Kantipur Engineering College

(Affiliated to Tribhuvan University)

Dhapakhel,Lalitpur

A PROPOSAL ON

IMAGE-CAPTION GENERATION SYSTEM

Submitted by:

Sabin Khadka [KAN077BCT067]

Sushil Kandel [KAN077BCT092]

Srijan Maharjan [KAN077BCT088]

Sujar Khanal [KAN077BCT090]

A PROPOSAL SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF BACHELOR IN COMPUTER ENGINEERING

Submitted to:

Department of Computer and Electronics Engineering

June, 2024

IMAGE-CAPTION GENERATION SYSTEM

Submitted by:

Sabin Khadka [KAN077BCT067]

Sushil Kandel [KAN077BCT092]

Srijan Maharjan [KAN077BCT088]

Sujar Khanal [KAN077BCT090]

A PROPOSAL SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF BACHELOR IN COMPUTER ENGINEERING

Submitted to:

Department of Computer and Electronics Engineering

Kantipur Engineering College

Dhapakhel,Lalitpur

June, 2024

KANTIPUR ENGINEERING COLLEGE
DEPARTMENT OF COMPUTER AND ELECTRONICS ENGINEERING

APPROVAL LETTER

The undersigned certify that they have read and recommended to the Institute of Engineering for acceptance, a project report entitled ”Image-Caption Generation System” submitted by

Sabin Khadka [KAN077BCT067]

Sushil Kandel [KAN077BCT092]

Srijan Maharjan [KAN077BCT088]

Sujar Khanal [KAN077BCT090]

in partial fulfillment for the degree of Bachelor in Computer Engineering.

.....
External Examiner
External’s Name
External’s Designation
Second Line of Designation (if required)

.....
Er. Rabindra Khati
Head of Department
Department of Computer and Electronics Engineering

Date: June 06, 2024

ABSTRACT

An image caption generation system is an AI-based solution that combines computer vision and natural language processing to automatically create descriptive text for images. It typically involves using a Convolutional Neural Network (CNN) to extract visual features from an image and a Long Short-Term Memory (LSTM) network to generate coherent and contextually relevant captions based on those features. This system can assist visually impaired individuals by converting visual content into accessible textual descriptions.

Whether accessed through mobile applications, web interfaces, or integrated with smart home devices, image caption generation system ensures seamless accessibility, enabling users to interact with the system and to know what is happening to one's sorrunding by generating the caption.

The ability to analyze the state, properties, and relationship between these objects is required for the meaningful description generation process of high-level picture semantics. Using CNN -LSTM architectural models on the captioning of a graphical image, we hope to detect things and inform people via text messages in this project In this project, we follow a variety of important concepts of image captioning and its standard processes, as this work develops a generative CNN-LSTM model that outperforms human baselines.

Our system can also be applied in various fields like robotic vision, business analytics, and medical diagnostics. This project not only demonstrates the practical application of advanced AI techniques but also aims to make a positive impact on society by enhancing accessibility for the visually impaired.

Keywords *– CNN,LSTM,RNN,AI*

ACKNOWLEDGMENT

We express our sincere gratitude to Er. Bishal Thapa, Project Head of the Computer Department at Kantipur Engineering College, and Er. Pralhad Chapagain, Senior Lecturer, whose invaluable guidance and unwavering support made this thesis possible. Their expertise and encouragement propelled us through every phase of our project, from inception to completion.

Special appreciation is extended to our esteemed lecturers, Er. Nishan Khanal, Er. Pawan Acharya, and Er. Kabin Devkota, whose unwavering support and understanding played a crucial role in the development of our research and the crafting of this thesis. Their dedication to our academic pursuits has been instrumental in shaping our scholarly endeavors.

We would also like to express our gratitude to the Head of the Department of Computer and Electronics Engineering, Er. Rabindra Khati, for his continuous support and encouragement throughout this journey. His guidance and leadership have been invaluable in facilitating our academic and professional growth.

We are indebted to these individuals for their mentorship, wisdom, and encouragement throughout this journey. Their contributions have been invaluable, and we are deeply grateful for their commitment to our academic and professional growth.

Thank you.

Sabin Khadka [KAN077BCT067]
Sushil Kandel [KAN077BCT092]
Srijan Maharjan [KAN077BCT088]
Sujar Khanal [KAN077BCT090]

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AI Artificial Intelligence

CNN Convolutional Neural Network

LSTM Long Short-Term Memory

NLP Natural Language Processing

RNN Recurrent Neural Network

AI,AI AIfeffef

NLP,NLP LSTM,LSTM CNN,CNN RNN,RNN

CHAPTER 1
INTRODUCTION

1.1 Background

Sure, here's the corrected version of your text:

As engineering students, our goal is to constantly develop projects or apply theories to real-world situations. Finding fresh, innovative ideas to launch a new project might be difficult. We hope to develop a program that will make it easier to generate captions from images. The goal is to develop a program that can automatically generate descriptive captions for images, making visual content more accessible and useful. This is particularly beneficial for visually impaired people as it provides them with a textual understanding of visual scenes.Content creators can also benefit from it, as it saves content creators a significant amount of time and effort by automatically generating captions for their images. This can be especially helpful for social media posts, where captions are essential for engagement. Businesses can use it to improve the accessibility of websites and marketing materials. They can also use it to generate captions for product images on e-commerce websites. People with learning disabilities may find it difficult to understand complex visual information. Image caption generators can provide a clear and concise explanation of what's in an image, making it easier for them to understand. Multilingual audiences can further use it to translate captions into multiple languages, making content more accessible to a global audience.

Image caption generators significantly enhance user experiences with images by improving accessibility, engagement, searchability, language learning, and social media interactions. These technologies provide valuable support to visually impaired individuals, add meaningful context to images, aid in accurate image searches, offer educational benefits through bilingual captions, and increase social media engagement. As a result, the integration of image caption generators into various applications demonstrates the profound impact of deep learning on making visual content more inclusive, informative, and enjoyable for a diverse audience.

1.2 Problem Statement

Automatically describing the content of images is a fundamental yet challenging task that has gained significant attention with the advent of powerful tools and extensive datasets. Humans can effortlessly describe their surroundings and provide detailed accounts of images at a glance, a capability that remains complex for machines. Despite significant advancements in computer vision, such as object recognition, action classification, image classification, attribute classification, and scene recognition, generating human-like descriptions of images in natural language is a relatively new and complex task. The goal of image captioning is to capture the semantics of images and express them in a coherent and contextually accurate manner. This technology can significantly benefit visually impaired individuals by providing descriptive narratives of images available on the internet. To achieve this, the image caption generator model combines Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). CNNs, particularly pre-trained models like Xception, are utilized for feature extraction from images, and the extracted features are then used by Long Short-Term Memory (LSTM) networks to generate descriptive sentences.

1.3 Objective

- To develop an automated system for generating image captions .
- Enhance accessibility for visually impaired individuals through image descriptions.

1.4 Application and Scope

The Image Caption Generator project has wide-ranging applications and significant scope, particularly in enhancing accessibility and improving user experience across various domains. This technology can be pivotal for visually impaired individuals, providing detailed descriptions of images and enabling better understanding of visual content on the internet. It can also be utilized in social media platforms for automatic tagging and captioning of photos, improving searchability and engagement. E-commerce plat-

forms can benefit by generating descriptive captions for product images, aiding in better product discovery and customer experience. Additionally, image captioning can be employed in digital archiving, content management systems, and education, where it can assist in organizing and describing large collections of visual data. The integration of advanced models combining CNNs and RNNs, such as Xception for feature extraction and LSTM for language generation, underscores the potential to create highly accurate and contextually relevant image descriptions, thus broadening the scope of practical applications in various fields.

1.5 Features

Features of this project are listed below:

- user login system.
- Provides real-time feedback for enhanced user experience.
- Design user-friendly interface for seamless image uploading and caption retrieval.

1.6 Feasibility Study

Feasibility is the determination of whether project is worth doing the process followed and making this determination is called feasibility study. This study helps us know if the project is viable and will also help govern different situations where there may arise problems regarding implementation. The feasibility of this project is discussed under the following headings:

1.6.1 Technical Feasibility

The technical feasibility of this system is strong, since it is based on advancements in technologies like supervised machine learning, where algorithms like cnn,rnn are employed to generate the cpation of the imagee provided. As it is driven by user-friendly and accessible technologies, with careful planning, robust infrastructure, and continuous improvement mechanisms, the system can be effectively developed, deployed, and maintained, delivering a reliable and efficient system and support platform.

1.6.2 Operational Feasibility

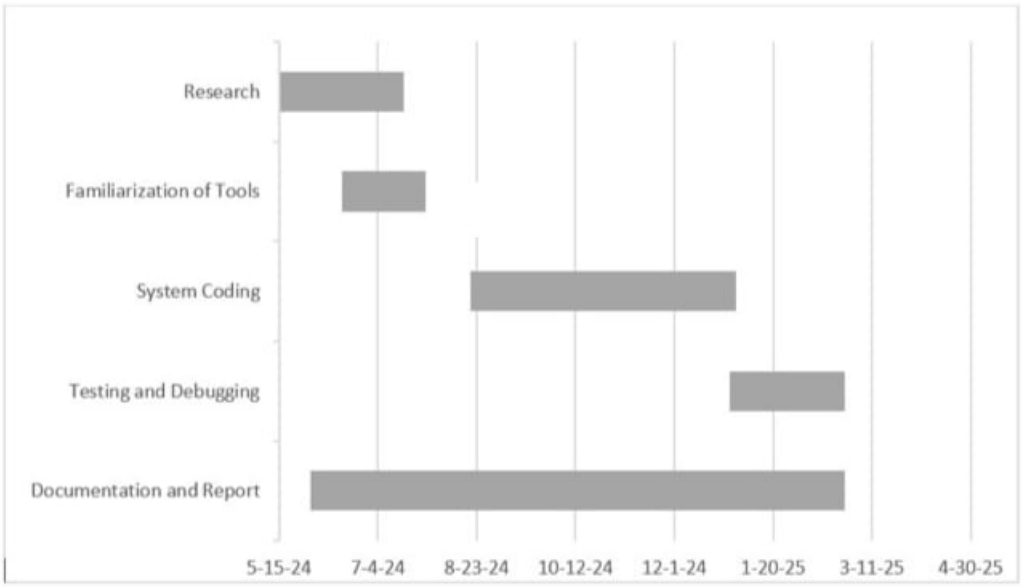
The operational feasibility of the proposed system is supported by factors such as user adoption and acceptance, stakeholder collaboration, technical support and maintenance, scalability and performance, integration with existing processes, and continuous improvement. By addressing these considerations, the system can be effectively implemented , ensuring user satisfaction, seamless integration, reliable operation, and the ability to adapt to evolving needs. This operational feasibility contributes to the system's practicality, effectiveness, and successful utilization .

1.6.3 Economic Feasibility

This project is economically feasible.Initially startup cost is high since we require datasets,development tools,however to balance these expenses, potential revenue can be generated through various streams such as Partnerships with organizations that support visually impaired individuals can further expand the market and provide additional revenue through specialized assistive technology applications. By effectively managing costs and leveraging diverse revenue opportunities, the project can achieve long-term economic feasibility.

1.6.4 Schedule Feasibility

Schedule Feasibility is defined as the probability of a project to be completed within its scheduled time limits, by a planned due date. If a project has a high probability to be completed on-time, then its schedule feasibility is appraised as high. Schedule feasibility ensures that a project can be completed before the technology becomes unnecessary. Since there are many features in our project but can be implemented in a quality way it has a very high probability to be completed on time.



‘

Figure 1.1: Gantt Chart

1.7 System Requirements

The system requirements for this project are as follows:

1.7.1 Development Requirements

Table 1.1: Development Requirements

Hardware Requirements	Software Requirements
CPU: Intel Core i3 or i5, or AMD Ryzen. RAM: Minimum 8GB. Storage: SSD minimum 512 GB. GPU: Accelerate performance in ML.	Python for backend development. CSS. HTML Javascript.

1.7.2 Deployment Requirements

Table 1.2: Deployment Requirements

Hardware Requirements	Software Requirements
Computer for development and testing purposes Adequate RAM and CPU	libraries and frameworks flask. flask-login SQLAlchemy ,Bootstrap

CHAPTER 2
LITERATURE REVIEW

2.1 Related Research

This paper presents a model based on CNN-LSTM neural networks that automatically detects objects in images and generates corresponding descriptions. By leveraging pre-trained models through Transfer Learning for object detection, the system combines Convolutional Neural Networks (CNNs) and Long-Short Term Memory (LSTM) networks to perform two key operations: object detection and image captioning. The interface for this model is developed using Flask REST API, a Python web development framework, with the primary aim of assisting visually impaired individuals in understanding their environment. The model utilizes CNNs, such as VGG, Inception, or YOLO, for training and detecting objects, and employs RNN-based LSTM networks to generate captions from the detected objects. Due to the large amount of data involved, traditional machine learning algorithms are inadequate, necessitating the use of deep learning and GPU-based computing for effective performance. The ResNet architecture is used for encoding image features, while LSTMs handle decoding. During training, the model is exposed to image features extracted by ResNet and a vocabulary built from training captions.[1]

This paper explores the use of computer vision and machine translation for image captioning, focusing on recognizing objects, actions, and attributes within images and identifying their relationships to generate descriptive text. Utilizing an encoder-decoder framework, the input image is encoded into an intermediary representation and then decoded into descriptions. The project employs the Flickr8k dataset and Python, aiming to develop a Flutter-based app that extracts image features and generates accurate descriptions, which can significantly aid visually impaired individuals and automate radiological tasks. The methodology includes image feature extraction using a CNN to capture key elements and context, text generation using an RNN (such as LSTM or GRU) to progressively form cohesive captions, and an attention mechanism to align image regions with generated words. The language model of the RNN is trained on a

dataset of images and captions, predicting the next word based on image features and preceding words. Evaluation metrics like BLEU, METEOR, and CIDEr assess the generated captions' quality, with refinements made to improve performance incrementally. Through these components, the project aims to create a robust image captioning system with practical applications in enhancing accessibility and automating medical tasks.[2].

This paper introduces an approach to image description and caption generation through a deep learning model combining computer vision and machine translation techniques. The goal is to detect objects within images, discern the relationships between these objects, and subsequently generate coherent captions. Leveraging the Flickr8k dataset and implementing Python3, also employing Transfer Learning, specifically utilizing the Xception model, to conduct the experiments. The paper delves into the architecture and functionalities of the neural networks utilized. Image caption generation holds significant importance in both Computer Vision and Natural Language Processing domains, with potential applications ranging from image segmentation in platforms like Facebook and Google Photos to aiding visually impaired individuals. This model utilizes a combination of Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks. An "encoder" RNN maps variable-length source sentences into fixed-length vector representations, serving as the initial hidden state for a "decoder" RNN responsible for generating meaningful captions. Notably, it also propose replacing the traditional RNN with a deep CNN, as it can provide a comprehensive representation of input images, pre-trained for image classification tasks, with its last hidden layer serving as input for the RNN decoder, ultimately enhancing sentence generation.[3]

This paper presents a model based on CNN-LSTM neural networks that automatically detects objects in images and generates corresponding descriptions. By leveraging pre-trained models through Transfer Learning for object detection, the system combines Convolutional Neural Networks (CNNs) and Long-Short Term Memory (LSTM) networks to perform two key operations: object detection and image captioning. The interface for this model is developed using Flask REST API, a Python web development framework, with the primary aim of assisting visually impaired individuals in understanding their environment. The model utilizes CNNs, such as VGG, Inception, or YOLO, for training and detecting objects, and employs RNN-based LSTM networks

to generate captions from the detected objects. Due to the large amount of data involved, traditional machine learning algorithms are inadequate, necessitating the use of deep learning and GPU-based computing for effective performance. The ResNet architecture is used for encoding image features, while LSTMs handle decoding. During training, the model is exposed to image features extracted by ResNet and a vocabulary built from training captions. Initial training phases showed low accuracy and irrelevant captions, but as training progressed to at least 20 epochs, the generated captions became more relevant. By 50 epochs, the model demonstrated significantly improved accuracy and coherence in its image descriptions.[4]

2.2 Proposed System

The proposed system for the image caption generator integrates advanced AI techniques with a user-friendly interface which will be built through several key phases, starting with requirement analysis and data collection, utilizing diverse datasets such as COCO and Flickr30k to provide a rich training foundation. Data preprocessing will involve standardizing image sizes, normalizing pixel values, and augmenting data.Text preprocessing will include tokenization, lowercasing, and the removal of punctuation and special characters, ensuring uniformity and reducing vocabulary size.

The system will utilize pre-trained Convolutional Neural Networks (CNNs) to extract high-level image features, which will then be fed into Recurrent Neural Networks (RNNs), specifically Long Short Term Memory (LSTM) units, for sequential caption generation. This approach ensures that the model can generate accurate and contextually relevant captions by learning the intricate relationships between images and their textual descriptions.

The architecture of the system will be designed to handle large volumes of data efficiently, incorporating data pipelines and storage solutions. The user interface will feature a login page, dashboards, and dedicated pages for different functionalities, for a responsive and interactive user experience.Extensive testing, including unit and integration testing, will ensure each component functions correctly and cohesively. This comprehensive and structured approach promises to deliver a powerful, accurate, and

user-friendly image captioning solution.

CHAPTER 3
METHODOLOGY

This methodology explained below provides insight to the various processes and methods which can be used to develop this image caption generator. It includes the AI integrated features in addition to the features inspired from Discord and Google Classroom. It guides us through requirement analysis and data collection phase, data preprocessing, design, development, testing phase, deployment and maintenance phase ensuring the development of quality and user-friendly image caption generator.

3.1 Data Sources

Data sources are where you get your data from for your project or analysis. They can be databases, websites, surveys, or any other place where you collect information.In image captioning, data sources provide pictures and their descriptions for training the model. These sources include diverse image datasets like COCO and Flickr30k.

3.2 Data Preprocessing

Data preprocessing for an image captioning generation system is a crucial step to ensure the model can effectively learn the relationship between images and their corresponding textual descriptions. This process involves several key steps:

- **Image preprocessing:** All the images are adjusted to a uniform size to maintain consistency. Images are cropped in such a way tha it doesnot affects the original image and helps in focusing the main content of image.

- **Normalization:** In the context of image data, normalization typically involves scaling the pixel values to a specific range and sometimes adjusting the distribution based on statistical properties of the dataset. Convert pixel values from the range [0, 255] to the range [0, 1] or [-1, 1], depending on the model requirements.

- **Data Augmentation:** Data augmentation is a technique used in machine learning, particularly in the context of computer vision and image processing, to artificially increase the size and diversity of a training dataset without actually collecting new data. This is achieved by applying various transformations to the existing data, creating modified versions of the original images.

3.3 Text Preprocessing

- **Tokenization:** In tokenization,split the captions into individual words or tokens. This can be done using simple whitespace tokenization or more sophisticated methods like subword tokenization.

- **Lowercasing:** Convert all words to lowercase to maintain uniformity and reduce the vocabulary size.

- **Removing Punctuation and Special Characters:** Strip out punctuation and special characters unless they are relevant to the context (e.g., apostrophes in contractions).

- **Truncation and Padding:** Truncation and padding are techniques used in text preprocessing to ensure that all text sequences (captions, sentences, etc.) in a dataset have the same length. This uniformity is crucial for batch processing and for feeding data into machine learning models, particularly neural networks, which often require inputs of fixed dimensions.

3.4 Dataset preparation

- **Image-Caption Pairing:** Ensuring each image is paired with its corresponding captions. Some images might have multiple captions, so the dataset should reflect these one-to-many relationships.

- **Encoding Captions:** Encoding captions is the process of transforming textual captions into a numerical format that can be fed into machine learning models. This involves converting each word (or token) in a caption into a corresponding numerical representation.

- **Split Dataset:**
  Splitting the dataset is a crucial step in preparing data for training machine learning models. It typically involves dividing the dataset into three subsets: training, validation, and test sets.It involves loading the dataset,creating data splits and maintaining consistency.

3.5 Feature Extraction for Images

- **Using Pre-trained CNNs:** Using a pre-trained Convolutional Neural Network (CNN) is a common and effective practice in image captioning tasks. Pre-trained CNNs, like those trained on the ImageNet dataset, are used to extract high-level features from images. These features are then fed into a sequence model (e.g., an LSTM or Transformer) to generate captions.

- **Storing Features:** Storing features extracted from a pre-trained CNN can help speed up the training process of your image captioning model, as you don't have to repeatedly extract features from the same images. Storing these feature vectors, which will be used as inputs to the captioning model during training and inference.

3.6 System Design and Architecture

Fig 3.1 describes the structure of the system. The Image Captioner system is designed to automatically generate textual descriptions of the image. To bridge the gap between

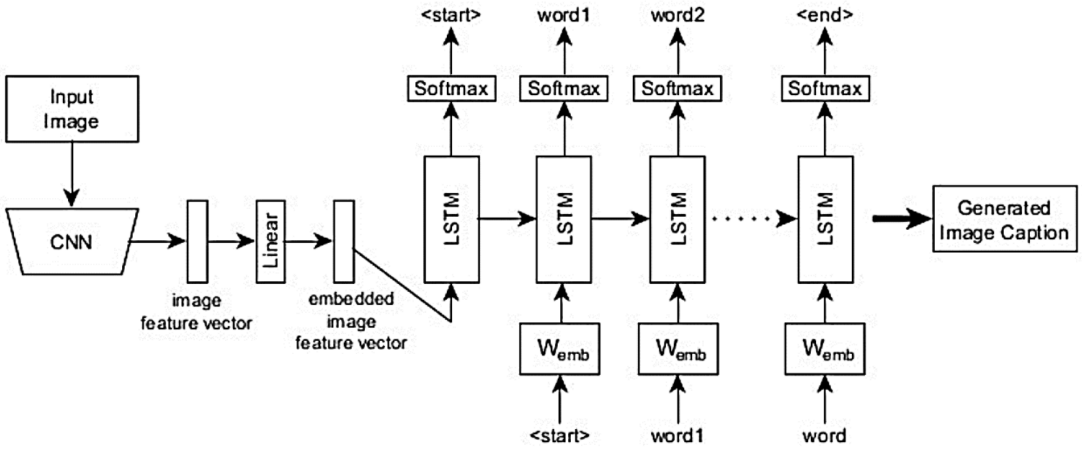


Figure 3.1: System Architecture

visual content and natural language, it employs deep learning techniques, particularly Recurrent Neural Networks (RNNs).

3.6.1 System Flowchart

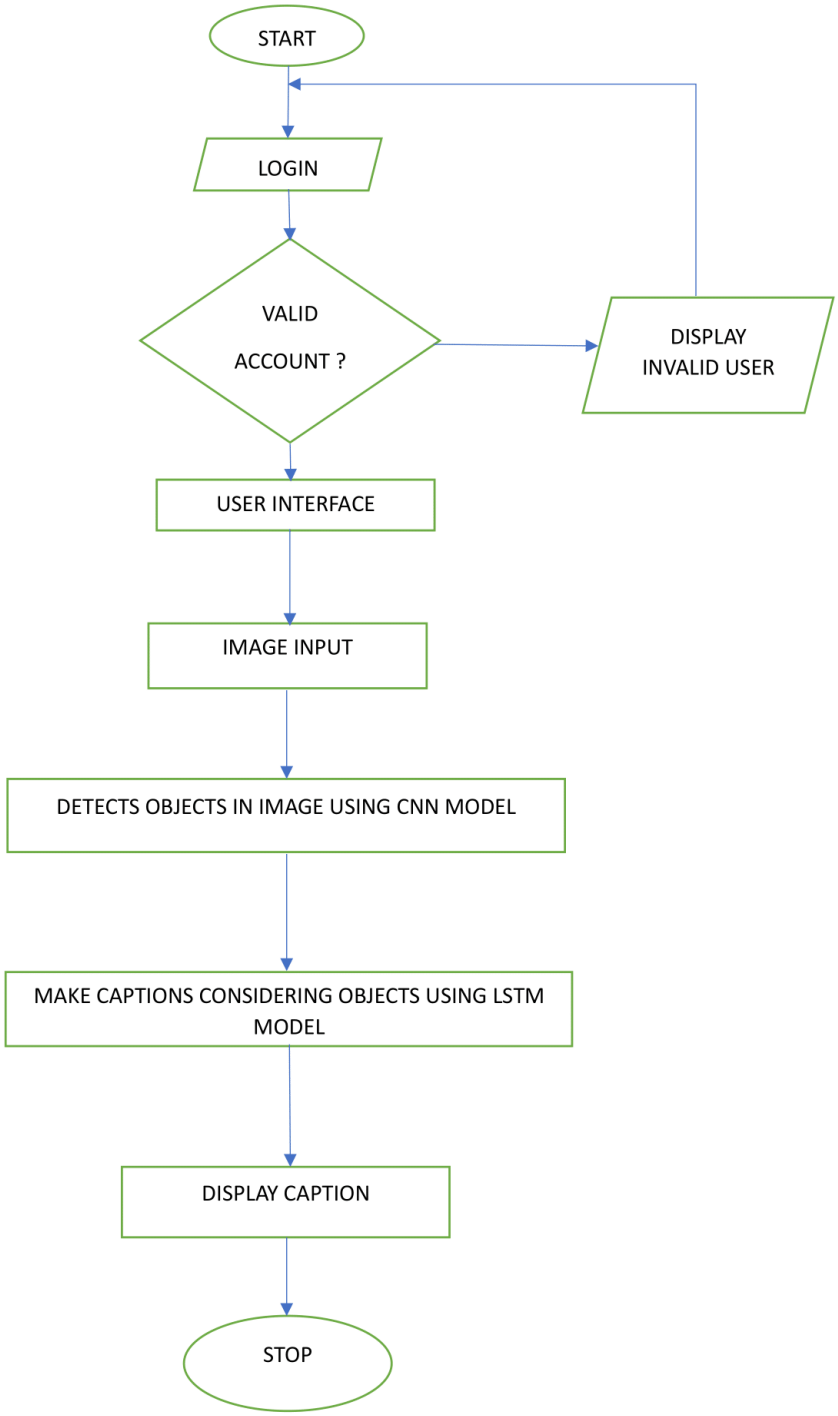


Figure 3.2: System flowchart

3.7 Algorithm

3.7.1 Convolutional Neural Network (CNN)

Convolutional Neural Networks (CNNs)are specialized deep neural networks which processes the data that has input shape like a 2D matrix. CNN works well with images and are easily represented as a 2D matrix. Image classification and identification can be easily done using CNN. It can determine whether an image is a bird, a plane or Superman, etc. Important features of an image can be extracted by scanning the image from left to right and top to bottom and finally the features are combined together to classify images. It can deal with the images that have been translated, rotated, scaled and changes in perspective. By integrating CNNs into image captioning systems, these models can effectively leverage the rich visual information present in images to generate coherent and semantically relevant captions.

3.7.2 Recurrent neural Network (RNN)

Recurrent Neural Networks (RNNs) are used in the caption generation stage to generate sequential descriptions of images based on the visual features extracted by Convolutional Neural Networks. The RNN takes the visual features as input and proceeds to generate words one at a time, considering both the previous word generated and its own internal state. This process continues until the model generates an end-of-sentence token or reaches a predefined maximum length for the caption. Throughout training, the parameters of both the CNN and the RNN are optimized jointly to minimize the difference between the generated captions and the ground truth captions provided in the training data.

3.7.3 Long Short Term Memory (LSTM)

LSTM are type of RNN (recurrent neural network) which is well suited for sequence prediction problems. We can predict what the next words will be based on the previous text. It has shown itself effective from the traditional RNN by overcoming the limi-

tations of RNN. LSTM can carry out relevantinformation throughoutthe processing, it discards non-relevant information.

3.7.4 Loss Function

Loss function quantifies the difference between the captions generated by the model and the actual captions provided in the training data. It guides the model to produce more accurate and contextually relevant captions during training by adjusting the model's parameters to minimize this discrepancy. Common loss functions include cross-entropy loss for word prediction and sequence loss for sequential tasks. The goal is to optimize the model to generate captions that closely resemble human-written captions for the input images.//

$$\text{Loss} = -\sum_{t=1}^{T} \log P(y_t \mid y_1, y_2, \ldots, y_{t-1}, \text{Image}) \tag{3.1}$$

where:

- T is the length of the sequence (caption).
- y_t is the actual word at time step t.
- $P(y_t \mid y_1, y_2, \ldots, y_{t-1}, \text{Image})$ is the predicted probability of the actual word y_t given the previous words $y_1, y_2, \ldots, y_{t-1}$ and the image features.

3.8 Design

- **System Architecture:** It includes designing the architecture of our system incorporating various features that makes our system more effective. It also includes planning the various data pipelines and storage solutions to handle large volume of data. The system architecture and flowchart of our system is shown above in system design phase.

- **AI Model Selection:** It is the step to select the most suitable and efficient machine learning models and algorithm for each of the image-captioning features.

- **UI/UX Design:** Designing the user interface is one of the most important steps for development of our system. We will be adding login page, dashboards and separate page for our features.

3.9 Development

3.9.1 Frontend Development

It is the development of the graphical user interface of the website using html, css and nextjs. Based on the UI design, we will develop our front end. We will implement interactive and responsive design principles for better user experience.

3.9.2 Backend Development

It refers to the development and maintenance of server-side logic and database that power the frontend of the web application. We will use Django for it. Similarly, implementation of web socket for real time updates will be done.

3.10 Testing

3.10.1 Unit Testing

Unit testing involves testing individual units or components of the software in isolation to verify that each unit performs as expected. In the context of an image captioning project, unit testing would involve testing specific functions or modules responsible for tasks such as image preprocessing, caption generation, or API endpoints.

3.10.2 Integration Testing

Integration testing would involve testing interactions between frontend and backend components, as well as interactions with external services or databases.

3.11 Deployment and Monitoring

Deploying and Monitoring the Model

- **Deployment:** We deploy the best performing model of our system to our web application.

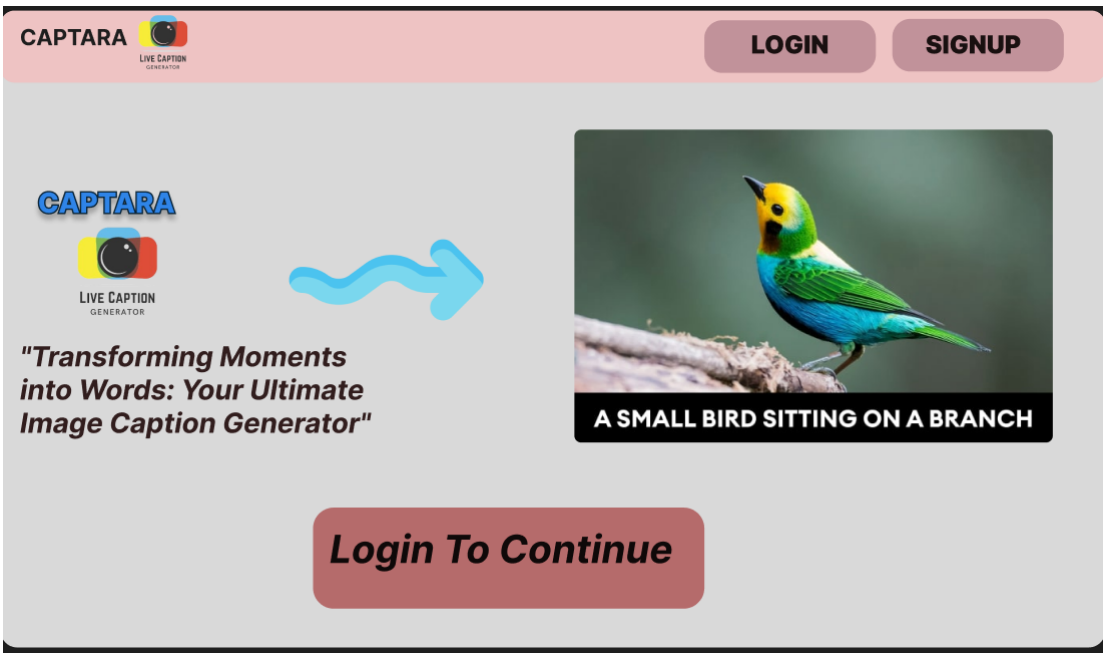
- **Monitoring:** Monitoring refers to the process of observing and tracking the performance, health, and behavior of a system or application in real-time. In image caption generation system, monitoring involves setting up tools and mechanisms to continuously monitor various aspects of the system.

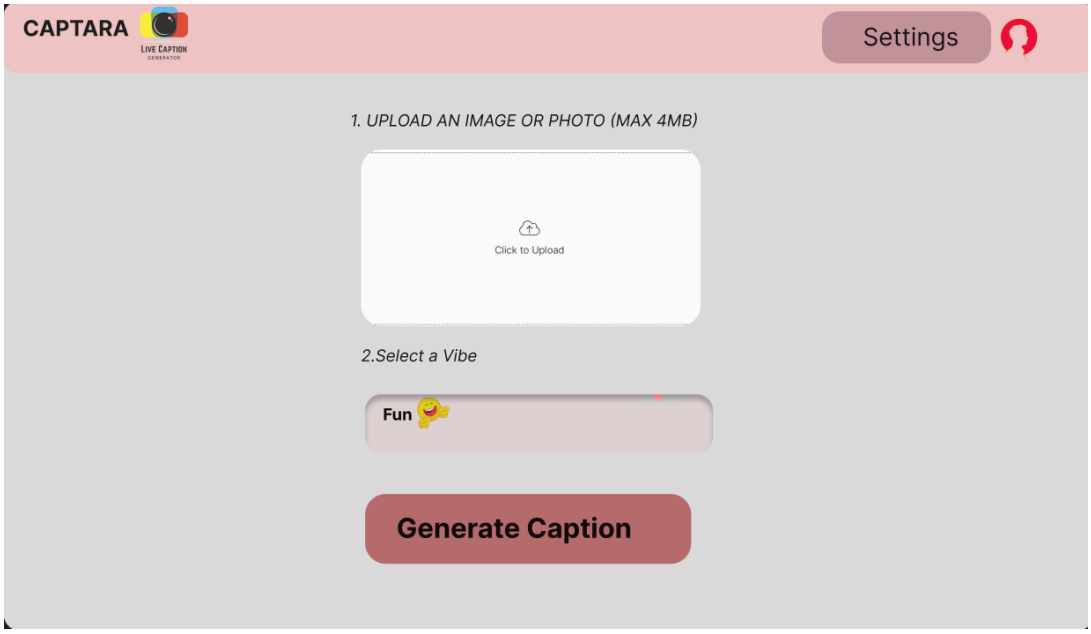
CHAPTER 4
EPILOGUE

4.1 Expected Outcome

The image caption generator developed through this methodology will provide accurate and contextually appropriate captions for a wide range of images. The system will be capable of understanding and describing images with high precision, leveraging advanced AI techniques and pre-trained models to ensure the generated captions are both coherent and semantically relevant. This will significantly enhance the usability of the system for various applications, such as assisting visually impaired users, automating image tagging, and improving search engine indexing of images.

Additionally, the final product will feature an intuitive and user-friendly interface, inspired by the design principles of Discord and Google Classroom. Users will benefit from a seamless experience with features like a login page, interactive dashboards, and dedicated pages for different functionalities. The system will be robust and scalable, capable of handling large volumes of data and providing real-time updates. The integration of comprehensive monitoring tools will ensure the system's performance and reliability, making it a valuable tool for end-users seeking efficient and accurate image captioning solutions.


CAPTARA
LIVE CAPTION
GENERATOR
LOGIN
SIGNUP
CAPTARA
LIVE CAPTION
GENERATOR
"Transforming Moments into Words: Your Ultimate Image Caption Generator"
A SMALL BIRD SITTING ON A BRANCH
Login To Continue


CAPTARA
LIVE CAPTION
GENERATOR
Settings
1. UPLOAD AN IMAGE OR PHOTO (MAX 4MB)
Click to Upload
2.Select a Vibe
Fun
Generate Caption

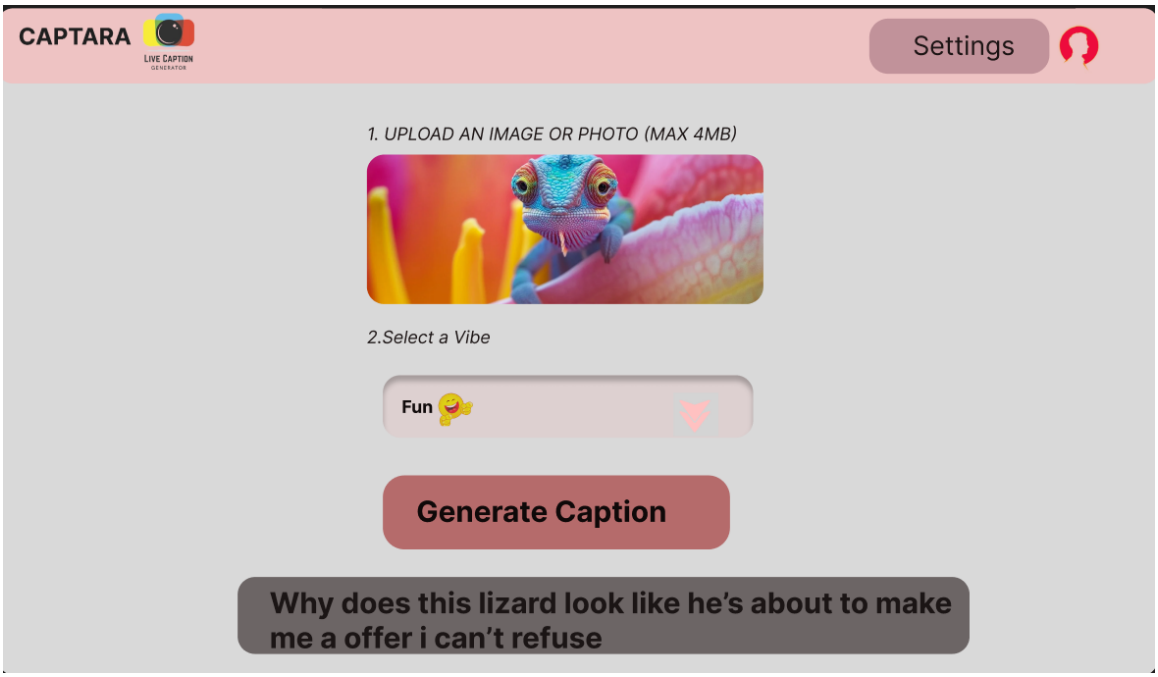


Figure 4.1: expected outcome

REFERENCES

[1] ghorbanali2022ensembleGhorbanali, A., Sohrabi, MK. Yaghmaee, F. 2022. Ensemble transfer learning-based multimodal sentiment analysis using weighted convolutional neural networks Ensemble transfer learning-based multimodal sentiment analysis using weighted convolutional neural networks. Information Processing & Management593102929.

[2] articleMohamed, A. 202005. Image Caption using CNN LSTM Image caption using cnn lstm.

[3] panicker2021imagePanicker, MJ., Upadhayay, V., Sethi, G. Mathur, V. 2021. Image caption generator Image caption generator. International Journal of Innovative Technology and Exploring Engineering (IJITEE)10387–92.

[4] vallabhaneni2023segmentationVallabhaneni, N. Prabhavathy, P. 2023. Segmentation quality assessment network-based object detection and optimized CNN with transfer learning for yoga pose classification for health care Segmentation quality assessment network-based object detection and optimized cnn with transfer learning for yoga pose classification for health care. Soft Computing1–23.