RSET
RAJAGIRI SCHOOL OF
ENGINEERING & TECHNOLOGY
(AUTONOMOUS)

Mini Project Report On

Pics-Talk

Submitted in partial fulfillment of the requirements for the award of the degree of

Bachelor of Technology

in

Computer Science & Engineering

By

Ajith Varghese Abraham(U2103018)

Anna Prince (U2103040)

Arjun Martin (U2103047)

Under the guidance of

Ms.Seema Safar

Department of Computer Science & Engineering
Rajagiri School of Engineering & Technology (Autonomous)
(Affiliated to APJ Abdul Kalam Technological University)
Rajagiri Valley, Kakkanad, Kochi, 682039
May 2024

CERTIFICATE

This is to certify that the mini project report entitled ***"Pics-Talk"*** *is a bonafide record of the work done by* ***Ajith Varghese Abraham (U2103018)***, ***Anna Prince(U2103040)***, ***Arjun Martin (U2103047)***,*submitted to the APJ Abdul Kalam Technological University in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology (B. Tech.) in Computer Science and Engineering during the academic year 2023-2024.*

Ms.Seema Safar
Asst. Professor
Dept. of CSE
RSET

Mr.Harikrishnan M
Asst. Professor
Dept. of CSE
RSET

Dr.Preetha K G
Head of Department
Dept. of CSE
RSET

ACKNOWLEDGEMENTS

I wish to express my sincere gratitude towards Dr P. S. Sreejith, Principal of RSET, and Dr. Preetha K.G., Head of the Department of Computer Science and Engineering for providing me with the opportunity to undertake my mini project, "Pics-Talk".

I am highly indebted to my project coordinators, **Mr.Harikrishnan M**, Asst. Professor, Department of Computer Science and Engineering and **Ms.Sherine Sebastian**, Asst. Professor, Department of Computer Science and Engineering for their valuable support.

It is indeed my pleasure and a moment of satisfaction for me to express my sincere gratitude to my project guide **Ms.Seema Safar**, Asst. Professor, Department of Computer Science and Engineering for her patience and all the priceless advice and wisdom she has shared with me.

Last but not the least, I would like to express my sincere gratitude towards all other teachers and friends for their continuous support and constructive ideas.

Ajith Varghese Abraham

Anna Prince

Arjun Martin

Abstract

In the modern era of ever rising demand, image captioning has become one of the most widely required tools to identify an unknown object with relative ease. In this project, we propose an innovative approach for automatically generating descriptive captions for images, leveraging recent advancements in both computer vision and natural language processing. This is done by recognizing the important objects, their attributes, the relationships among the various components present in the image and thereby generating corresponding descriptions. Here, we integrate deep learning models Convolution Neural Network (CNN) based on Residual Network (RESNET) and Long Short-Term Memory (LSTM) to describe images and generate captions using computer vision and machine translation. The datasets and the programming language used are FLICKR-8K, MS COCO and Python3 respectively. This project also delves deep into the basic functioning and structure of Neural networks involved. In addition, we are also incorporating the feature of converting the generated captions into an audio file.

Contents

List of Figures

Chapter 1

Introduction

1.1 Background

Images are visual representations used for describing the physical world. Images are extensively used for conveying enormous amount of information over internet and social media. Therefore, there is an ever-rising demand for image data analysis to design efficient information processing systems. This has in turn led to the development of systems with capability to automatically analyze the details contained in an image (image recognition) and to express it in meaningful natural language sentences (caption generation).

Image recognition is the ability of computers to identify and classify specific objects, places, people, text and actions within digital images and videos. Using image recognition, a computer vision system recognizes patterns and regularities and computes numerical data corresponding to things like people or objects. It essentially automates the innate human ability to look at an image, identify objects within it and respond accordingly. As an application of computer vision, image recognition software works by analyzing and processing the visual content of an image.

Image caption generation is an integral part of many useful systems and applications such as visual question answering machines, surveillance video analyzers, video captioning, automatic image retrieval, assistance for visually impaired people, biomedical imaging, robotics and so on. A good captioning system will be capable of highlighting the contextual information in the image similar to human cognitive system. In the recent years, several techniques for automatic caption generation in images have been proposed that can effectively solve many computer vision challenges.

1.2 Problem Definition

This project aims to create a model that can recognise the components in an image and generate corresponding image descriptions.In addition, the generated captions are converted into an audio file.

1.3 Scope and Motivation

This project aims to create a system capable of automatically generating descriptive textual captions for images. This involves various tasks like including data collection, pre-processing, model architecture design, training, evaluation, integration, and deployment. These processes are based on deep learning, a subcategory of machine learning that uses multi-layered structures of algorithms called neural networks to continually analyse data and draw conclusions about it, similar to the working of the human brain. In case of image recognition, neural networks are fed with as many pre-labelled images as possible in order to teach them how to recognize similar images which implies that larger the dataset used, greater will be the accuracy of the model developed. The captions are generated corresponding to the analysed data which is in turn converted into an audio. Ultimately, the goal is to develop an efficient and scalable system capable of accurately describing the content of diverse images from large-scale datasets.

1.4 Objectives

- Collection of the required dataset.
- Identification of the different components present in an image.
- Determination of the relationship between the components in the image.
- Generation of descriptions corresponding to the analysed data.
- Conversion of the generated captions into an audio file.

1.5 Challenges

One of the most common challenges faced while developing an image captioning model is the accuracy of the captions generated.The precision of captions depends on the quality of images used and the diversity of the dataset to a large extent.

1.6 Assumptions

It is assumed that the images to be analysed will be have a good quality.It is also assumed that the data set is diverse enough to easily recognise most of the objects in the physical world.

1.7 Societal / Industrial Relevance

- Content Management: In industries dealing with vast amounts of visual content such as media, publishing, and advertising, image caption generators can automatically generate descriptions for images, making content management more efficient. It can help in organizing, categorizing, and searching through large image databases.
- Social Media Management: Social media platforms generate enormous amounts of visual content every day. Image caption generators can automatically generate captions for images uploaded by users, making content more accessible to search engines and aiding in content discovery.
- E-commerce: In e-commerce, image caption generators can automatically generate product descriptions based on product images. This can improve Search Engine Optimization by providing text-based content that search engines can index.
- Healthcare: In medical imaging, image caption generators can automatically generate descriptions for medical images such as X-rays, MRIs, and CT scans. This can aid healthcare professionals in quickly interpreting and documenting medical images, leading to more efficient diagnosis and treatment planning.

.

1.8 Organization of the Report

The contents of the Project Report are organized as follows:

Chapter 2 consists of the Software Requirements Specifications(SRS) document of the project. It includes topics such as Purpose, Overall Description, External Interface Requirements, System Features, and other Non-Functional Requirements.

Chapter 3 consists of the System Architecture and Design document of the project. It includes topics such as System Overview, Architectural Design, Identified Dataset, Proposed Methodology/Algorithms, User Interface Design, Implementation Strategies, Module Division, and Gantt Chart.

Chapter 2

Software Requirements Specification

2.1 Introduction

2.1.1 Purpose

This Software Requirements Specification (SRS) document outlines the requirements for the development of a system capable of generating descriptive captions for images using a combination of Convolutional Neural Network (CNN) for image feature extraction and Long Short-Term Memory (LSTM) for caption generation. Additionally, the system will include a feature to convert the generated captions into audio

2.1.2 Product Scope

The system aims to provide an innovative solution for enhancing the process of generation of captions based on images and later converting these captions into audio. It will cater to users who may have visual impairments or prefer an auditory form of content consumption. Furthermore, it focuses on generating captions that can be used for obtaining better descriptions of the images which can in turn be used to obtain better search results. However, the accuracy of the generated descriptions can vary depending upon various external factors like the quality and complexity of the image used.

2.2 Overall Description

2.2.1 Product Perspective

The proposed system exists within the context of image processing, natural language processing and accessibility enhancement. This model is also designed to integrate with existing applications or platforms that involve image content and processing. The system

functions as a standalone image captioning and audio conversion service but can also be integrated into broader software ecosystems.

2.2.2 Product Functions

The primary functions of the system include:

Image Caption Generation:

Convolutional Neural Networks (CNNs) is utilized for robust image feature extraction along with Long Short-Term Memory (LSTM) networks for generating descriptive captions from the extracted image features [3].The CNN-LSTM architecture exploits the CNN's capacity to capture detailed visual information and the LSTM's proficiency in sequential modeling to generate captions that exhibit both semantic coherence and contextual relevance. During training, the CNN extracts features from input images, which are then passed to the LSTM network as initial states. The LSTM sequentially generates words, conditioning each word prediction on both the extracted image features and the previously generated words.

Audio Conversion:

The generated captions are converted into audio format through a Text-to-Speech (TTS) module. The process involves the utilization of specialized modules such as gTTS (Google Text-to-Speech) and Pyttsx3 to facilitate the transformation of textual descriptions into audio format.The module operates by sending the textual input to Google's servers, where it undergoes text analysis and synthesis, resulting in the creation of audio files containing the synthesized speech corresponding to the provided captions.Additionally, the Pyttsx3 module serves as an alternative text-to-speech solution, offering flexibility and customization options. Built upon the cross-platform TTS engine, Pyttsx3 enables to generate speech directly within the Python environment. This module provides various configuration parameters for controlling aspects such as voice selection, speech rate, and volume, allowing for fine-tuning of the audio output according to specific requirements.

2.2.3 Operating Environment

The system is designed to operate in various environments:

Supported Platforms:

The model is compatible with common operating systems (Windows, Linux, macOS). Furthermore, web-based deployment is implemented for online accessibility.

Hardware Requirements:

A machine with GPU efficient for supporting CNN-based image feature extraction (Intel Core i5 or AMD Ryzen 5 and above) is the primary requirement. Further, SSDs are required to access and handle the humongous data set with ease. On top of these, primary memory of 16 Gb or above may prove to be helpful as it increases the executional efficiency of the model as a whole.

Software Dependencies:

The software used are Tensorflow, Pytorch (Deep Learning Frameworks), OpenCV (Image ProcessingLibraries), NumPy- Numerical Computing Functionalities, gTTS or Pyttxs3(Text-to-speech libraries).

2.2.4 Design and Implementation Constraints

Computational Resources:

The efficiency of the image captioning process may be constrained by the availability of computational resources, particularly during feature extraction.

Technology Limitations:

The effectiveness of the system is subject to the performance of pre-trained models and external libraries used for image processing and TTS.

2.2.5 Assumptions and Dependencies

Assumptions:

It is assumed that users will provide images in standard formats compatible with the system (e.g., JPEG, PNG). The quality and relevance of generated captions depend on the quality of pre-trained models and the diversity of the training dataset.

Dependencies:

The system relies on external libraries and frameworks for image processing (e.g., TensorFlow, PyTorch) and TTS (e.g., Google Text-to-Speech API).

2.3 External Interface Requirements

2.3.1 User Interfaces (UI)

The user interface type for this project is a web application. This ensures optimal viewing across different devices such as desktops and mobile phones. The screen layout consists of an input area dedicated to uploading images and a text output area where the generated caption will be displayed. Standard buttons and functions include an upload button, allowing users to select and upload images, and a generate caption button. The generate caption button triggers the image processing mechanism and subsequently displays the generated text output. In addition, an audio button is also present which triggers the conversion of the generated captions into speech.

2.3.2 Software Interfaces (SI)

The application is designed to be platform-independent(it can run smoothly on all major operating systems). It is primarily programmed using Python, a language commonly chosen for deep learning frameworks. For building and deploying image captioning models, the preferred deep learning frameworks are TensorFlow or PyTorch, both of which offer robust support. Additionally, the application utilizes various libraries: OpenCV for image processing, either Keras/TensorFlow orPyTorch.nn for deep learning model construction, and NumPy for handling numerical computations. The data flow of the application begins with the user uploading image data using their web browser. Subsequently, the application

utilizes a pre-trained ResNet model to extract features from the uploaded image. These features are then forwarded to the Long Short-Term Memory (LSTM) network for the generation of a caption. Then, the resulting text caption is transmitted back to the user's web browser for display. Finally, the generated captions are converted into an audio file using text-to-speech library.

2.3.3 Communications Interfaces (CI)

No external communication protocols are needed for this application. Data transfer is confined to uploading images and displaying text output within the web application. For text-to-speech conversion, the application interfaces with a text-to-speech library such as gTTS or Pyttsx3. The generated caption text is sent to the text-to-speech library for audio generation. The synthesized speech audio can then be played within the web browser using HTML5 audio elements.

2.4 System Features

2.4.1 Caption Generation

This image captioning system integrates deep learning models, Convolutional Neural Network (CNN) based on the Residual Network (ResNet) architecture and Long Short-Term Memory (LSTM) network [4]. The ResNet-based CNN extracts high-level visual features from input images, capturing details and complex representations of objects. This architecture easily combines the spatial information from the CNN with the sequential processing capabilities of the LSTM, ensuring an overall understanding of the image context. The system is trained on a large dataset, aligning visual features with machine-generated captions, by transferring the process of learning from a pre-trained ResNet-based CNN model. The user-friendly interface allows users to effortlessly upload images and receive linguistically appealing rich captions, showcasing the system's effectiveness in combining computer vision and machine translation for a comprehensive image captioning solution.

2.4.2 Text-to-Speech

In addition to image captioning capabilities, this system has an innovative feature for audio generation from texts derived from input images. Building upon the powerful

combination of a Convolutional Neural Network (CNN) based on the Residual Network (ResNet) and a Long Short-Term Memory (LSTM) network, this integrated solution offers a seamless transition from visually descriptive captions to audio output. The LSTM's processing not only contributes to generation of textual descriptions but also facilitates the conversion of these texts into audio representations, through the incorporation of a text-to-speech (TTS) [5] module. The system transforms the generated captions into English language. Users can now not only receive detailed image captions but also experience the content through audio narration, demonstrating the system's versatility in translating visual information into a more inclusive user interface.

2.5 Other Nonfunctional Requirements

2.5.1 Performance Requirements

Accuracy of Captioning:

The generated captions should be accurate and contextually relevant to the content of the input image. The system's accuracy is measured against established benchmarks and user expectations, emphasizing the importance of precise image understanding.

Audio Quality:

In audio generation, the quality of the synthesized speech is paramount. The system should produce clear, natural-sounding audio with proper intonation, pacing, and pronunciation to enhance the overall user experience.

Robustness to Noisy Inputs:

The system should exhibit robustness to noisy or imperfect input images. It should gracefully handle situations where the image quality is compromised, ensuring that the generated captions and audio remain reliable and relevant.

2.5.2 Software Quality Attributes

Reliability:

The system should consistently and accurately generate captions and audio from input images, ensuring a reliable and trustworthy user experience. Minimize errors and unexpected behaviors to enhance overall system reliability.

Accuracy:

Achieve a good amount of precision and correctness in generating captions that precisely describe the content of the input image. The audio synthesis should faithfully represent the intended speech, providing an accurate and meaningful auditory experience.

Usability:

Design a user-friendly interface that allows users to easily upload images, understand generated captions, and access audio outputs. Ensure simplicity and clarity to enhance overall usability

Chapter 3

System Architecture and Design

3.1 System Overview

The system integrates image capturing, text generation, and audio generation functionalities to enhance user experience. Leveraging the FLICKR-8K dataset for training image captioning models, TensorFlow and Keras serves as the backbone for machine learning operations. The system further employs a Text-to-Speech (TTS) library to convert generated text captions into audio.

In its architecture, the Image Capturing Module facilitates image capture through various compatible devices such as smartphones or cameras, providing users with a user-friendly interface for both taking photos and selecting images from the gallery. The Text Generation Module, trained on the FLICKR-8K dataset, employs deep learning techniques, potentially based on Convolutional Neural Networks (CNNs), to generate descriptive text captions for the captured images. Meanwhile, the Audio Generation Module converts these generated text captions into audio utilizing the chosen TTS library.

The User Interface aspect of the system offers an interface for image capturing, text generation, and audio playback. It enhances the user experience by supporting features like voice commands and compatibility with screen readers.

In terms of data flow, images captured by the user undergo processing by the text generation module, resulting in the production of descriptive captions. These captions are then passed to the audio generation module, where they are converted into audio using the TTS library. The synthesized audio can then be played back to the user or stored for future use.

The technology stack comprises TensorFlow for image processing and text generation, specifically utilizing Convolutional Neural Networks (CNNs). For text generation, TensorFlow handles model deployment while the TTS library, pyttsx3, facilitates the conversion

of text to audio. The user interface is developed using web technologies such as HTML, CSS, JavaScript, and Python.

For future enhancements, considerations include integrating additional datasets to enhance caption generation accuracy, expanding language support for both text generation and audio synthesis, and incorporating real-time object recognition to enable more context-aware captions.

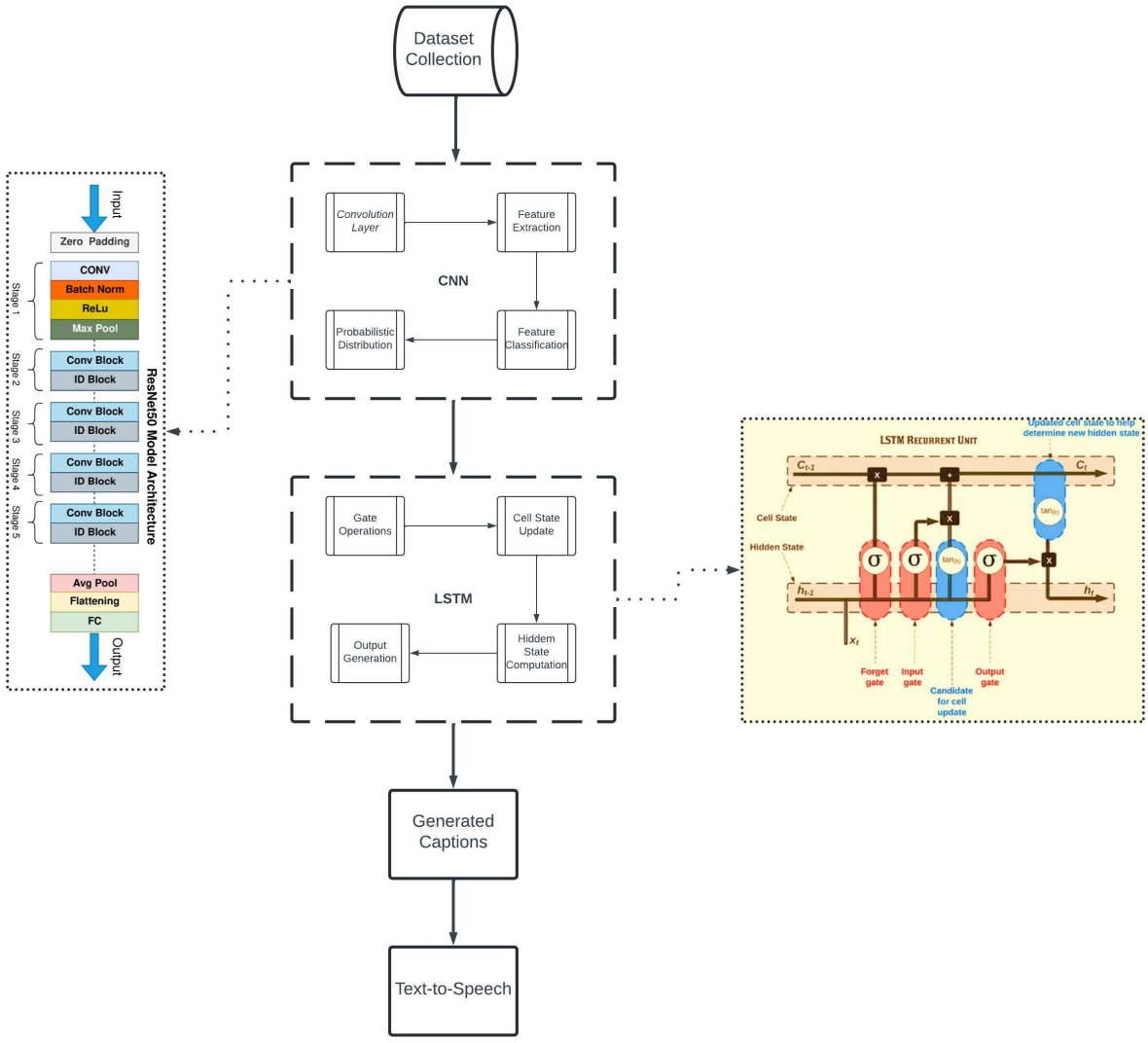


Figure 3.1: Architecture Diagram

3.2 Architectural Design

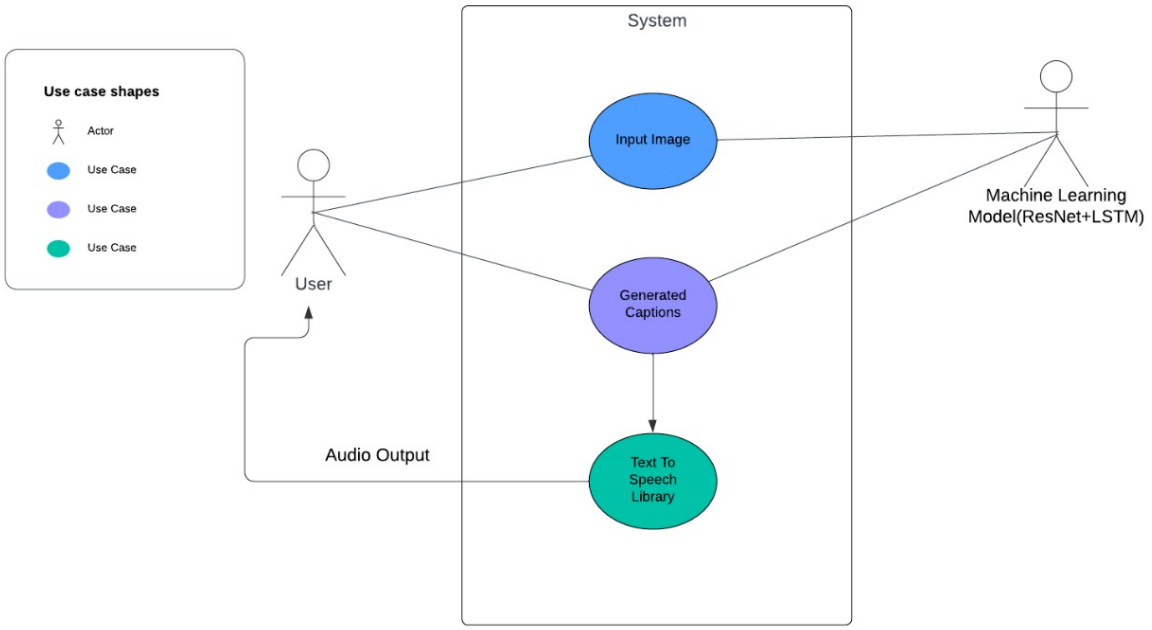


Figure 3.2: Use Case Diagram

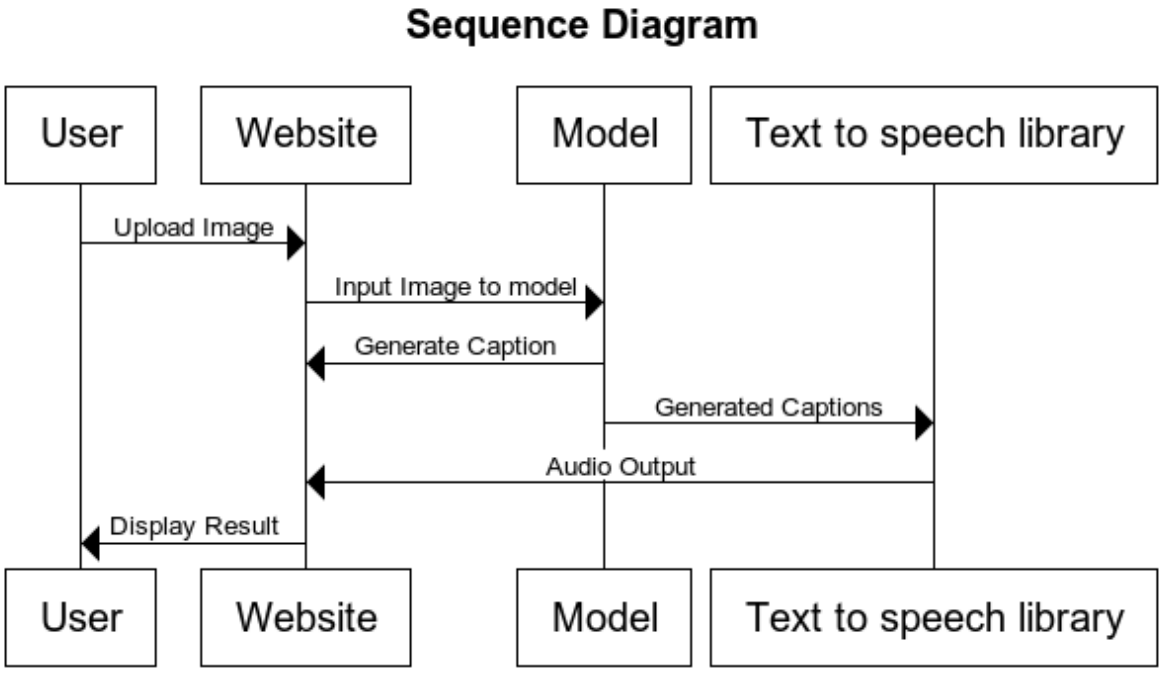


Figure 3.3: Sequence Diagram

3.3 Dataset Identified

The COCO [2] and FLICKR-8K [1] dataset consists of an extensive collection of images that depict everyday scenes with various objects in different contexts.Each image is carefully annotated with bounding box coordinates and segmentation masks, providing precise information about the location and shape of the objects present in the scene.In addition to object annotations, the COCO and FLICKR-8K dataset includes captions for a subset of the images. These captions are human-generated textual descriptions that provide an understanding of the objects and activities depicted in the images.The data in COCO is stored in a JSON file structured into sections such as info, licenses, categories, images, and annotations. Generating individual JSON files for training, testing, and validation is possible, allowing for distinct datasets for each specific purpose.The captions in the COCO and FLICKR-8K dataset provide textual descriptions that go beyond the object annotations. They offer additional information about the scene, activities, object relationships, and other contextual details. This makes the dataset valuable for training models to understand images at a semantic level and generate meaningful and accurate captions.

3.4 Proposed Methodology/Algorithms

The project is executed by using two prominent machine learning algorithms: Convolution Neural Networks (CNN) and Long Short Term Memory (LSTM) [3].

Convolution Neural Network (CNN) is a category of machine learning model, namely a type of deep learning algorithm well suited to analyse visual data. CNNs (also referred to as Convnets) use principles from linear algebra, particularly convolution operations, to extract features and identify patterns within images. Although CNNs are predominantly used to process images, they can also be adapted to work with audio and other signal data. CNNs use a series of layers, each of which detects different features of an input image. Depending on the complexity of its intended purpose, a CNN can contain dozens, hundreds or even thousands of layers, each building on the outputs of previous layers to recognize detailed patterns. There are different types of CNN models which are specialized in performing certain tasks. In this project we are making use of ResNet.ResNet (Residual Networks) is a type of CNN algorithm that is particularly well-suited for image recognition

and processing tasks.ResNets are known for their ability to train very deep networks. They are particularly used for key point detection of specific points of an object in an image.

Long Short-Term Memory (LSTM) is an improved version of recurrent neural network suitable for sequence prediction tasks and capturing long-term dependencies LSTMs are able to process and analyse sequential data, such as time series, text, and speech. They use a memory cell and set of gates to control the flow of information, allowing them to selectively retain or discard information as and when required. LSTMs are utilized for natural language processing tasks such as machine translation, language modelling, and text summarization. By understanding the relationships between words in a sentence, they can be trained to construct meaningful and grammatically correct sentences. LSTMs are also widely used for speech recognition and time series forecasting

3.5 User Interface Design

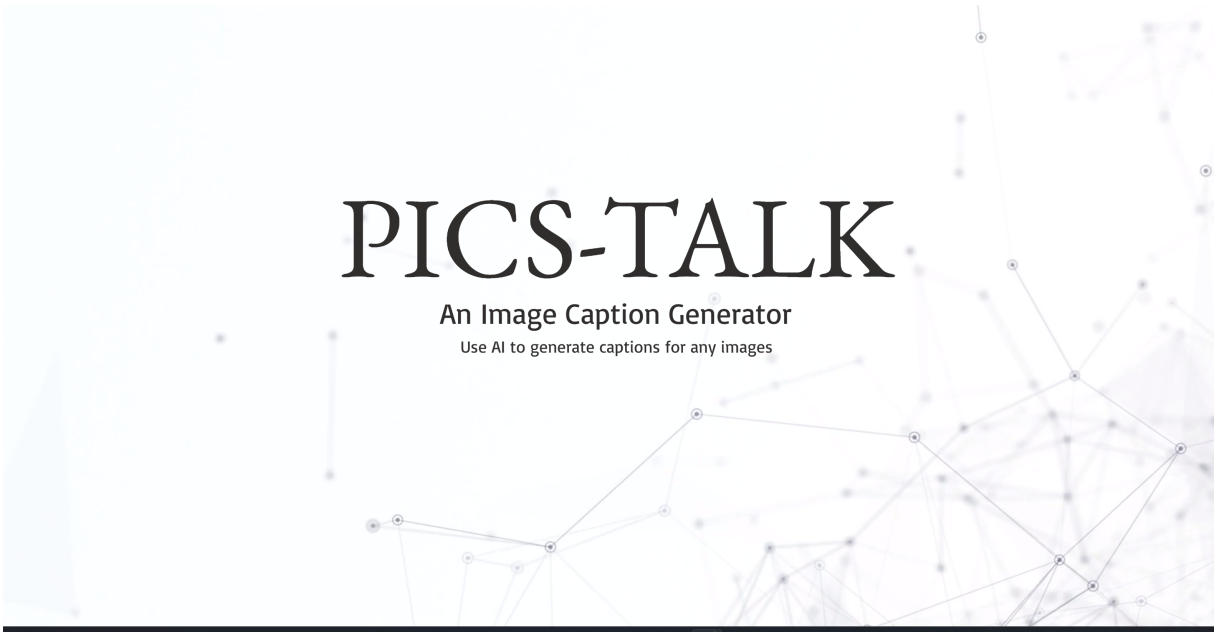


Figure 3.4: User Interface 1

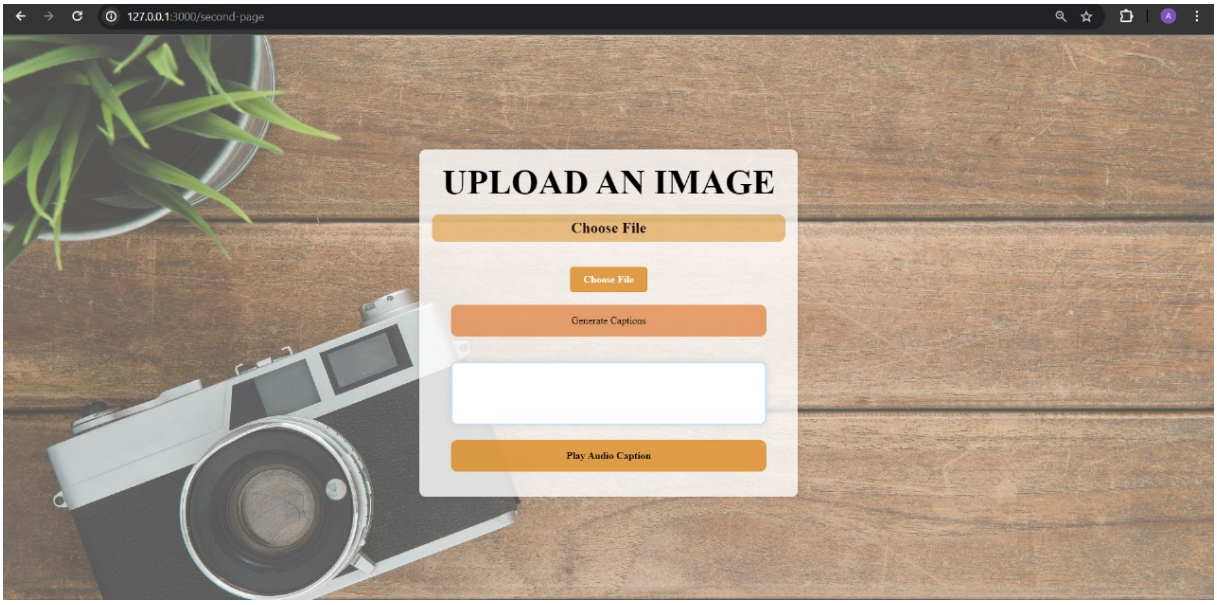


Figure 3.5: User Interface 2

3.6 Description of Implementation Strategies

Residual Networks based on Convolution Neural Networks and Long Short Term Memory are the implementation strategies used in the project. CNN is a deep learning algorithm which is used to analyse data in the form of visual representations. Depending on the complexity of its intended purpose, a CNN can contain dozens, hundreds or even thousands of layers, each building on the outputs of previous layers to recognize detailed patterns. There are different types of CNN models which are specialized in performing certain tasks. This project uses ResNet (Residual Networks), a type of CNN algorithm that is particularly well-suited for image recognition and processing tasks. CNNs contain a series of layers, each of which detects the different features of the input image. LSTMs are able to process and analyse sequential data. They use a memory cell and set of gates to control the flow of information, allowing them to selectively retain or discard information as and when required. By understanding the relationships between words in a sentence, they can be trained to construct meaningful and grammatically correct sentences. In short, Long Short Term Memory algorithm is used to analyse visual information which is in turn used to generate appropriate captions .Further a Text-to-Speech module is used to convert the generated captions into an audio.

3.7 Module Division

1. **Image Recognition using ResNet-50**

   The Image Recognition module serves as the foundation of the system. Using the ResNet-50 convolutional neural network to analyse images. ResNet-50 is a pre-trained model known for its accuracy in identifying and extracting features from images. In this module, ResNet-50 processes the input images, extracting key visual elements like shapes, textures, and objects. These features are then used as a detailed representation of the image, which becomes the input for the next stages of our system.

2. **Caption Generation using LSTM**

   Following the image feature extraction process, the Caption Generation module employs a Long Short-Term Memory (LSTM) neural network to generate descriptive captions for the input images [4]. LSTMs are a type of recurrent neural network (RNN) specifically designed to model sequential data. In this context, the LSTM network is trained on a dataset containing pairs of images and corresponding captions. During inference, the LSTM network takes the extracted image features as input and generates captions by predicting the most likely sequence of words. This process allows the system to generate coherent and contextually relevant captions that describe the content of the input images.

3. **Audio Output using pyttsx3 Library**

   Once the captions are generated by the LSTM network, they are passed to the Audio Output module for conversion into speech . This module utilizes the pyttsx3 library [5], a text-to-speech synthesis library for Python, to transform the textual captions into audio output. By leveraging pyttsx3, the system is able to provide users with audio descriptions of the images, enhancing accessibility and to perceive and understand the visual content.

4. **User Interface using Webpage**

The User Interface module serves as the interface through which users interact with the image captioning system. Implemented as a web application, the user interface provides a seamless and intuitive platform for users to upload images and receive captions and audio descriptions in real-time. Users can simply navigate to the web page, upload an image through the provided interface, and instantly receive descriptive captions along with corresponding audio output.

3.8 Work Schedule - Gantt Chart

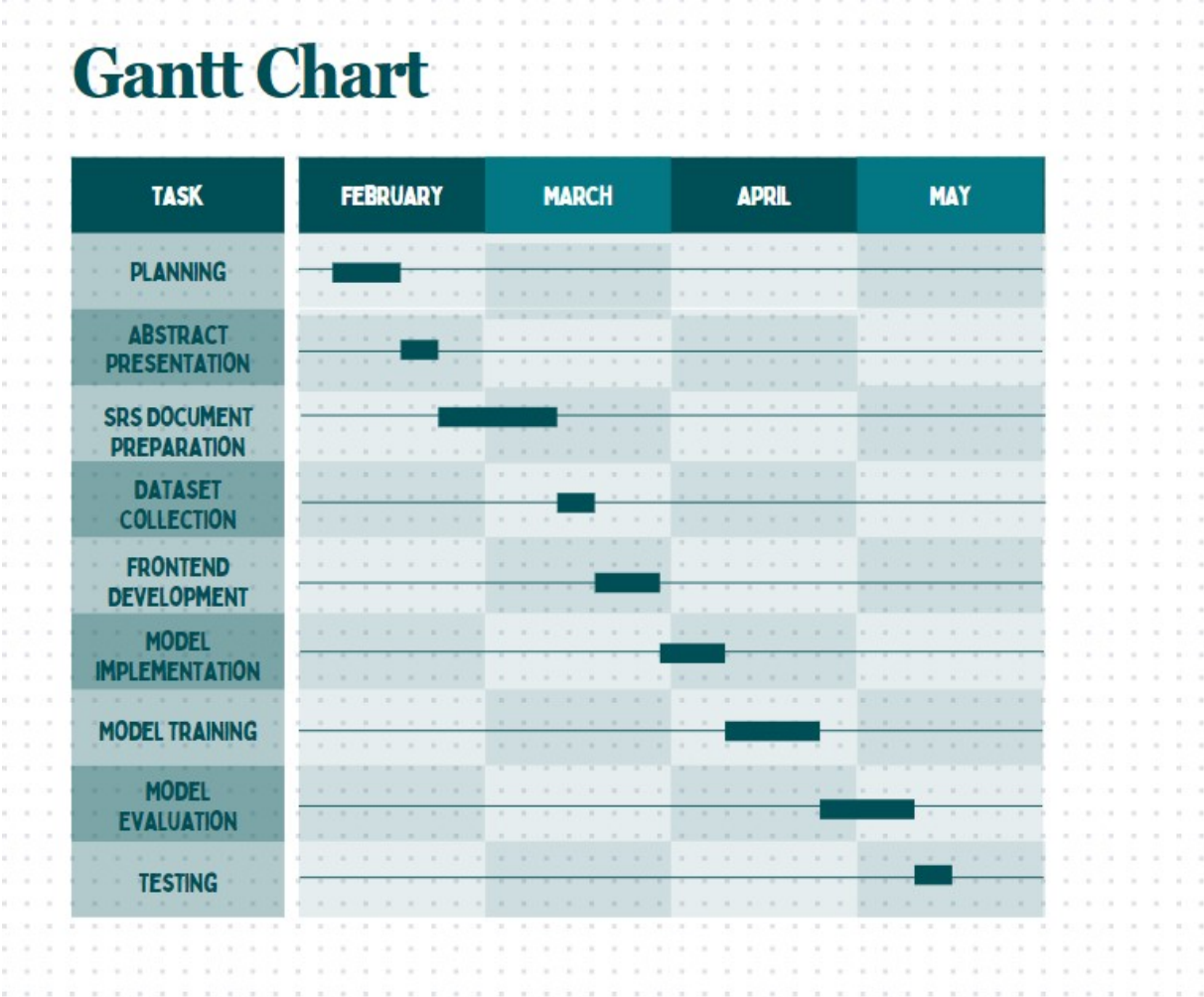


Figure 3.6: Gantt Chart

Chapter 4

Results and Discussions

4.1 Overview

This project focuses on extracting features from a given input image and generating appropriate captions from it. The front end is developed using HTML-CSS, while Python is utilized in the back end. The front and back end are integrated using Python Flask. The necessary dataset was identified, downloaded (FLICKR-8K dataset) and was used to train the model for the desired output. Although the predictions are not always cent percent accurate, the model provides an overall description of the given input image. Besides generating captions, it also converts the generated texts to audio output

Further, the Resnet50 and LSTM model was also trained using the MS COCO dataset. However, the results were much better while using the FLICKR-8K dataset.The precision and accuracy were considerably higher and loss was much lower while using the FLICKR-8K Dataset.

This project also demonstrates the integration of modern web technologies with machine learning, providing a seamless user experience. By borrowing the FLICKR-8K dataset, the model is exposed to a wide variety of images, enhancing its ability to identify and extract features. The Flask framework facilitates smooth communication between the front-end interface and the back-end processing, ensuring efficient data handling and response generation.

4.2 Testing

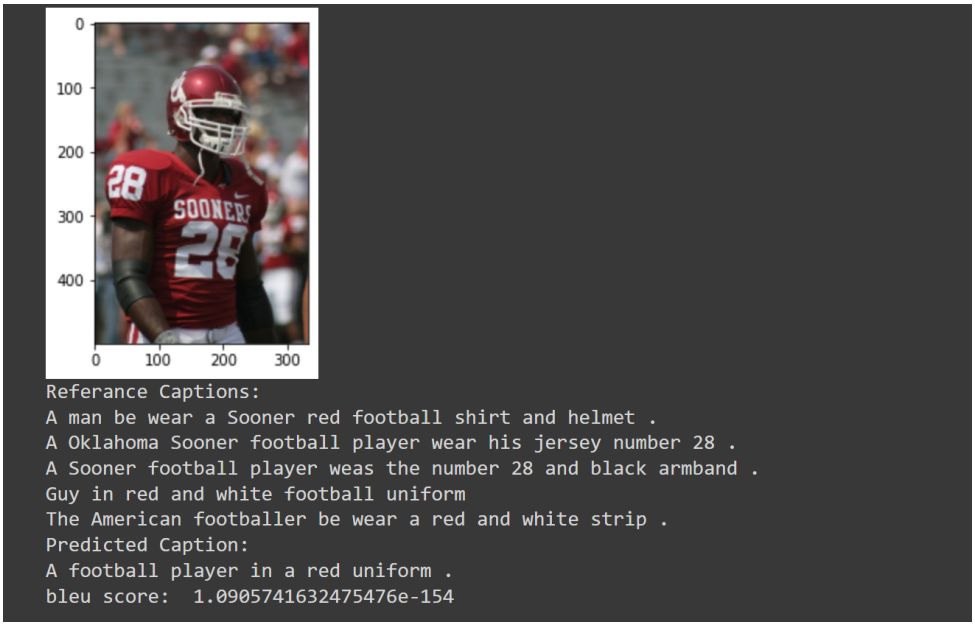


Figure 4.1: FLICKR-8K TEST 1

Figure 4.2: FLICKR 8K TEST 2

Figure 4.3: MS COCO TEST 1

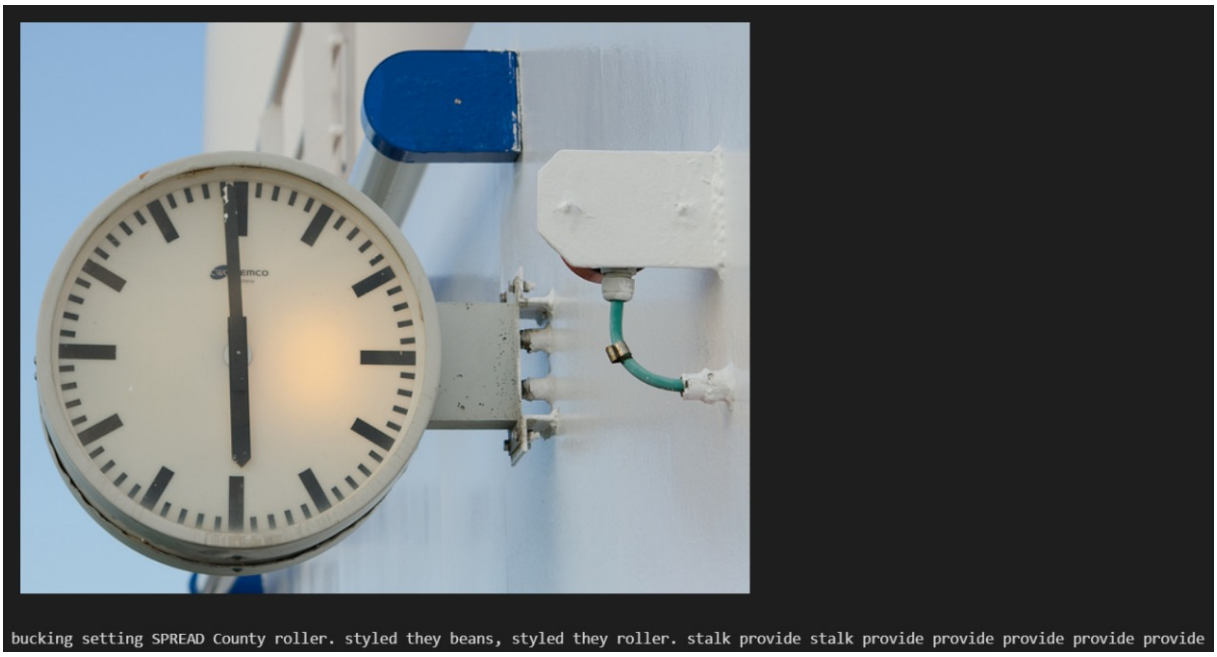


Figure 4.4: MS COCO TEST 2

4.3 Quantitative Results

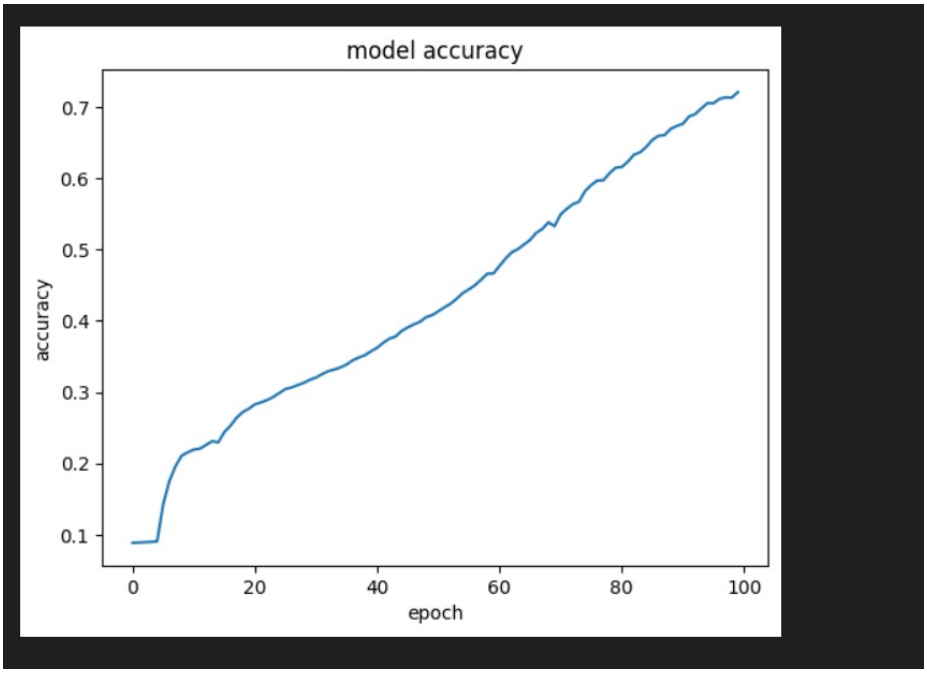


Figure 4.5: Accuracy

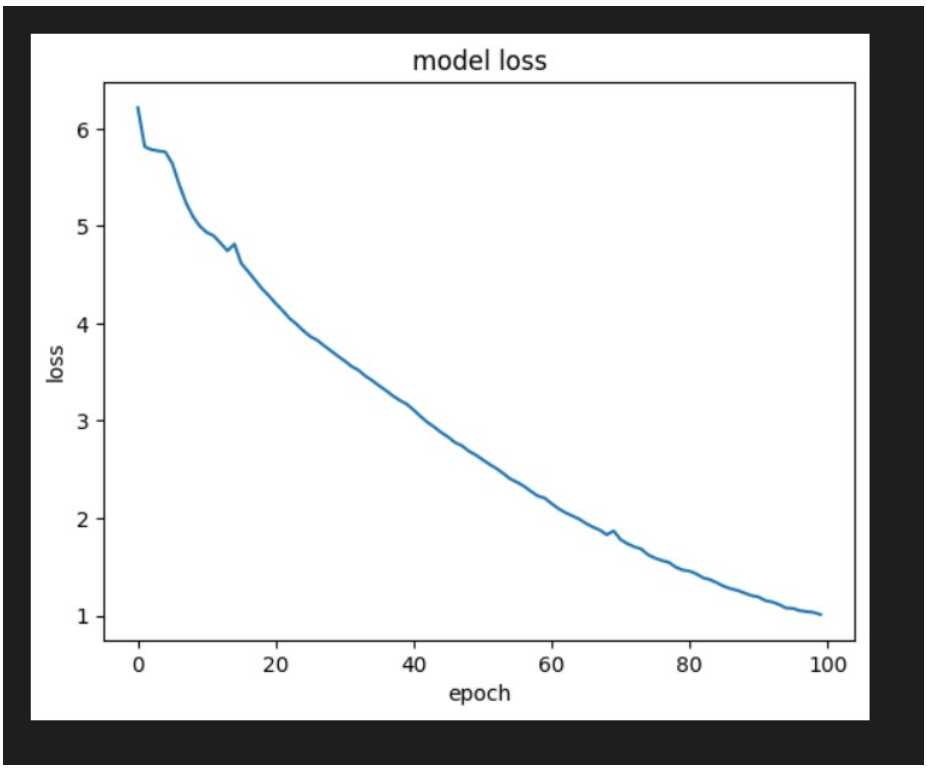


Figure 4.6: Loss

4.4 Discussion

The primary objective of this project is to detect and identify images, generating captions in both text and audio formats. This goal has been successfully realized by developing a suitable model. The desired model was created by training ResNet-50 and LSTM with an appropriate dataset .This project recognises and extracts features within an image and generates corresponding captions and audio. The accuracy and precision of the generated captions are mostly satisfactory . However, the predictions are not always cent percent accurate.

To address these challenges, we have explored various techniques to enhance model performance. This includes experimenting with different datasets and fine-tuning the model's parameters to improve accuracy. Moreover, implementing advanced pre-processing techniques has shown promise in enhancing object detection capabilities. Despite these efforts, achieving one hundred percent accuracy still remains a significant challenge due to the complexity and variability of real-world images. Nevertheless, we have strived to develop the most accurate model possible, creating two to three versions during the process with the goal of achieving the best results we could.

Chapter 5

Conclusion

5.1 Conclusion

In conclusion, this project integrates image detection, caption generation and audio output through the utilization of HTML CSS, Python, and Flask. The model is competent in generating descriptive captions for a variety of images, with some limitations in accuracy. Additionally, the model demonstrates the capability to recognize and categorize objects in images not included in the training dataset, highlighting its robustness.It has also been observed that the model gives better results while using FLICKR-8K when compared to MS COCO Dataset.

The implementation includes a user-friendly interface that allows individuals to upload images and receive immediate feedback in the form of captions and audio descriptions.Furthermore, the audio output is generated using text-to-speech technology, providing a clear and natural listening experience.By exploiting Flask, the project ensures efficient back-end processing and seamless integration with the web interface. The use of ResNet-50, known for its deep learning capabilities, ensures high-quality image recognition by object detection and LSTM guarantees generation of comprehensible and precise captions corresponding to the extracted features.

Overall, the project showcases the potential of merging modern web technologies with machine learning to create an automated image understanding tool.

5.2 Future Scope

This project is currently a website but it can be transformed into an application.With a certain level of automation, this project can also prove to be useful for visually impaired

people. Additionally, incorporating a mobile-friendly interface would make the application more accessible on various devices. Integrating real-time image processing could also enhance the responsiveness and usability of the tool. Furthermore, audio control can be used to access pictures from the gallery which can improve the ease of use. Implementing multilingual support for audio output could also broaden the user base and increase accessibility globally.

Bibliography

[1] Megha J Panicker, Vikas Upadhayay, Gunjan Sethi, and Vrinda Mathur. Image caption generator. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, 10(3):87–92, 2021.

[2] Zhou Ren, Xiaoyu Wang, Ning Zhang, Xutao Lv, and Li-Jia Li. Deep reinforcement learning-based image captioning with embedding reward. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 290–298, 2017.

[3] Satish Kumar Satti, Goluguri NV Rajareddy, Prasad Maddula, and NV Vishnumurthy Ravipati. Image caption generation using resnet-50 and lstm. In *2023 IEEE Silchar Subsection Conference (SILCON)*, pages 1–6. IEEE, 2023.

[4] Alex Sherstinsky. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Physica D: Nonlinear Phenomena*, 404:132306, 2020.

[5] Puli Venkatesh Kumar, Puli Nitish Kumar, Kanagasabapathi MG, et al. Enhancing transmission of voice in real-time applications. *International Research Journal on Advanced Science Hub*, 5(05S):271–279, 2023.

Appendix A: Presentation

PICS-TALK

Project Guide:
Ms. Seema Safar
Asst. Professor
Dept. of CSE

Group Members:
Ajith Varghese Abraham
Anna Prince
Arjun Martin

CONTENTS

INTRODUCTION

- **Images**

-Visual representations of physical world

-Conveys enormous amounts of information across the internet and social media

-Image Recognition

-Image Processing

-Image Classification

PROBLEM DEFINITION

- Develop a model that can extract features from an image and generate corresponding Captions
- Convert captions into Audio

OBJECTIVES

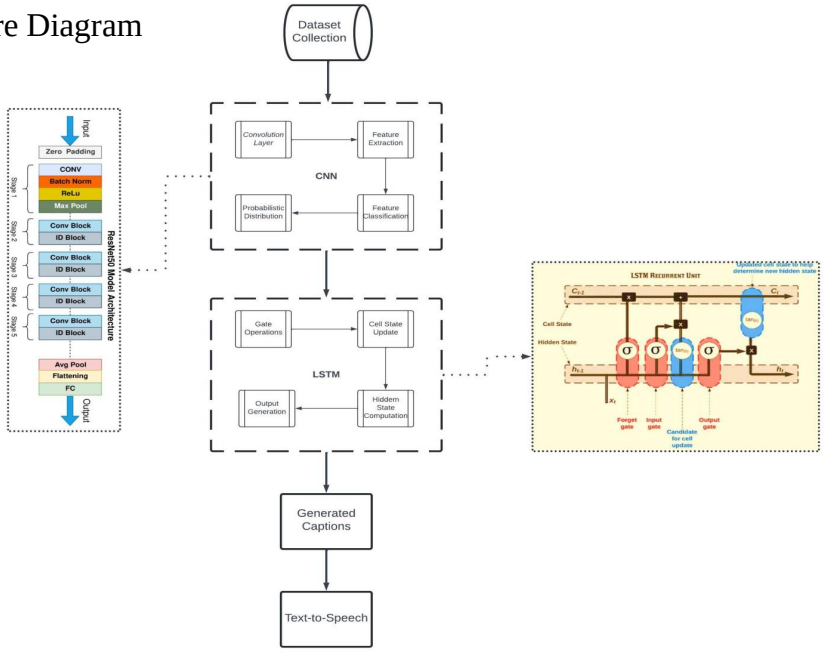
- Collection of the required dataset.
- Identification of the different components present in an image.
- Determination of the relationship between the components in the image.
- Generation of descriptions corresponding to the analysed data.
- Conversion of the generated captions into an audio file.

SCOPE AND RELEVANCE

- Improves accessibility and user experience
- Search Optimization
- Content Organization and Descriptions
- Increases cost and time efficiency-Automation

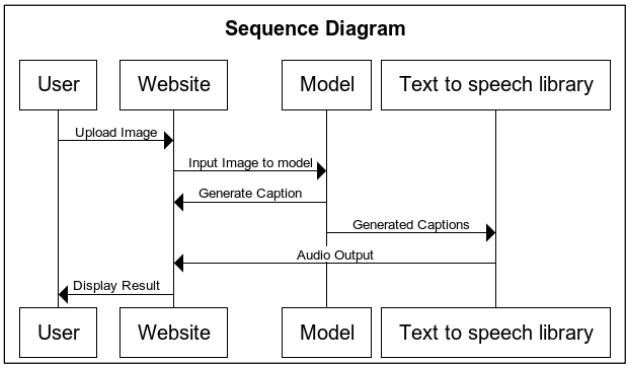
SYSTEM DESIGN

- Architecture Diagram

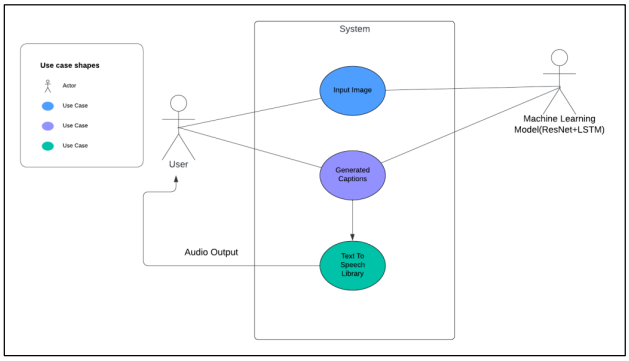


SYSTEM DESIGN

- Sequence Diagram



- Use-Case Diagram



I. IMAGE RECOGNITION USING RESNET-50

- Powerful Image Classification Model
- Type of convolution neural network (CNN) that forms networks by stacking residual blocks
- 50-layer convolutional neural network
- Divided into four main parts:

  -The Convolutional Layers : Extracts features from the input image

  -**The Identity Block** : Processing Features

  -**The Convolutional Block** : Transforming Features

  -The Fully Connected Layers : Final Feature Classification

II. CAPTION GENERATION USING LSTM

- A type of Recurrent Neural Network (RNN)
- Specifically designed to model sequential data
- Uses the extracted image features as input and generates captions by predicting the most likely sequence of words
- Divided into four parts:

  -Input Gate : Controls the flow of information into the memory cell

  -Forget Gate : Selectively forget or remember information from previous memory cells

  -Memory Cell : Stores information that can be selectively modified by the input and forget gates.

  -Output Gate : Controls the flow of information from the memory cell to the current hidden state and output

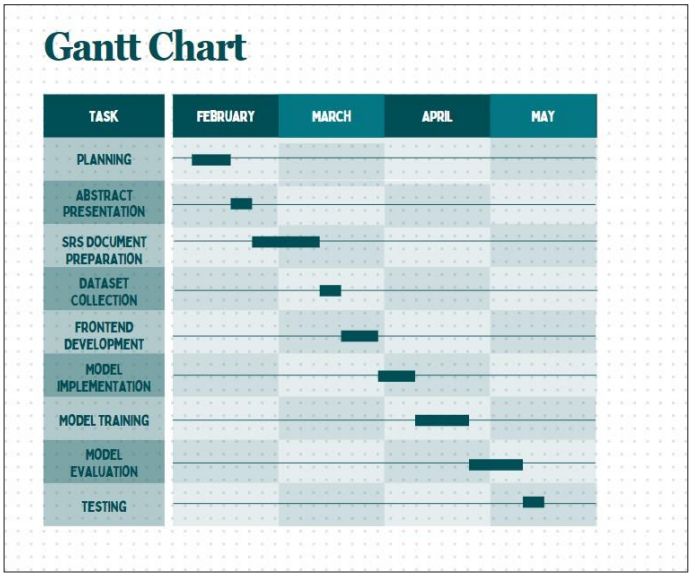
III. AUDIO OUTPUT USING PYTTSX3 LIBRARY

- Utilizes the pyttsx3 library
- Transform the textual captions into audio output
- Enhances accessibility and understanding of visual content

DATA-SET

- Uses **FLICKR 8K Dataset**
- Extensive collection of 8091 images that depict everyday scenes
- Includes five Human-generated textual descriptions for each image
- COCO (Common Objects in Context) Data-set

WORK DIVISION



SOFTWARE-HARDWARE REQUIREMENTS

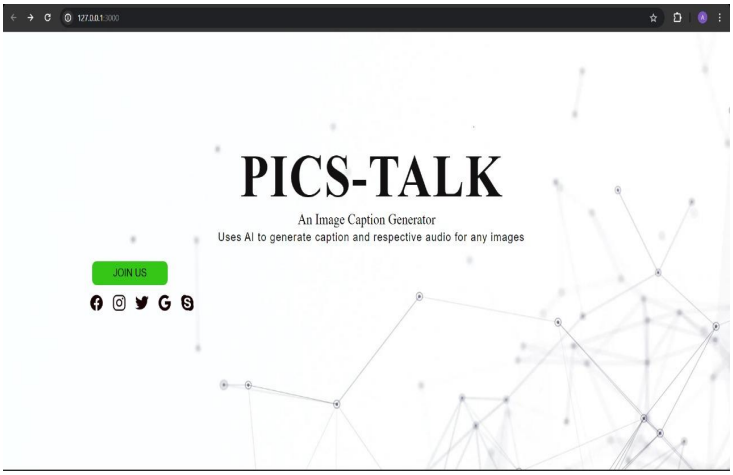
Software

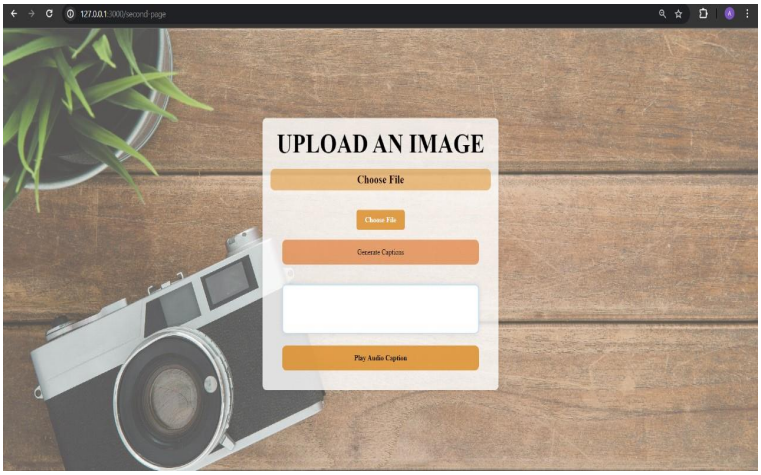
- Deep Learning Frameworks –Tensorflow, Keras
- Python
- Text to Speech- pyttsx3
- Image Processing Libraries -OpenCV

Hardware

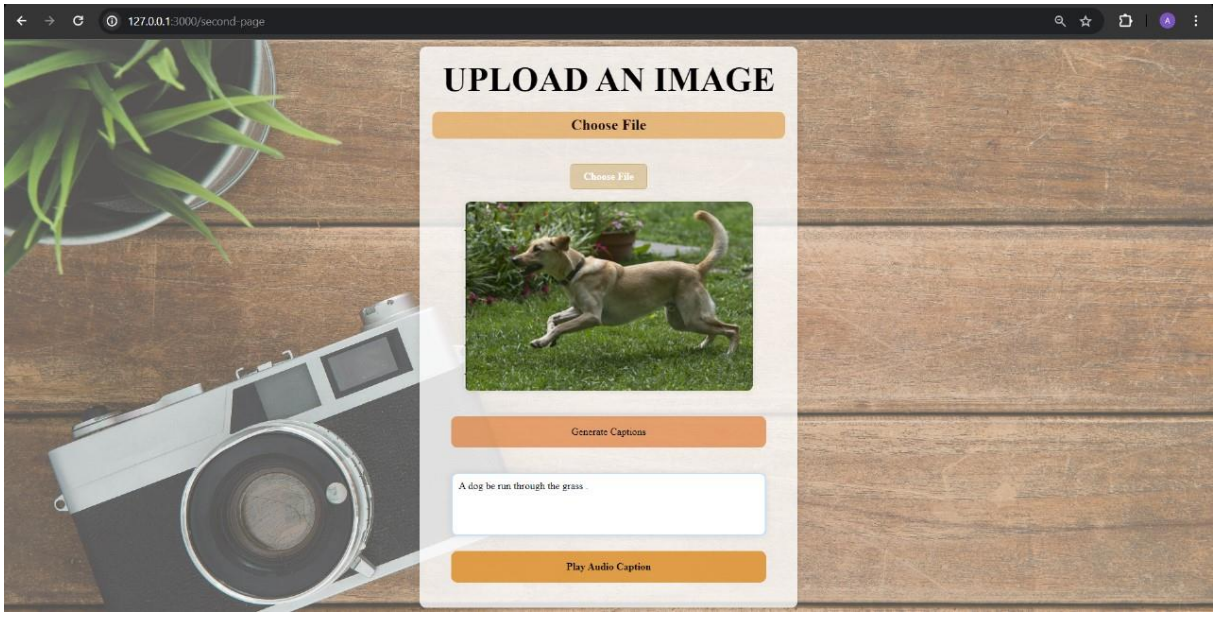
- **CPU-**Intel Core i5 , AMD Ryzen 5 or higher
- **GPU-**NVIDIA GTX 1060 , AMD Radeon RX 6800 XT or higher
- RAM-16 GB or higher
- SSD-Depending on size of dataset

RESULTS



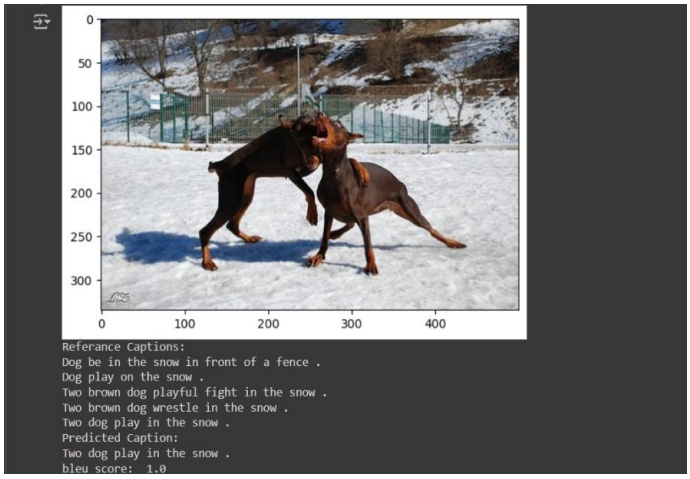


RESULTS



RESULTS

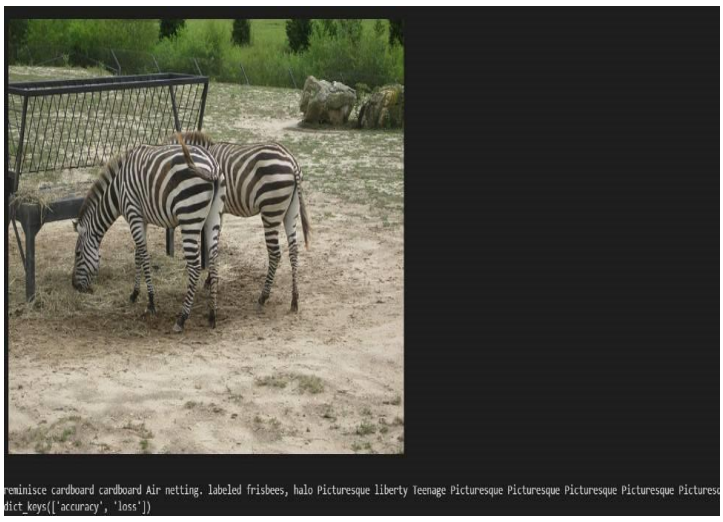
- FLICKR 8K Dataset



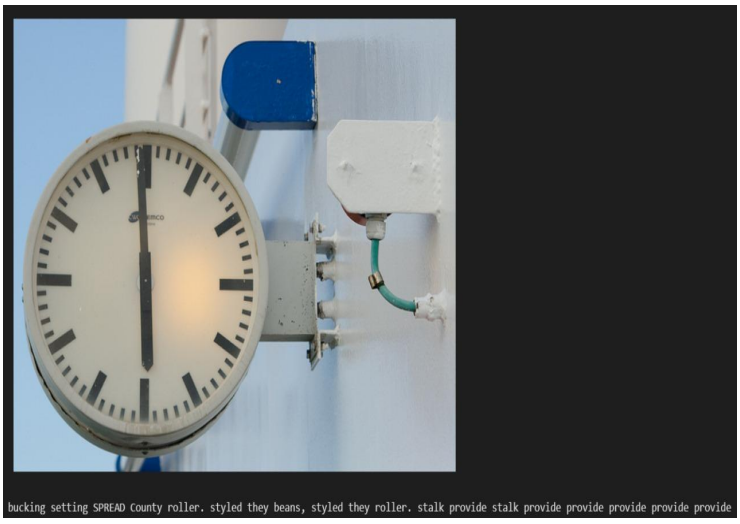
'loss': [0.8245726227760315], **'accuracy':** [0.8121230602264404], '**precision':** 0.9584 , **recall:** 0.7623100

RESULTS

- MS COCO Dataset



'Accuracy': 0.9170 , **'loss':** 0.2040



'Accuracy': 0.7282 , **' loss':** 0.9859

CONCLUSION

- Pics-Talk aims to :

  -Analyse Images

  -Generate Captions

  -Convert Captions into audio

FUTURE ENHANCEMENTS

- Transformed into an application for enhancing accessibility
- Can be automated- helps visually impaired
- Multi-lingual Support
- Future work could focus on improving accuracy with larger datasets, fine-tuning specific image domains and exploring other advanced features

REFERENCES

- Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2015. Going deeper with convolutions.
- Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhutdinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention.
- Murk Chohan, Adil Khan, Muhammad Saleem Mahar Saif Hassan, Abdul Ghafoor, Mehmood Khan Department of Computer Science Sukkur IBA University Pakistan,Y.2020. Image Captioning using Deep Learning: A Systematic Literature Review.
- Zhou Ren, Xiaoyu Wang, Ning Zhang, Xutao Lv, Li-Jia Li; Deep Reinforcement Learning-Based Image Captioning With Embedding Reward by Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 290-298.
- Mao, J.; Wei, X.; Yang, Y.; Wang, J.; Huang, Z.; and Yuille, A. L. 2015a. Learning like a child: Fast novel visual concept learning from sentence descriptions of images.

Appendix B: Vision, Mission, Programme Outcomes and Course Outcomes

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
RAJAGIRI SCHOOL OF ENGINEERING & TECHNOLOGY (AUTONOMOUS)
RAJAGIRI VALLEY, KAKKANAD, KOCHI, 682039
(Affiliated to APJ Abdul Kalam Technological University)

Vision, Mission, Programme Outcomes and Course Outcomes

Institute Vision

To evolve into a premier technological institution, moulding eminent professionals with creative minds, innovative ideas and sound practical skill, and to shape a future where technology works for the enrichment of mankind.

Institute Mission

To impart state-of-the-art knowledge to individuals in various technological disciplines and to inculcate in them a high degree of social consciousness and human values, thereby enabling them to face the challenges of life with courage and conviction.

Department Vision

To become a centre of excellence in Computer Science and Engineering, moulding professionals catering to the research and professional needs of national and international organizations.

Department Mission

To inspire and nurture students, with up-to-date knowledge in Computer Science and Engineering, ethics, team spirit, leadership abilities, innovation and creativity to come out with solutions meeting societal needs.

Programme Outcomes (PO)

Engineering Graduates will be able to:

1. Engineering Knowledge: Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.

2. Problem analysis: Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.

3. Design/development of solutions: Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.

4. Conduct investigations of complex problems: Use research-based knowledge including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.

5. Modern Tool Usage: Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations.

6. The engineer and society: Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal, and cultural issues and the consequent responsibilities relevant to the professional engineering practice.

7. Environment and sustainability: Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.

8. Ethics: Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.

9. Individual and Team work: Function effectively as an individual, and as a member or leader in teams, and in multidisciplinary settings.

10. Communication: Communicate effectively with the engineering community and with society at large. Be able to comprehend and write effective reports documentation. Make effective presentations, and give and receive clear instructions.

11. Project management and finance: Demonstrate knowledge and understanding of engineering and management principles and apply these to one's own work, as a member and leader in a team. Manage projects in multidisciplinary environments.

12. Life-long learning: Recognize the need for, and have the preparation and ability to engage in independent and lifelong learning in the broadest context of technological change.

Programme Specific Outcomes (PSO)

A graduate of the Computer Science and Engineering Program will demonstrate:

PSO1: Computer Science Specific Skills

The ability to identify, analyze and design solutions for complex engineering problems in multidisciplinary areas by understanding the core principles and concepts of computer science and thereby engage in national grand challenges.

PSO2: Programming and Software Development Skills

The ability to acquire programming efficiency by designing algorithms and applying standard practices in software project development to deliver quality software products meeting the demands of the industry.

PSO3: Professional Skills

The ability to apply the fundamentals of computer science in competitive research and to develop innovative products to meet the societal needs thereby evolving as an eminent researcher and entrepreneur.

Course Outcomes

After the completion of the course the student will be able to:

CO1:

Identify technically and economically feasible problems (Cognitive Knowledge Level: Apply)

CO2:
Identify and survey the relevant literature for getting exposed to related solutions and get familiarized with software development processes (Cognitive Knowledge Level: Apply)

CO3:
Perform requirement analysis, identify design methodologies and develop adaptable & reusable solutions of minimal complexity by using modern tools & advanced programming techniques (Cognitive Knowledge Level: Apply)

CO4:
Prepare technical report and deliver presentation (Cognitive Knowledge Level: Apply)

CO5:
Apply engineering and management principles to achieve the goal of the project (Cognitive Knowledge Level: Apply)

Appendix C: CO-PO-PSO Mapping

COURSE OUTCOMES:

After completion of the course the student will be able to

SL. NO	DESCRIPTION	Blooms' Taxonomy Level
CO1	Identify technically and economically feasible problems (Cognitive Knowledge Level: Apply)	Level 3: Apply
CO2	Identify and survey the relevant literature for getting exposed to related solutions and get familiarized with software development processes (Cognitive Knowledge Level: Apply)	Level 3: Apply
CO3	Perform requirement analysis, identify design methodologies and develop adaptable & reusable solutions of minimal complexity by using modern tools & advanced programming techniques (Cognitive Knowledge Level: Apply)	Level 3: Apply
CO4	Prepare technical report and deliver presentation (Cognitive Knowledge Level: Apply)	Level 3: Apply
CO5	Apply engineering and management principles to achieve the goal of the project (Cognitive Knowledge Level: Apply)	Level 3: Apply

CO-PO AND CO-PSO MAPPING

	PO 1	PO 2	PO 3	PO 4	PO 5	PO 6	PO 7	PO 8	PO 9	PO 10	PO 11	PO 12	PSO 1	PSO 2	PSO3
CO1	3	3	3	3		2	2	3	2	2	2	3	2	2	2
CO2	3	3	3	3	3	2		3	2	3	2	3	2	2	2
CO3	3	3	3	3	3	2	2	3	2	2	2	3			2
CO4	2	3	2	2	2			3	3	3	2	3	2	2	2
CO5	3	3	3	2	2	2	2	3	2		2	3	2	2	2

3/2/1: high/medium/low

JUSTIFICATIONS FOR CO-PO MAPPING

MAPPING	LOW/ MEDIUM/ HIGH	JUSTIFICATION
101003/CS622T.1-PO1	**HIGH**	Identify technically and economically feasible problems by applying the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.
101003/CS622T.1-PO2	**HIGH**	Identify technically and economically feasible problems by analysing complex engineering problems reaching substantiated conclusions using first principles of mathematics.
101003/CS622T.1-PO3	**HIGH**	Design solutions for complex engineering problems by identifying technically and economically feasible problems.
101003/CS622T.1-PO4	**HIGH**	Identify technically and economically feasible problems by analysis and interpretation of data.
101003/CS622T.1-PO6	**MEDIUM**	Responsibilities relevant to the professional engineering practice by identifying the problem.
101003/CS622T.1-PO7	**MEDIUM**	Identify technically and economically feasible problems by understanding the impact of the professional engineering solutions.
101003/CS622T.1-PO8	**HIGH**	Apply ethical principles and commit to professional ethics to identify technically and economically feasible problems.
101003/CS622T.1-PO9	**MEDIUM**	Identify technically and economically feasible problems by working as a team.
101003/CS622T.1-PO10	**MEDIUM**	Communicate effectively with the engineering community by identifying technically and economically feasible problems.
101003/CS622T.1-P011	**MEDIUM**	Demonstrate knowledge and understanding of engineering and management principles by selecting the technically and economically feasible problems.
101003/CS622T.1-PO12	**HIGH**	Identify technically and economically feasible problems for long term learning.
101003/CS622T.1-PSO1	**MEDIUM**	Ability to identify, analyze and design solutions to identify technically and economically feasible problems.
101003/CS622T.1-PSO2	**MEDIUM**	By designing algorithms and applying standard practices in software project development and Identifying technically and economically feasible problems.
101003/CS622T.1-PSO3	**MEDIUM**	Fundamentals of computer science in competitive research can be applied to Identify technically and economically feasible problems.
101003/CS622T.2-PO1	**HIGH**	Identify and survey the relevant by applying the knowledge of mathematics, science, engineering fundamentals.

101003/CS622T.2-PO2	**HIGH**	Identify, formulate, review research literature, and analyze complex engineering problems get familiarized with software development processes.
101003/CS622T.2-PO3	**HIGH**	Design solutions for complex engineering problems and design based on the relevant literature.
101003/CS622T.2-PO4	**HIGH**	Use research-based knowledge including design of experiments based on relevant literature.
101003/CS622T.2-PO5	**HIGH**	Identify and survey the relevant literature for getting exposed to related solutions and get familiarized with software development processes by using modern tools.
101003/CS622T.2-PO6	**MEDIUM**	Create, select, and apply appropriate techniques, resources, by identifying and surveying the relevant literature.
101003/CS622T.2-PO8	**HIGH**	Apply ethical principles and commit to professional ethics based on the relevant literature.
101003/CS622T.2-PO9	**MEDIUM**	Identify and survey the relevant literature as a team.
101003/CS622T.2-PO10	**HIGH**	Identify and survey the relevant literature for a good communication to the engineering fraternity.
101003/CS622T.2-PO11	**MEDIUM**	Identify and survey the relevant literature to demonstrate knowledge and understanding of engineering and management principles.
101003/CS622T.2-PO12	**HIGH**	Identify and survey the relevant literature for independent and lifelong learning.
101003/CS622T.2-PSO1	**MEDIUM**	Design solutions for complex engineering problems by Identifying and survey the relevant literature.
101003/CS622T.2-PSO2	**MEDIUM**	Identify and survey the relevant literature for acquiring programming efficiency by designing algorithms and applying standard practices.
101003/CS622T.2-PSO3	**MEDIUM**	Identify and survey the relevant literature to apply the fundamentals of computer science in competitive research.
101003/CS622T.3-PO1	**HIGH**	Perform requirement analysis, identify design methodologies by using modern tools & advanced programming techniques and by applying the knowledge of mathematics, science, engineering fundamentals.
101003/CS622T.3-PO2	**HIGH**	Identify, formulate, review research literature for requirement analysis, identify design methodologies and develop adaptable & reusable solutions.

101003/CS622T.3-PO3	**HIGH**	Design solutions for complex engineering problems and perform requirement analysis, identify design methodologies.
101003/CS622T.3-PO4	**HIGH**	Use research-based knowledge including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.
101003/CS622T.3-PO5	**HIGH**	Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools.
101003/CS622T.3-PO6	**MEDIUM**	Perform requirement analysis, identify design methodologies and assess societal, health, safety, legal, and cultural issues.
101003/CS622T.3-PO7	**MEDIUM**	Understand the impact of the professional engineering solutions in societal and environmental contexts and Perform requirement analysis, identify design methodologies and develop adaptable & reusable solutions.
101003/CS622T.3-PO8	**HIGH**	Perform requirement analysis, identify design methodologies and develop adaptable & reusable solutions by applying ethical principles and commit to professional ethics.
101003/CS622T.3-PO9	**MEDIUM**	Function effectively as an individual, and as a member or leader in teams, and in multidisciplinary settings.
101003/CS622T.3-PO10	**MEDIUM**	Communicate effectively with the engineering community and with society at large to perform requirement analysis, identify design methodologies.
101003/CS622T.3-PO11	**MEDIUM**	Demonstrate knowledge and understanding of engineering requirement analysis by identifying design methodologies.
101003/CS622T.3-PO12	**HIGH**	Recognize the need for, and have the preparation and ability to engage in independent and lifelong learning in the broadest context of technological change by analysis, identify design methodologies and develop adaptable & reusable solutions.
101003/CS622T.3-PSO3	**MEDIUM**	The ability to apply the fundamentals of computer science in competitive research and prior to that perform requirement analysis, identify design methodologies.
101003/CS622T.4-PO1	**MEDIUM**	Prepare technical report and deliver presentation by applying the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.
101003/CS622T.4-PO2	**HIGH**	Identify, formulate, review research literature, and analyze complex engineering problems by preparing technical report and deliver presentation.

101003/CS622T.4-PO3	**MEDIUM**	Prepare Design solutions for complex engineering problems and create technical report and deliver presentation.
101003/CS622T.4-PO4	**MEDIUM**	Use research-based knowledge including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions and prepare technical report and deliver presentation.
101003/CS622T.4-PO5	**MEDIUM**	Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools and Prepare technical report and deliver presentation.
101003/CS622T.4-PO8	**HIGH**	Prepare technical report and deliver presentation by applying ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.
101003/CS622T.4-PO9	**HIGH**	Prepare technical report and deliver presentation effectively as an individual, and as a member or leader in teams, and in multidisciplinary settings.
101003/CS622T.4-PO10	**HIGH**	Communicate effectively with the engineering community and with society at large by prepare technical report and deliver presentation.
101003/CS622T.4-PO11	**MEDIUM**	Demonstrate knowledge and understanding of engineering and management principles and apply these to one's own work by prepare technical report and deliver presentation.
101003/CS622T.4-PO12	**HIGH**	Recognize the need for, and have the preparation and ability to engage in independent and lifelong learning in the broadest context of technological change by prepare technical report and deliver presentation.
101003/CS622T.4-PSO1	**MEDIUM**	Prepare a technical report and deliver presentation to identify, analyze and design solutions for complex engineering problems in multidisciplinary areas.
101003/CS622T.4-PSO2	**MEDIUM**	To acquire programming efficiency by designing algorithms and applying standard practices in software project development and to prepare technical report and deliver presentation.
101003/CS622T.4-PSO3	**MEDIUM**	To apply the fundamentals of computer science in competitive research and to develop innovative products to meet the societal needs by preparing technical report and deliver presentation.
101003/CS622T.5-PO1	**HIGH**	Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.
101003/CS622T.5-PO2	**HIGH**	Identify, formulate, review research literature, and analyze complex engineering problems by applying engineering and management principles to achieve the goal of the project.

101003/CS622T.5-PO3	**HIGH**	Apply engineering and management principles to achieve the goal of the project and to design solutions for complex engineering problems and design system components or processes that meet the specified needs.
101003/CS622T.5-PO4	**MEDIUM**	Apply engineering and management principles to achieve the goal of the project and use research-based knowledge including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.
101003/CS622T.5-PO5	**MEDIUM**	Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools and to apply engineering and management principles to achieve the goal of the project.
101003/CS622T.5-PO6	**MEDIUM**	Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal, and cultural issues and the consequent responsibilities by applying engineering and management principles to achieve the goal of the project.
101003/CS622T.5-PO7	**MEDIUM**	Understand the impact of the professional engineering solutions in societal and environmental contexts, and apply engineering and management principles to achieve the goal of the project.
101003/CS622T.5-PO8	**HIGH**	Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice and to use the engineering and management principles to achieve the goal of the project.
101003/CS622T.5-PO9	**MEDIUM**	Function effectively as an individual, and as a member or leader in teams, and in multidisciplinary settings and to apply engineering and management principles to achieve the goal of the project.
101003/CS622T.5-PO11	**MEDIUM**	Demonstrate knowledge and understanding of engineering and management principles and apply these to one's own work, as a member and leader in a team. Manage projects in multidisciplinary environments and to apply engineering and management principles to achieve the goal of the project.
101003/CS622T.5-PO12	**HIGH**	Recognize the need for, and have the preparation and ability to engage in independent and lifelong learning in the broadest context of technological change and to apply engineering and management principles to achieve the goal of the project.
101003/CS622T.5-PSO1	**MEDIUM**	The ability to identify, analyze and design solutions for complex engineering problems in multidisciplinary areas. Apply engineering and management principles to achieve the goal of the project.

101003/CS622T.5-PSO2	**MEDIUM**	The ability to acquire programming efficiency by designing algorithms and applying standard practices in software project development to deliver quality software products meeting the demands of the industry and to apply engineering and management principles to achieve the goal of the project.
101003/CS622T.5-PSO3	**MEDIUM**	The ability to apply the fundamentals of computer science in competitive research and to develop innovative products to meet the societal needs thereby evolving as an eminent researcher and entrepreneur and apply engineering and management principles to achieve the goal of the project.