augmentation - [1]

- Five images generated using random combination of augmentation techniques
  - Horizontal flipping
  - Brightness and contrast
  - Gamma adjustments
  - Gaussian noise and blur
  - Rotation

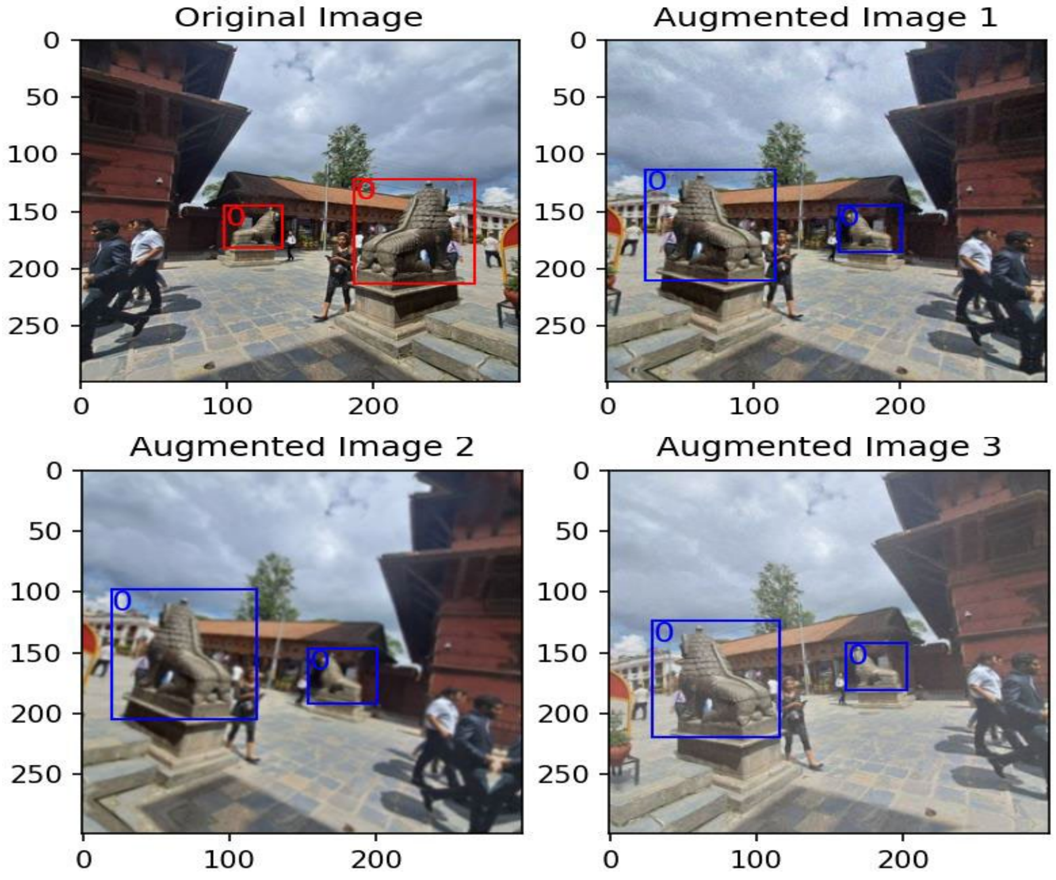


Figure: Dataset Augmentation

Dataset Analysis - [4]

- Question answering pairs has been illustrated along with about 18 QA pairs per object in average
- Around 218 question answer pair has been created till now

{} qa_0.json > [ ] qa_pairs

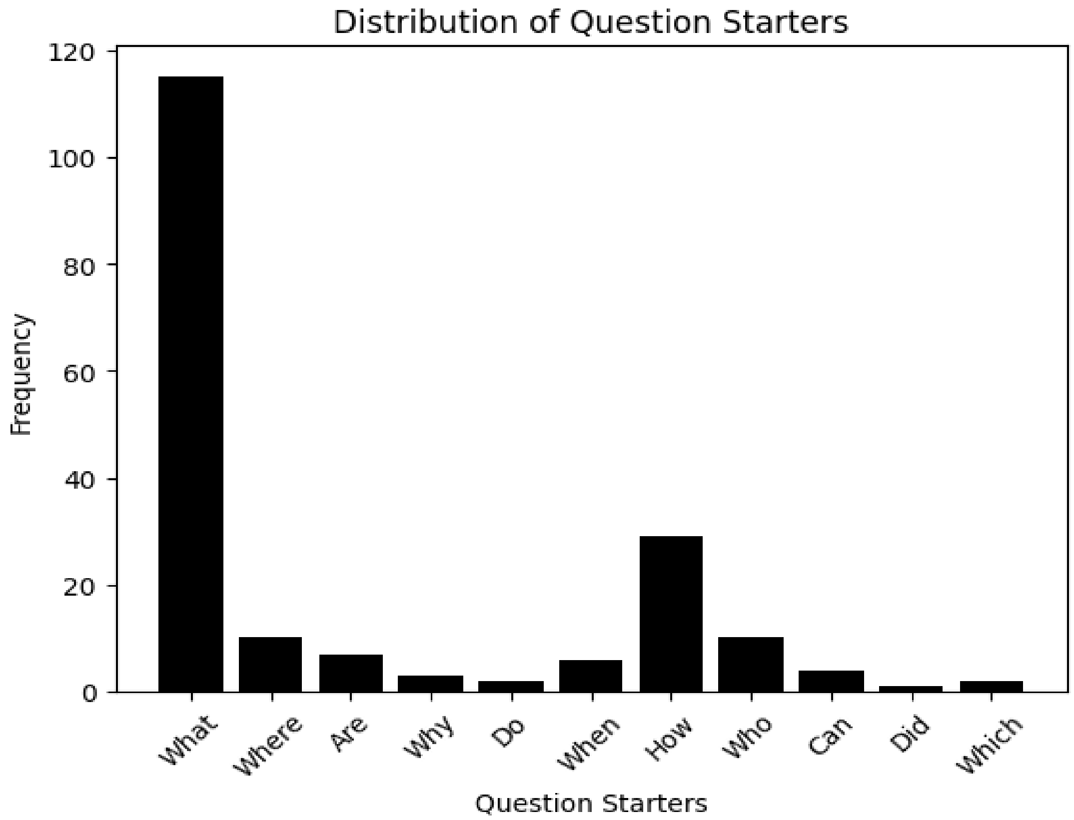

```json
{
  "image_id": 0,
  "qa_pairs": [
  {
    "question": "What is a Yali?",
    "answer": "A Yali is a mythological creature with features of multiple animals."
  },
  {
    "question": "What is the origin of Yali?",
    "answer": "Yali originates from ancient Hindu mythology."
  },
  {
    "question": "Where are these sculptures found?",
    "answer": "They are often sculpted onto the pillars and entrance gates of Hindu temples."
  },
  {
    "question": "What do Yalis symbolize?",
    "answer": "Yalis symbolize protection and man's struggle with nature."
  },
```

```json
{
  "image_id": 4,
  "qa_pairs": [
  {
    "question": "What is a prayer wheel?",
    "answer": "A prayer wheel is a cylindrical wheel used for Buddhist recitation."
  },
  {
    "question": "Where are prayer wheels commonly found?",
    "answer": "Prayer wheels are common in Tibet and areas with Tibetan culture."
  },
  {
    "question": "What is usually written on a prayer wheel?",
    "answer": "The mantra Om mani padme hum is commonly written on prayer wheels."
  },
  {
    "question": "What materials are prayer wheels made from?",
    "answer": "They can be made from metal, wood, stone, leather, or cotton."
  },
```


Figure: Question Answer Pairing

Dataset Analysis - [5]

- The graphs representing the question types and most frequently used words in the dataset



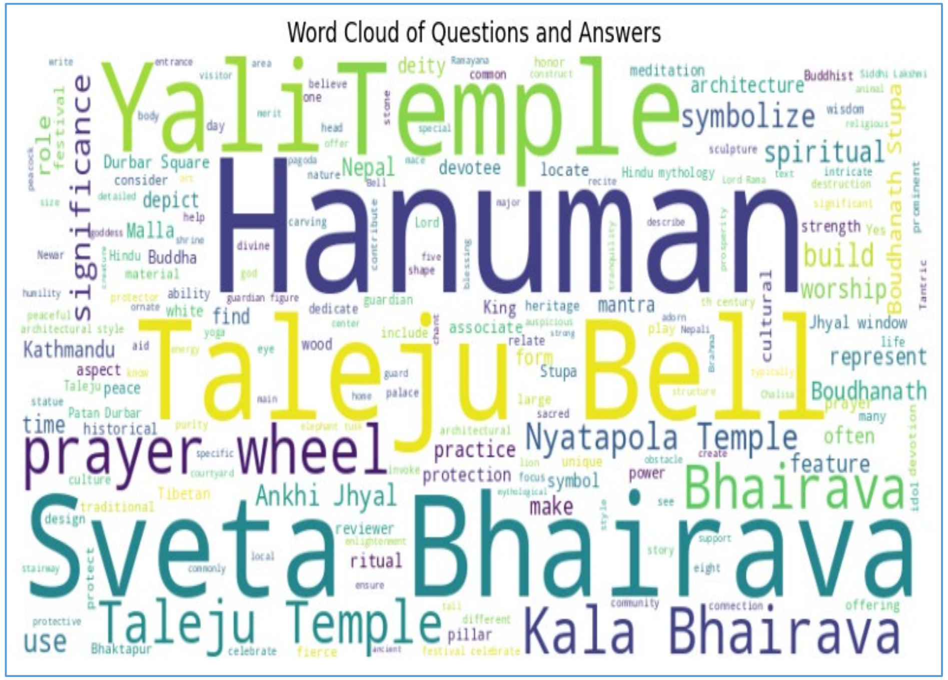


Figure: Graph of question answer types

Results - [1]

Object Detection Module - [1]

- The inference results of YOLOv8 are:

Figure: YOLOv8 inference results

Results - [1]

Object Detection Module - [2]

- Losses and metrics curve

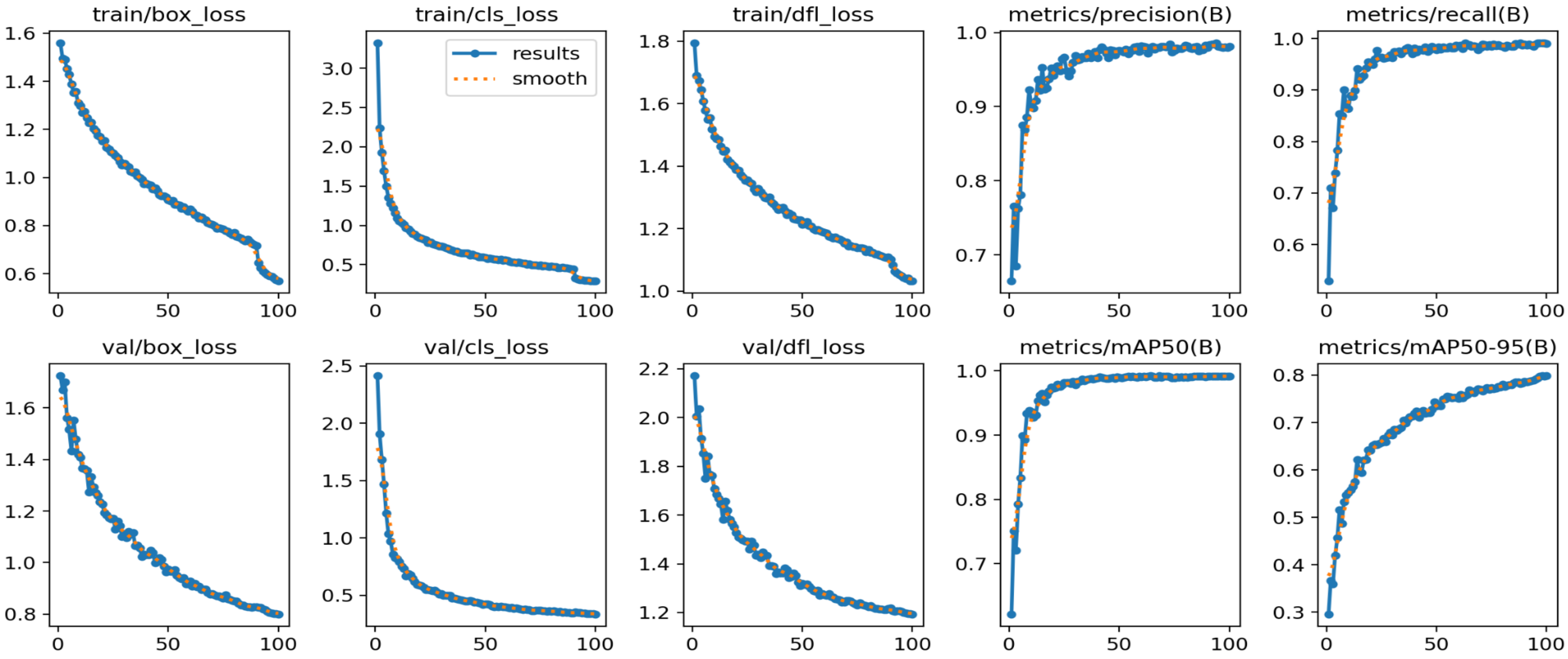


Figure: Losses and metrics curve

Object Detection Module - [3]

Results - [1]

- Precision Recall curve

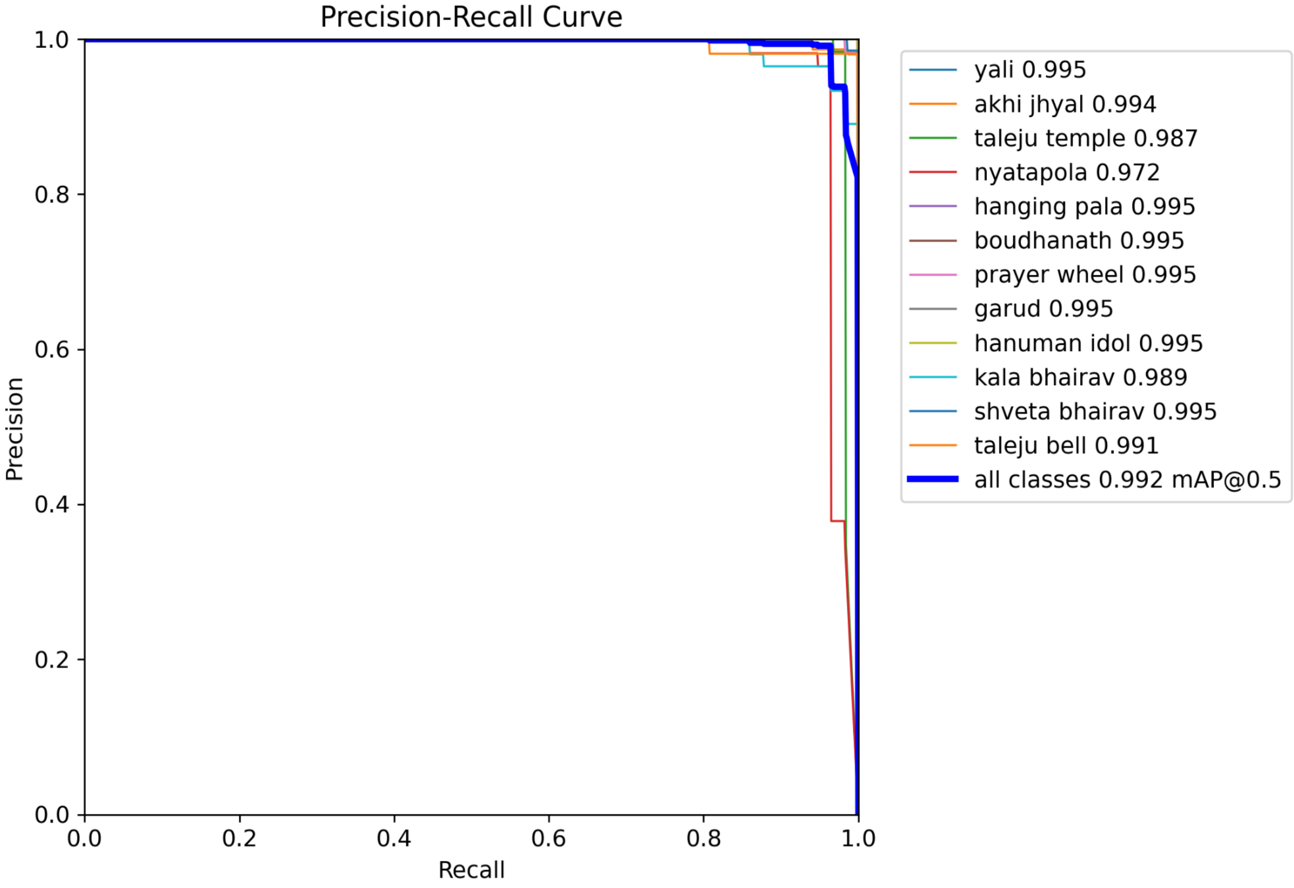


Figure: Precision recall curve

Object Detection Module - [4]

Results - [1]

- F1-score curve

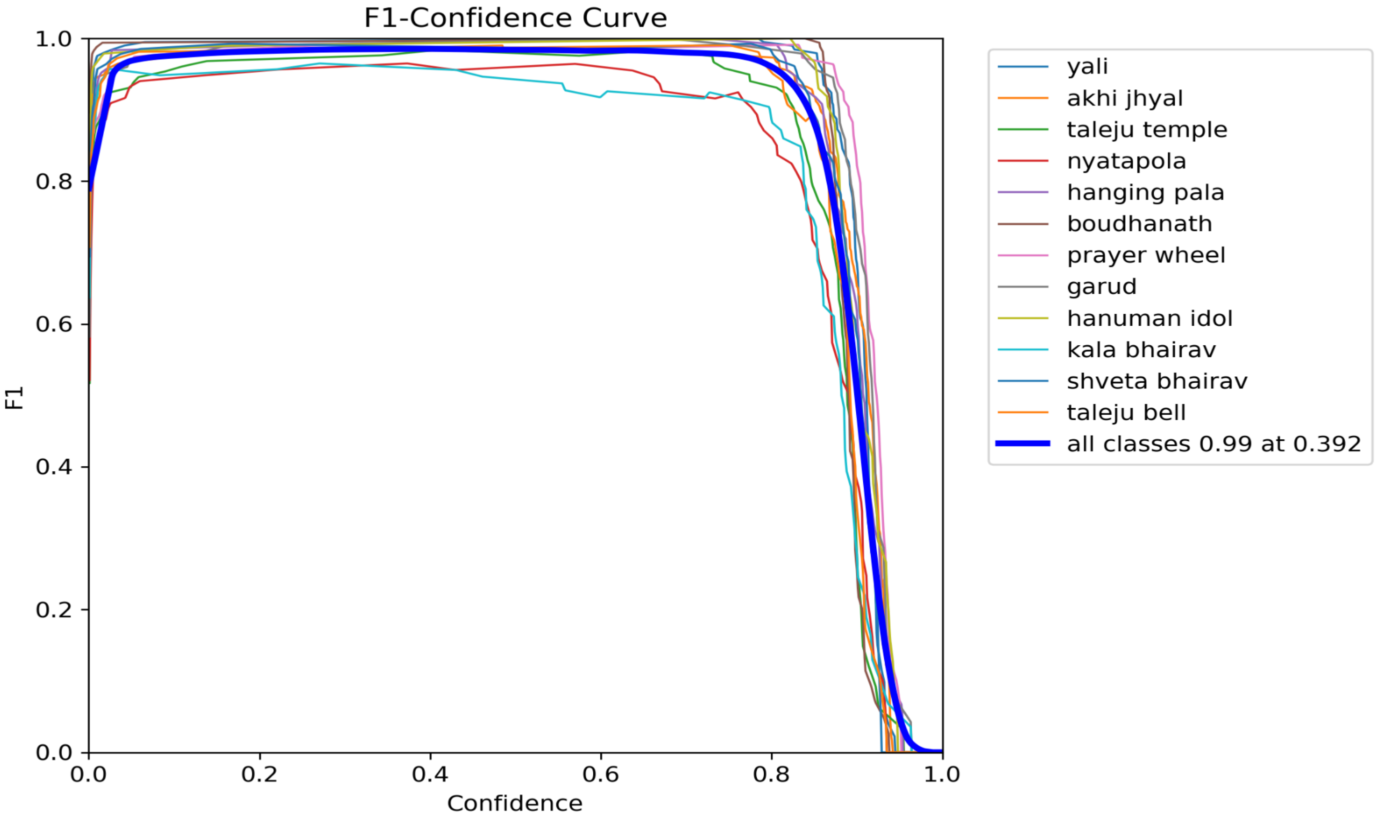


Figure: F1 score curve

Results - [1]

Object Detection Module - [5]

- Confusion Matrix

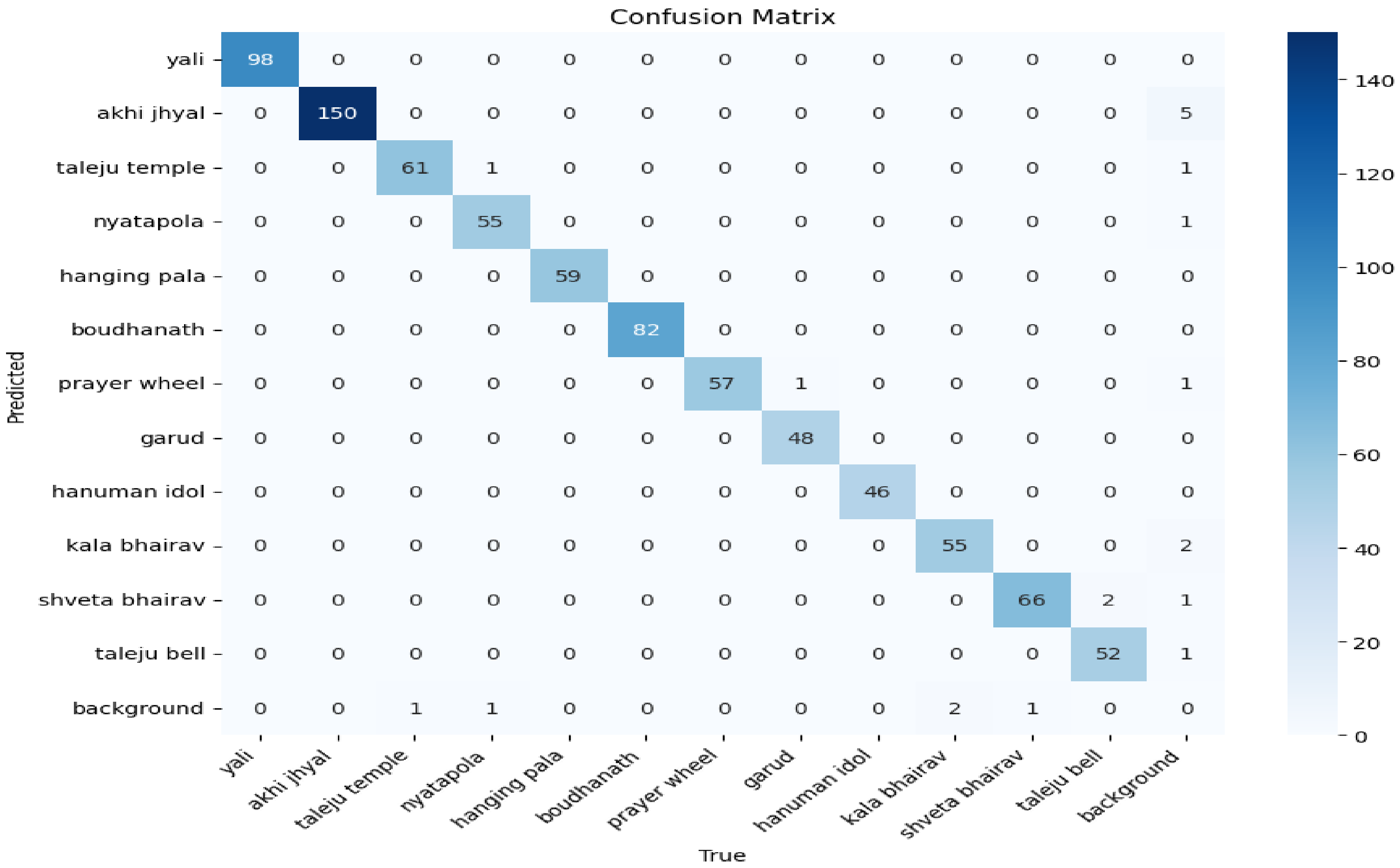


Figure: Confusion Matrix

Android Application

Results -[2]

- Application Interface

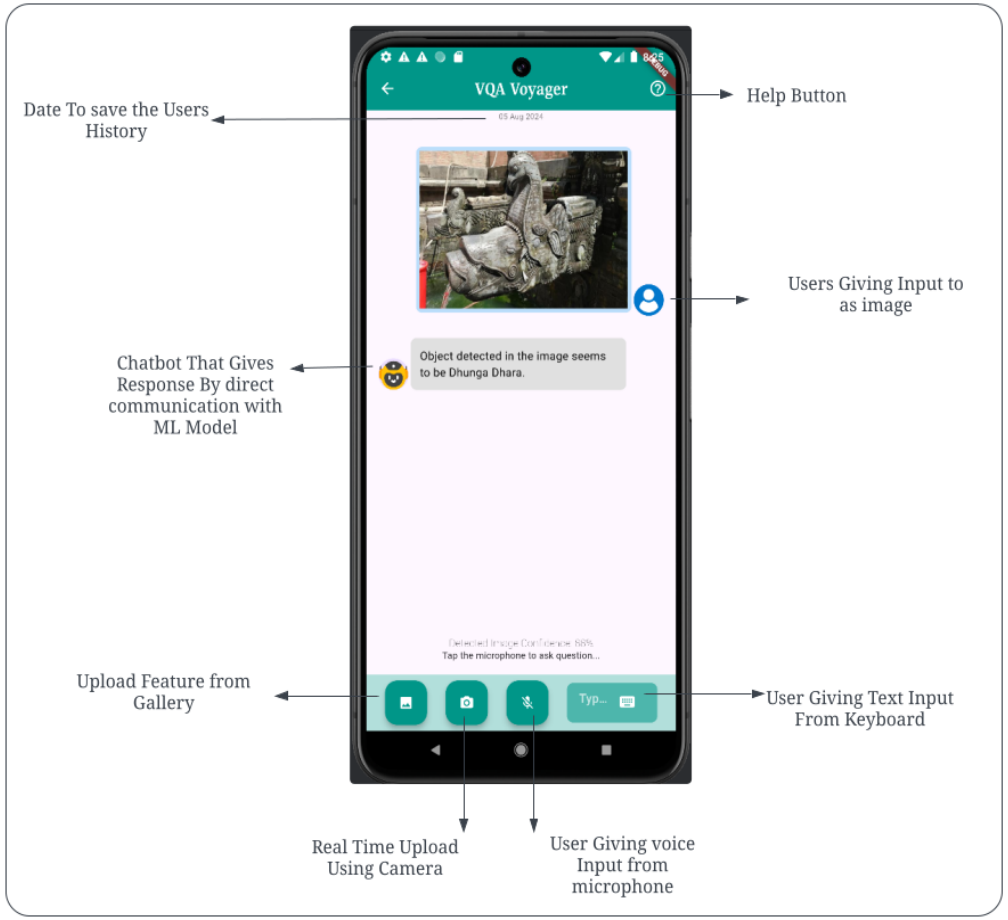


Figure: Application Interface

Rasa Chatbot

Results -[1]

- Chatbot Integration

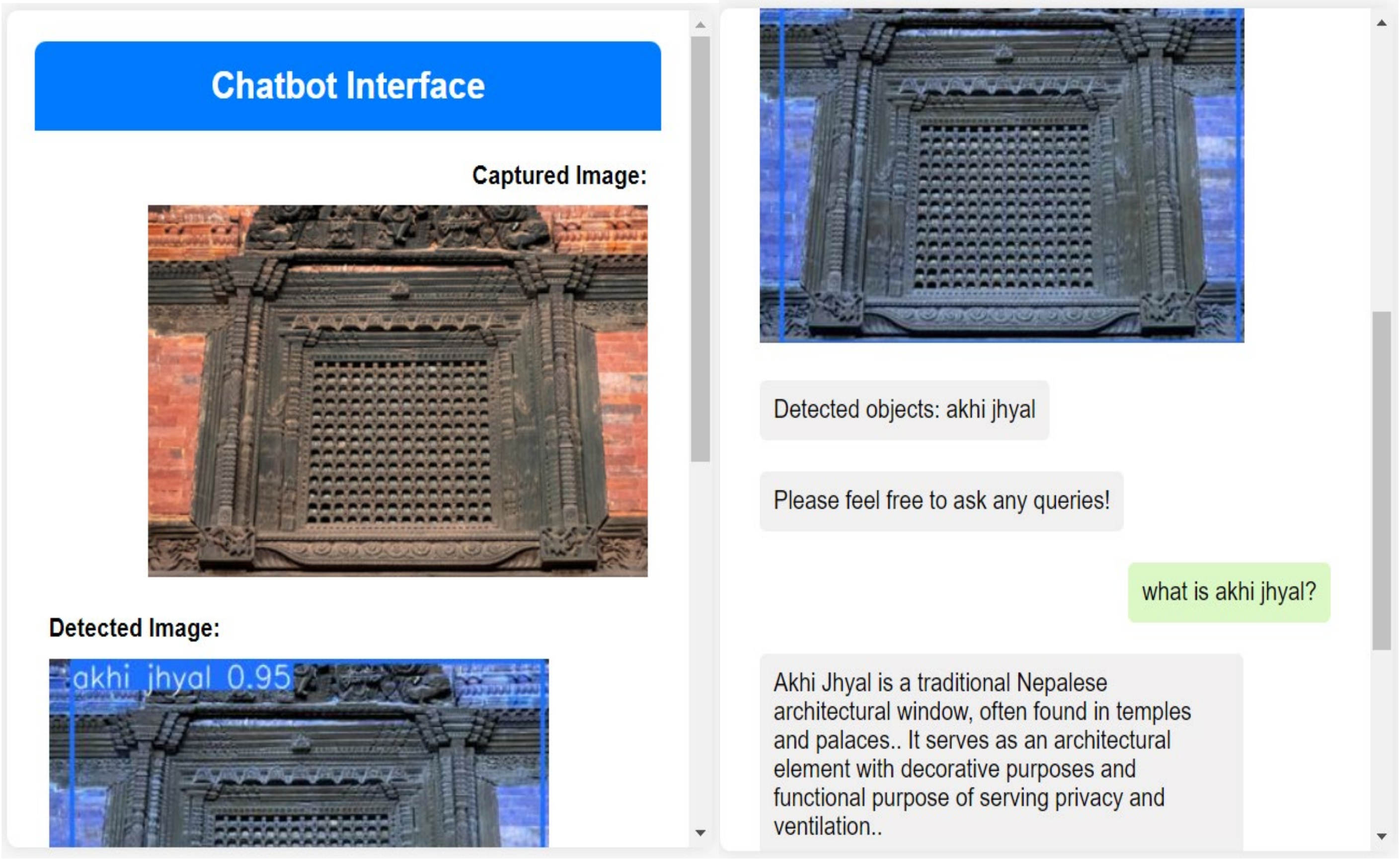


Figure: Chatbot Interface

Discussion and Analysis - [1]

- The maximum values of evaluation metrics were obtained to be as follows:
  - Precision: 98.62%
  - Recall: 99.17%
  - mAP50: 99.23%
  - mAP90: 79.89%
- For android “Speech_to_text:6.6.2” package is used for voice to text generation whereas “image_picker” package is used to open gallery and camera

Discussion and Analysis - [2]

- 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

- Augment and increase the size of QA-pairs
- Train and test the ViLT encoder and BART decoder on the custom dataset
- Create RESTful API
- Implement contextual management in chat
- Host the above-trained model in a server for Real-time Communication

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.