OMNIVISION: Image Captioning Model

TEAM NO.: 496

NAMES OF THE STUDENTS WHO PARTICIPATED IN THE TEAM:

AKSHAY KUSHAWAHA, (SECOND YEAR)

MUDIT GARG, (SECOND YEAR)

PRATHAM PANDEY, (SECOND YEAR)

ABHAY MISHRA, (THIRD YEAR)

SIDDHANTH SHAH, (THIRD YEAR)

COLLEGE: SYMBIOSIS INSTITUTE OF TECHNOLOGY

DEPARTMENT: COMPUTER SCIENCE AND ENGINEERING

CITY: PUNE

STATE: MAHARASHTRA

PROJECT MENTOR NAME: MR. MD ABDUL AZIZ

Vodafone Idea Foundation

_VOIS

Connecting Dreams Foundation

Project Details:

A High Accuracy Image captioning model with native language audio output to assist visually impaired people.

- AI / ML-based smart object detection and real-time text-to-speech engine to provide audio output.
- Capturing images and processing in real-time to detect the objects and provide audio output for the same.
- High Accuracy and speed of detection are critical for the intended purpose.

Problem Statement:

Visually Impaired persons face a lack of independence in their daily life resulting in a lack of confidence and unemployability. The current project aims to change this **dependence into independence** and provide them with the possibility of employment.

Need of Project:

- Sense of **independence** in Visually impaired persons in their daily life.
- Create something useful for the society solving a major problem
- Trained model can be used to create other supplementary models to help solve other problems.
  Ex: assisting police and doctors, crowd counting, etc

Proposed Solution:

- AI Model for Object detection – detecting multiple objects in the images
- Machine Learning Library for Objects identification – identifying each detected object and describing it in text format
- Text to Speech engine for Audio Output - converting each identification into an audio mode

Technology Used:

Python

HTML, CSS, Bootstrap

Flask

Python Libraries: Tensorflow, Keras, Numpy, gTTS, Pandas, Matplotlib

Project Outcomes:

A trained model which can predict the caption for an entered image describing it in just a sentence

Text to speech integration

Support for use in native language

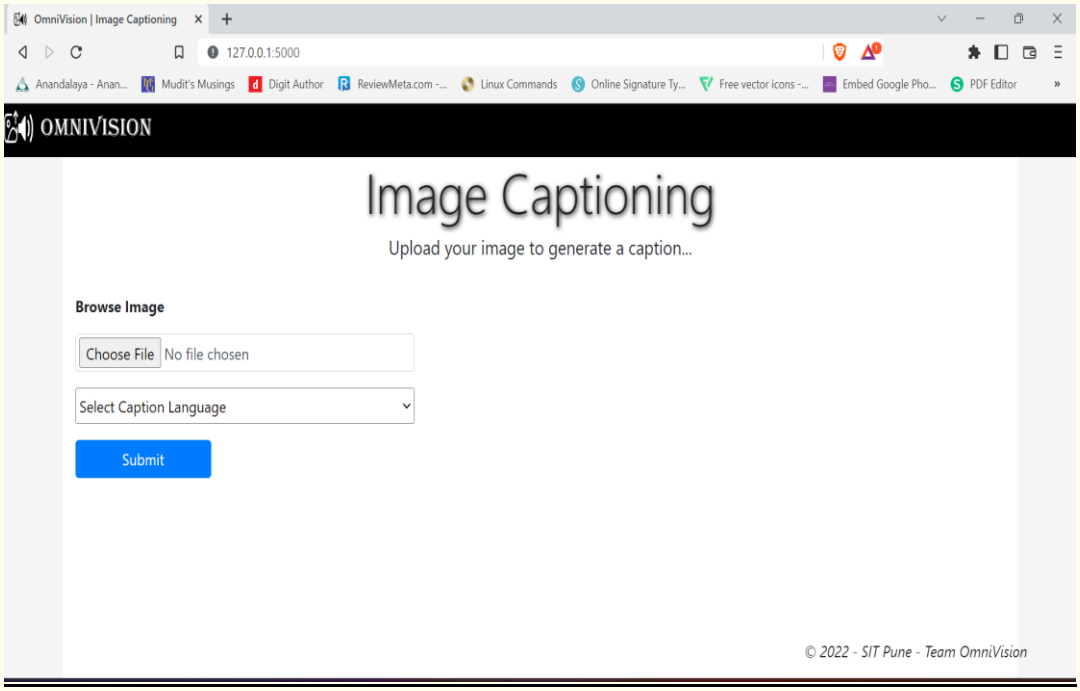
Modelling:

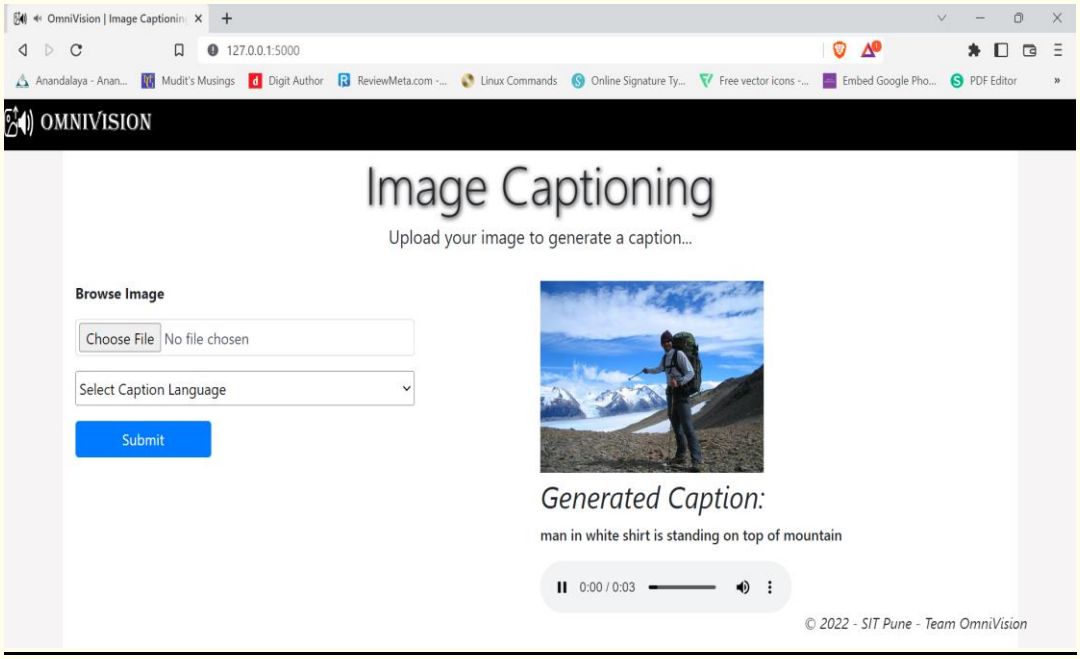
CNN LSTM – ENCODER DECODER MODEL

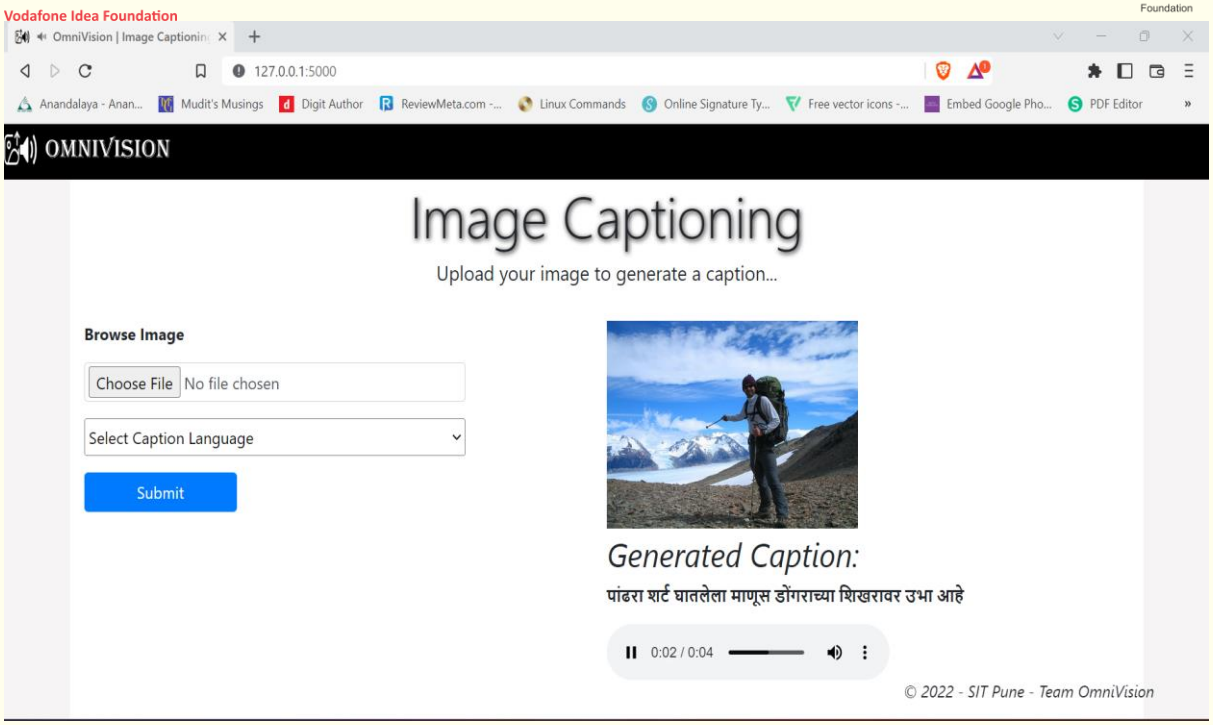
CNN MODEL- used to extract feature vectors of the input image.

LSTM MODEL- generates captions for the images after extracting features taking into consideration the state of the previous output and the present cell's input for the current output.

Results:







Future scope for project enhancement:

- Currently, as we are in the learning stage, we have just made a simple web interface as a prototype.
- converting web interface to a desktop/mobile application
- accessing the webcam to get the images from the live feed and generate real-time captions
- integrate video captioning
- provide better real-time knowledge to the user about the depth of the object, obstacles(stones, potholes, etc) at ground level