CAPTION-BOT FOR ASSISTIVE VISION

A REPORT ON PROJECT BASED LEARNING-II(SEMESTER-IV)

Submitted by-

E2 Group 2

NAMES OF THE CANDIDATES

1) **ATHARVA WALAKE (21205)**
2) **NIKHIL AWATADE (21206)**
3) **AKSHAY BACHKAR (21207)**
4) **TEJAS BENDKULE (21208)**

SECOND YEAR ENGINEERING

Society for Computer Technology and Research's

PUNE INSTITUTE OF COMPUTER TECHNOLOGY

DHANKAWADI, PUNE-43

A.Y. 2021-22

-CERTIFICATE –

This is to certify that the work incorporated in the report entitled **"CAPTION BOT FOR ASSISTIVE VISION**" is carried out by **Mr. Awatade Nikhil Nandkumar (Roll No. 21206) ,**who is part a group of students under the subject ***Project Based Learning-II*** during A.Y. **2021-2022 .**

Such material has not been submitted to any other University/ Institute for any financial support. The literature related to the problem investigated has been appropriately cited and duly acknowledged wherever facilities and suggestions have been availed of.

Date: **Name & Sign of Project Guide**

Place: PUNE

Name & Sign of PBL Coordinator **Name & Sign of Head of Department**

Prof. Parag Jambhulkar **Dr. Geetanjali Kale**

ABSTRACT

Image Captioning is a phenomenon that describes an image in the form of text. It is mainly used in applications where one needs information from any particular image in a textual format automatically. This project overcomes limitations by generating entire captions for describing pictures, which can tell detailed, unified stories. It develops a model that decomposes both images and paragraphs into their constituent elements, sleuthing linguistics regions in images with the help of LSTM model and NLP technique. It also presents the implementation of LSTM Method with additional features for a good performance. Gated Recurrent Unit (GRU) Method and LSTM Method are evaluated in this paper. According to the evaluation using BLEU Metrics LSTM is identified as the best method with 80% efficiency. This approach improves on the best results on the Visual Genome paragraph captioning dataset.

ACKNOWLEDGEMENT

Every undertaking like this project is impossible without the help and support of others. I would like to thank my fellow team members for working hard alongside me, encouraging me and helping me out.

We would like to express our deeply felt gratitude for Prof. Parag Jambhulkar who has been there with us every step of the way, guiding and supervising us. His valuable inputs and patient answering of our doubts have been instrumental in our second PBL journey.

We would also like to appreciate our college for giving us the opportunity to use our imaginations and explore different possibilities and ideas.

Place: Name of Student

PUNE NIKHIL AWATADE

TABLE OF CONTENT

INTRODUCTION

Training computers to be able to automatically generate descriptive captions for images is currently a very hot topic in Computer Vision and Machine Learning. This task is a combination of image scene understanding, feature extraction, and translation of visual representations into natural languages. This project shows some great promises such as building assistive technologies for visually impaired people. Three authors have equal contribution and listed by alphabetic order and help automating caption tasks on the internet. There are a series of relevant research papers attempting to accomplish this task in last decades, but they face various problems such as grammar problems, cognitive absurdity and content irrelevance. However, with the unparalleled advancement in Neural Networks, some groups started exploring Convolutional Neural Network and recurrent neural network to accomplish this task and observed very promising results. The most recent and most popular ones include Show and Tell: A Neural Image Caption Generator and Show, attend and tell: Neural image caption generator with visual attention. While both papers propose to use a combination of a deep Convolutional Neural Network and a Recurrent Neural Network to achieve this task, the second paper is built upon the first one by adding attention mechanism. As shown in Figure 1, this learnable attention layer allows the network to focus on a specific region of the image for each generated word.

MOTIVATION:

Having the ability to see is a wonderful gift. Individuals with vision are able to see and interpret the environment around them. Visual deficit is a state in which one is unable to recognise objects due to physiologic or neurological factors. Visual impairment can make it difficult for persons to carry out day-to-day tasks. As per a recent estimate, 253 million people worldwide suffer from visual loss. There are 36 billion people who are blind, and 217 million people who have moderate to severe vision problems. Individuals and their family suffer tremendously as a result of their loss of sight. This system is designed to help visually impaired persons live a more independent life. It is integrated with cutting-edge technology and is designed to let the visually impaired lead a life free of limitations. This is a visual-based application with a few

major components such as a camera, a system containing OpenCV, and speakers connected together, as well as additional internet-based working techniques. The project's input is an image/video (several frames), which will be captured and analysed using a camera connected to the computer. As a result, the object
gets identified, and audible data is transmitted to the blind individual through speakers or headphones.

OBJECTIVE:

To apply ml algorithms on the set of large data (images and their information) and predict captions for pictures on the basis of analyzed data. Basically, our project is all about capturing image and converting that image into text form as caption and further conversion text to speech.

The most recent and most popular ones include Show and Tell: A Neural Image Caption Generator and Show, attend and tell: Neural image caption generator with visual attention . While both papers propose to use a combination of a deep Convolutional Neural Network and a Recurrent Neural Network to achieve this task, the second paper is built upon the first one by adding attention mechanism. As shown in Figure 1, this learnable attention layer allows the network to focus on a specific region of the image for each generated word.

SCOPE OF PROJECT:

Image caption, automatically generating natural language descriptions according to the content observed in an image, is an important part of scene understanding, which combines the knowledge of computer vision and natural language processing. Automatic Captioning can help, make Google Image Search as good as Google Search, as then every image could be first converted into a caption and then search can be performed based on the caption.

CCTV cameras are everywhere today, but along with viewing the world, if we can also generate relevant captions, then we can raise alarms as soon as there is some malicious activity going on somewhere. This could probably help reduce some crime and/or accidents.

Self driving cars :: Automatic driving is one of the biggest challenges and if we can properly caption the scene around the car, it can give a boost to the self-driving system.

INTENDED AUDIENCE:

The project does not intend to only help visually impaired people but also helpful for research and students interested in the same domain development. The Algorithm gives different perspective of an image, we humans have various perspectives for same images.
This project can also be helpful for automated robots for finding objects or guiding the costumer in malls, research labs etc

OVER ALL DESCRIPTION

Functional Requirements:

1] LSTM model- LSTM stands for Long short-term memory; they are a type of RNN which is well suited for sequence prediction problems. Based on the previous text, we can predict what the next word will be. It has proven itself effective from the traditional RNN by overcoming the limitations of RNN which had short term memory. LSTM can carry out relevant information throughout the processing of inputs and with a forget gate, it discards non-relevant information.

2] Recurrent Neural Network- (RNN) are a type of Neural Network where the output from previous step is fed as input to the current step. In traditional neural networks, all the inputs and outputs are independent of each other, but in cases like when it is required to predict the next word of a sentence, the previous words are required and hence there is a need to remember the previous words. Thus, RNN came into existence, which solved this issue with the help of a Hidden Layer. The main and most important feature of RNN is Hidden state, which remembers some information about a sequence.

3] gTTS library -(Google Text-to-Speech), a Python library and CLI tool to interface with Google Translates text-to-speech API. Write spoken mp3 data to a file, a file-like object (byte string) for further audio manipulation, or stout. Or simply pre-generate Google Translate TTS request URLs to feed to an external program.

4]TensorFlow library - TensorFlow is an end-to-end open-source platform for machine learning. It has a comprehensive, flexible ecosystem of tools, libraries and community resources that lets researchers push the state-of-the-art in ML and developers easily build and deploy ML powered applications.

5] Keras - Keras is an API designed for human beings, not machines. Keras follows best practices for reducing cognitive load: it offers consistent & simple APIs, it minimizes the number of user actions required for common use cases, and it provides clear & actionable error messages. It also has extensive documentation and developer guides. Keras is the most used deep learning framework among top-5 winning teams on Kaggle. Because Keras makes it easier to run new experiments, it empowers you to try more ideas than your competition, faster. And this is how you win.

Non-Functional Requirements:

Compatibility:

The website will work smoothly on machines with Pentium-4 processor and higher. The machine should have at least 128MB of free RAM to load the Web Browser and thus the website.

Usability:

The website is made by keeping in mind the ease of use. The website is user friendly and can be used easily as any normal website available on the internet

Operating Environment:

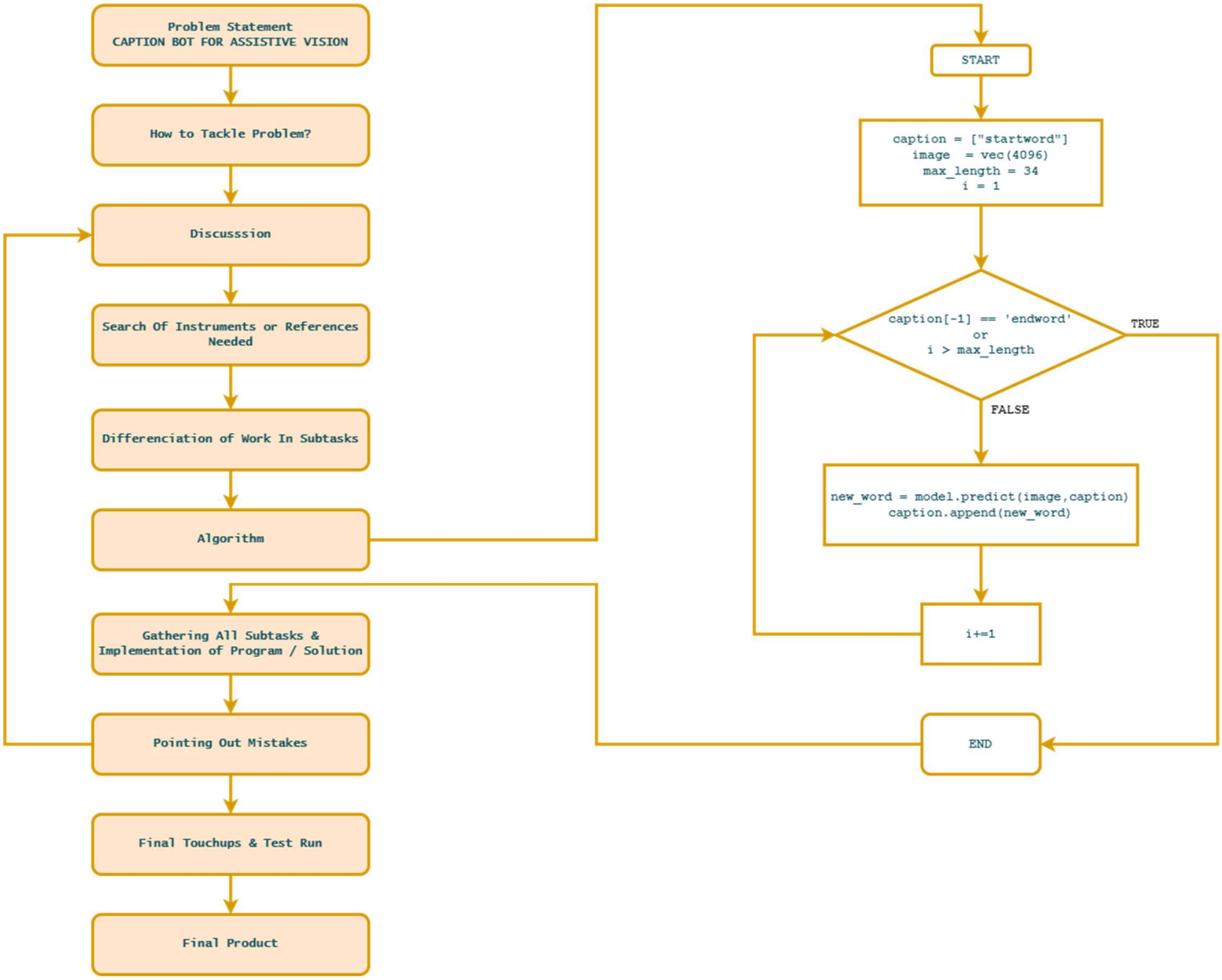
Hardware Requirements:(Minimum Requirement for a laptop)

- Audio Speakers
- Active Internet Connection
- RAM: 4GB
- Processor: Intel i5 10th Generation

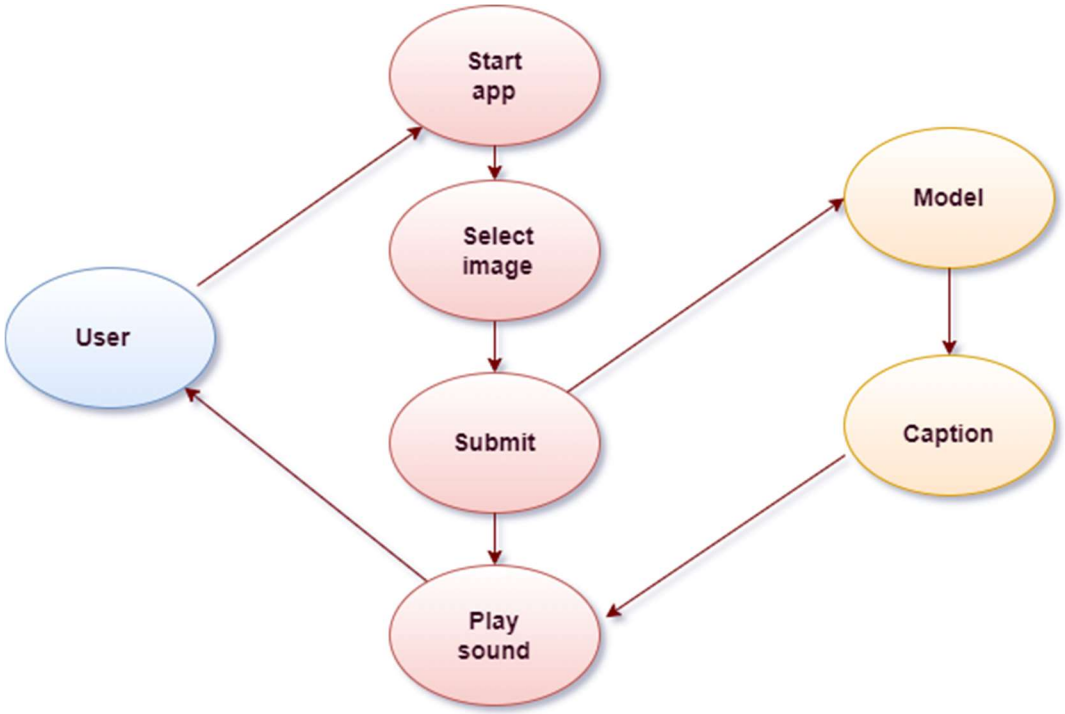
Software Requirements

- Web Browser: Any modern day internet browser such as Google Chrome, Microsoft Edge, Mozilla Firefox, Brave Browser
- VS Code: For writing and editing the source code
- Git and GitHub: for version control

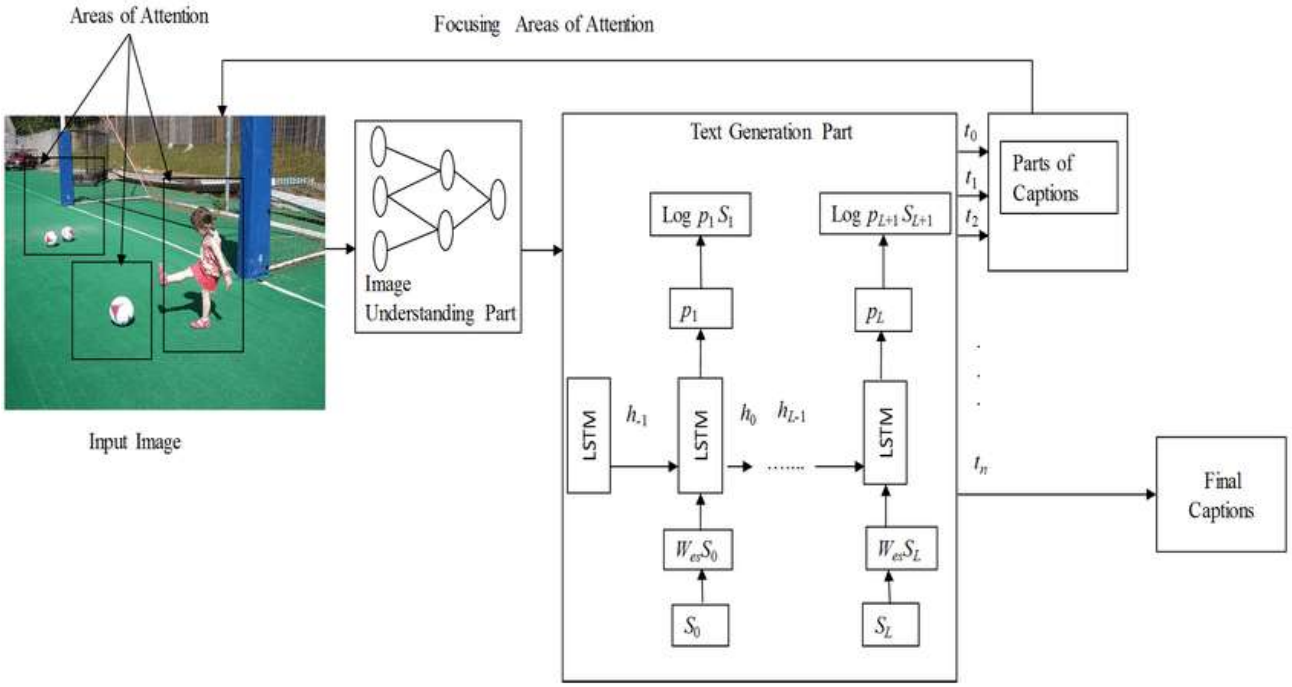
FLOW CHART


Problem Statement
CAPTION BOT FOR ASSISTIVE VISION
How to Tackle Problem?
Discusssion
Search Of Instruments or References Needed
Differenciation of Work In Subtasks
Algorithm
Gathering All Subtasks & Implementation of Program / Solution
Pointing Out Mistakes
Final Touchups & Test Run
Final Product
START
caption = ["startword"]
image = vec(4096)
max_length = 34
i = 1
caption[-1] == 'endword'
or
i > max_length
TRUE
FALSE
new_word = model.predict(image,caption)
caption.append(new_word)
i+=1
END

USE-CASE DIAGRAM



ARCHITECTURE



IMPLEMENTATION

- **FRONTEND:**

  Frontend consist of 2-page web interface. 1st page includes the image input mechanism, the next page (target page) includes caption display slot and text to speech (audio) tag. Languages used are HTML, CSS, and JavaScript.

- **INTEGRATION:**

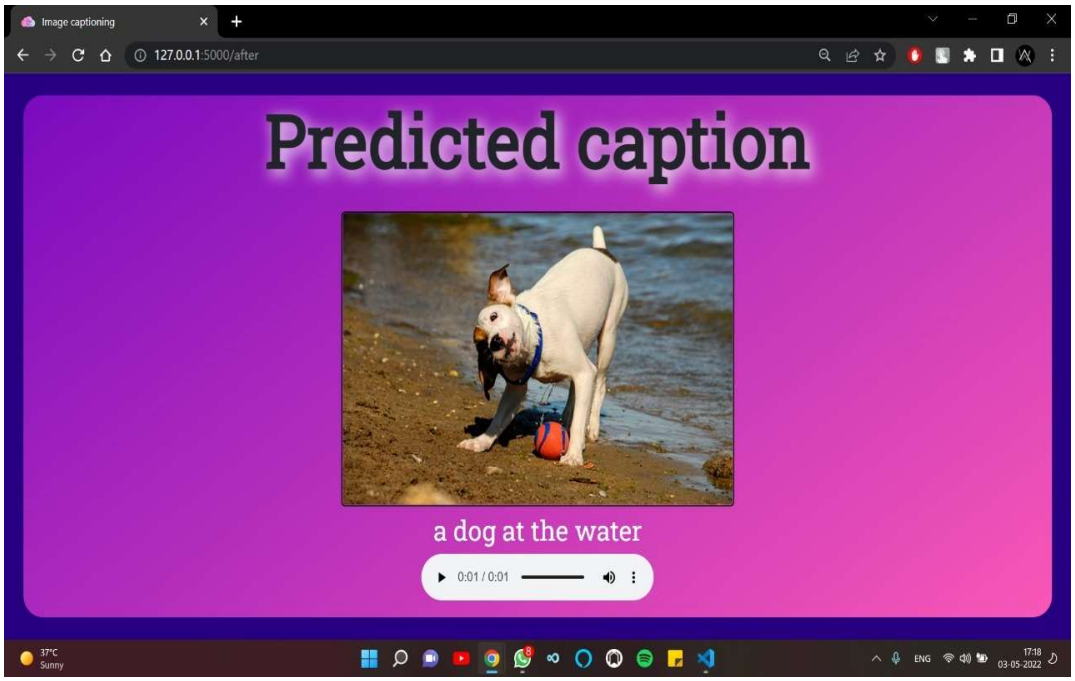
  Integration of frontend and backend is done using flask framework.
  Flask is a web framework, it's a Python module that lets you develop web applications easily.It was developed by Armin Ronacher, who led a team of international Python enthusiasts called Poocco.

- **BACKEND TECHNOLOGIES:**

  - Image recognition
  - Natural Language Processing
  - Convolutional Neural Networking (CNN).
  - LSTM based models.
  - Language models.
  - Open CV.
  - Transfer learning to use pre trained bracenet.
  - Data pre-processing .

- **Interface visuals:**

- **Implementation screen shots:**



DEMO Video :: https://youtu.be/2ISwjqWcR0Q

CONCLUSION

According to WHO, the number of visually impaired people of all ages worldwide is estimated to reach 285 million, with 39 million of them blind . People aged 50 and up account for 82% of all blind people. Uncorrected refractive errors (43%) and cataract (33%) are the leading causes of visual impairment; cataract is the leading cause of blindness (51%) . Our proposed plan enables this percent of the multitude to lead an independent and almost normal lifestyle and to ultimately aid them in situations were

there's nobody around. The suggested concept uses live photos to detect objects and provide captions that can be turned to voice to help the visually impaired understand what is going on around them. Subsequently, this technique can also be used on text images,

allowing a visually impaired individual to recognize text on drugs and other things with text. Furthermore, with the help of a virtual text reader, they would be able to obtain knowledge about current events as well as enjoyment.

REFERENCES

1. Programming Hub YouTube channel- click here
https://www.youtube.com/watch?v=-cT1m6NZYWc

2.Aditya Jain GitHub repo. - click here
https://github.com/adityajn105/image-caption-bot

3.Image dataset - click here
https://www.kaggle.com/datasets/srbhshinde/flickr8k-sau

4. Training model - www.kaggle.com