A Project report on

GENERATING IMAGE CAPTIONS BASED ON DEEP LEARNING AND NATURAL LANGUAGE PROCESSING

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

BACHELOR OF TECHNOLOGY
in
COMPUTER SCIENCE & ENGINEERING

By

M. BHARGAVI (204G1A0521)

Under the Guidance of

Mrs. P. Rohini M. Tech.,(Ph.D.)
Assistant Professor

Department of Computer Science & Engineering

SRINIVASA RAMANUJAN INSTITUTE OF TECHNOLOGY
(AUTONOMOUS)
Rotarypuram Village, B K Samudram Mandal, Ananthapuramu – 515701

2023-2024

SRINIVASA RAMANUJAN INSTITUTE OF TECHNOLOGY

(AUTONOMOUS)

(Affiliated to JNTUA, Accredited by NAAC with ‘A’ Grade, Approved by AICTE, New Delhi &
Accredited by NBA (EEE, ECE & CSE)
Rotarypuram Village, BK Samudram Mandal, Ananthapuramu-515701

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

Certificate

This is to certify that the Project report entitled GENERATING IMAGE CAPTIONS BASED ON DEEP LEARNING AND NATURAL LANGUAGE PROCESSING is the bonafide work carried out by **M. Bhargavi** bearing Roll Number **204G1A0521** in partial fulfilment of the requirements for the award of the degree of **Bachelor of Technology** in **Computer Science & Engineering** during the academic year 2023 - 2024.

Project Guide
Mrs. P. Rohini M.Tech., (Ph.D.)
Assistant Professor

Head of the Department
Mr. P. Veera Prakash M.Tech., (Ph.D.)
Assistant Professor

Date:

Place: Rotarypuram

External Examiner

DECLARATION

I, Ms. M. Bhargavi with reg no: 204G1A0521 student of SRINIVASA RAMANUJAN INSTITUTE OF TECHNOLOGY, Rotarypuram, hereby declare that the dissertation entitled **“GENERATING IMAGE CAPTIONS BASED ON DEEP LEARNING AND NATURAL LANGUAGE PROCESSING”** embodies the report of my project work carried out by me during IV year Bachelor of Technology under the guidance of **Mrs. P. Rohini,** Assistant Professor, Department of CSE, **SRINIVASA RAMANUJAN INSTITUTE OF TECHNOLOGY**, and this work has been submitted for the partial fulfillment of the requirements for the award of the Bachelor of Technology degree.

The results embodied in this project have not been submitted to any other Universities of Institute for the award of Degree.

M. BHARGAVI Reg no: 204G1A0521

VISION & MISSION OF THE INSTITUTION

Vision:

To become a premier Educational Institution in India offering the best teaching and learning environment for our students that will enable them to become complete individuals with professional competency, human touch, ethical values, service motto, and a strong sense of responsibility towards environment and society at large.

Mission:

- Continually enhance the quality of physical infrastructure and human resources to evolve in to a center of excellence in engineering education.
- Provide comprehensive learning experiences that are conducive for the students to acquire professional competences, ethical values, life-long learning abilities and understanding of the technology, environment and society.
- Strengthen industry institute interactions to enable the students work on realistic problems and acquire the ability to face the ever-changing requirements of the industry.
- Continually enhance the quality of the relationship between students and faculty which is a key to the development of an exciting and rewarding learning environment in the college.

VISION & MISSION OF THE DEPARTMENT OF CSE

Vision:

To evolve as a leading department by offering best comprehensive teaching and learning practices for students to be self-competent technocrats with professional ethics and social responsibilities.

Mission:

DM 1: Continuous enhancement of the teaching-learning practices to gain profound knowledge in theoretical & practical aspects of computer science applications.

DM 2: Administer training on emerging technologies and motivate the students to inculcate self-learning abilities, ethical values and social consciousness to become competent professionals.

DM 3: Perpetual elevation of Industry-Institute interactions to facilitate the students to work on real-time problems to serve the needs of the society.

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompany the successful completion of any task would be incomplete without the mention of people who made it possible, whose constant guidance and encouragement crowned our efforts with success. It is a pleasant aspect that we have now the opportunity to express our gratitude for all of them.

It is with immense pleasure that we would like to express our indebted gratitude to our Guide **Mrs. P. Rohini, Assistant Professor, Computer Science & Engineering**, who has guided us a lot and encouraged us in every step of the project work. We thank her for the stimulating guidance, constant encouragement and constructive criticism which have made possible to bring out this project work.

We express our deep felt gratitude to **Mr. C. Lakshminatha Reddy, Assistant Professor** and **Mr. M. Narasimhulu, Assistant Professor,** Project Coordinators for their valuable guidance and unstinting encouragement enabled us to accomplish our project successfully in time.

We are very much thankful to **Mr. P. Veera Prakash, Assistant Professor & Head of the Department, Computer Science & Engineering,** for his kind support and for providing necessary facilities to carry out the work.

We wish to convey our special thanks to **Dr. G. Balakrishna, Principal** of **Srinivasa Ramanujan Institute of Technology** for giving the required information in doing our project work. Not to forget, We thank all other faculty and non- teaching staff, and our friends who had directly or indirectly helped and supported us in completing our project in time.

We also express our sincere thanks to the Management for providing excellent facilities**.**

Finally, we wish to convey our gratitude to our families who fostered all the requirements and facilities that we need.

Project Associate
204G1A0521

ABSTRACT

Humans and computers are attempting to communicate because everything in today's society depends on systems like computers, mobile phones, etc. This is how our project is visualized. Our undertaking People with visual impairments can benefit from the creation of image captions. Computers are unable to distinguish objects, things, or activities with the same ease as humans. To recognize them, they require some training. The suggested method is used to identify activities or similar items. We offer several deep neural network-based models for creating captions for images, with a particular emphasis on CNNs (Convolutional Neural Networks) that extract characteristics from the image. Using LSTM (Long Short-Term Memory) techniques, RNNs (Recurrent Neural Networks) create captions based on the image's attributes. and examining how they affect the construction of sentences. Here, encoder-decoders are used to create a link between descriptions from natural language processing and visual information such as image features. The process of generating a caption's sequence is handled by the decoder, while the encoder extracts features. In order to determine which feature extraction and encoder model produces the best results and accuracy, we have also created captions for sample photos and compared them with one another. We also introduce Deep Voice, a text-to-speech system of production quality that uses only deep neural networks to generate captions based on visual attributes. The evaluation of our project will be conducted utilizing several machine learning methods and Python.

Keywords

CNN, RNN, LSTM, Encoder - Decoder.

CONTENTS

List of Figures

LIST OF ABBREVIATIONS

CNN	Convolutional Neural Network
LSTM	Long Short Term Memory
SRS	System Requirements Specification
RNN	Recurrent Neural Network
RGB	Red Green and Blue
ResNet	Residual Network
DLL	Dynamic Load Libraries
UML	Unified Modelling Language

CHAPTER 1

INTRODUCTION

It is relatively easy for humans to describe the environments they are living. It is normal for a human to be able to quickly describe vast number of details about an image. This is a fundamental human ability. The ability to identify objects and describe images is facilitated by the human brain. Artificial Intelligence introduces numerous algorithms that are based on the architecture of the brain. Here, human beings are employing these algorithms to mimic human visual world interpretation on computers. Despite significant advancements in computer vision fields including object identification, picture classification, attribute classification, and scene recognition. To allow a computer the automatically explains an image, which has been forwarded to it using a language that resembles a person, is a relatively new undertaking. Generating image captions is a process that automatically produces caption for the given image using a computer. It is a difficult task. Image captioning necessitates both a high-level comprehension of an image's semantic contents and the ability to convey the information in a sentence that sounds human because it connects computer vision with natural language processing. Giving computers visual access to the outside world will have a variety of benefits, such as facilitating human-robot communication, promoting early learning, facilitating knowledge retrieval, and providing assistance to the visually impaired. It is a significantly a new endeavor to make a computer automatically explain an image that is forwarded to it in a language that sounds human.

The mechanism of providing a human readable text for an image with a computer known as image captioning. It is challenging task. Because picture captioning combines Machine vision and natural language processing, that needs both high level of knowledge of an image's semantic contents and the ability to transmit the information in a human-sounding sentence. Full of reasonable applications, will become possible if computers are given access to the visual world.

Convolutional neural networks (CNN) algorithm is mainly used to obtain characteristics from the picture, while recurrent neural networks (RNN) used to produce sequence of words in a meaningful way which may easily understandable to human beings based on the image. During first stage, we have taken a novel technique to extracting features from a picture, which will provide us with details the smallest change among comparable images, instead of just detecting the living or non-living things present in the picture. The 16 convolutional layer model VGG-16 (Visual Geometry Group) has been employed by us for object recognition.

We must train the dataset's captions in the second stage. In order to construct our captions from the input photos, we use GRU (Gated Recurrent Unit) and LSTM (Long Short-Term Memory). We have measured the accuracy of different algorithms using the BLEU (Bilingual Evaluation Understudy) in order to estimate which architecture is best. For evaluating the caliber of translations produced by machines, BLUE offers a numerical score. Our goal is to achieve greater accuracy than previous efforts. We are utilizing the fickr30K dataset to guarantee more accuracy. Images are fed into a computer system as two-dimensional arrays, and the images are mapped to descriptive phrases or captions. In the last few years, the task of automatically produces captions for images has received a lot of curiosity.

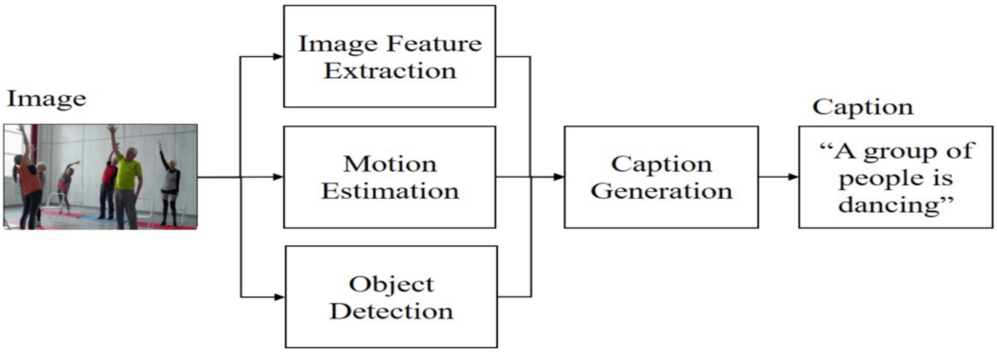


Fig1. Block Diagram

1.1Problem Statement

The project's objective is to create a model that can produce meaningful and cogent textual descriptions for supplied photos in order to tackle the problem of automatic image captioning. The particular challenge is establishing a productive synergy between natural language comprehension and visual content. Making use of the dataset, the challenge entails combining word embeddings from GloVe, extracting discriminative features using the ResNet50 CNN, and preprocessing and cleaning a variety of captions. The difficulty is in teaching a model to produce precise and contextually relevant captions by capturing the links between verbal and visual aspects. The model's performance is measured by BLEU scores, which indicate how well the model generates captions that correspond with human-generated references. Resolving this issue effectively advances natural language processing and computer vision, improving robots' ability to interpret and describe visual information.

The system consists of two main modules:
i. System Module – In this Creating database, Pre-Processing, Training will be done.
ii. User Module – In this module User can register and login with their credentials to use the system.

1.2 Objectives

To accomplish the project's purpose, the following particular objectives have been established.

i. To efficiently train the dataset utilizing the flickr30k dataset in order to develop a deep learning-based picture captioning system that is more accurate than advanced.

ii. To turn the generated image description into we use text-to-speech technology.

1.3 Scope of the Project

The following are the boundaries that have established in the proposed system which defines scope.

i. The project's goal is to improve accessibility by giving photos relevant and accurate captions. This could help those who are visually impaired by improving their ability to understand and engage with visual content.

ii. The study is in line with the advancement of multimodal AI systems that can comprehend and produce material in a variety of modalities, particularly text and visuals. This involves investigating the interrelationships between sequential modeling (LSTMs) and enhanced visual feature extraction (ResNet-50).

iii. The project's practical applications encompass domains like picture indexing, content retrieval, and alternative text descriptions for photographs on internet platforms. The project's results are intended to have concrete effects on enhancing user experiences across a range of industries.

iv. The project helps to comprehend the capabilities of these models in the particular context of picture captioning by employing cutting-edge deep learning architectures. The deeper field of deep learning study may benefit from the discoveries made during this investigation.

CHAPTER 2
LITERATURE SURVEY

[1] K Harika, Rajan Singh, M Sailaja, B Sridhar, In recent years, deep neural networks have enabled the captioning of images. Based on the dataset, the photo caption generator assigns a suitable title to an applied input image. The current study suggests a deep learning-based model and applies it to produce a caption for the input image. The model frames a statement associated with an input image by utilizing CNN and LSTM algorithms. This CNN model recognizes the objects in the picture, and the Long Short-Term Memory (LSTM) algorithm generates text as well as caption that fits the project. Thus, the primary goals of the suggested model are object recognition and title generation for the input images.

[2] Kalaivani S P, Karthika S, Karthika V, C S Kanimozhiselvi, Writing a description for an image is known as image captioning. It is among the most important and contemporary research problems at the moment. Every day, new methods for solving the problem are developed. Even with so many possibilities available, more work needs to be done in order to get more accurate and superior results. We thus considered developing an image captioning model that combines multiple Long Short-Term Memory and Convolutional Neural Network architecture configurations in order to attain superior results. Three CNN and LSTM combinations were combined to create the model. The proposed model is trained using three Convolutional Neural Network architectures: Xception, ResNet50, and Inception-v3. These networks are used to use LSTM to extract features from the image and generate pertinent captions. The three CNN and LSTM combinations that perform the best are selected based on the accuracy of the model. The model is trained on Flicker8k dataset.

[3] Vaishali Jabade, Amritkar Chetan, Artificial Intelligence automatically synthesizes the information of images using machine vision and natural language processing (NLP). A neuronal regeneration model is developed. Both machine translation and vision are involved. This process results in the creation of organic sentences that ultimately describe the visual. Recurrent neural networks (RNN) and convolutional neural networks (CNN) make up this paradigm. RNNs are used to construct sentences, whereas CNN is used to extract features from images. When given an input image, the model is trained to provide captions that, in essence, explain it. Various datasets are used to assess the language proficiency and accuracy of the model learned from picture descriptions. The results of these tests demonstrate that the model often provides precise descriptions of the input image.

[4] Vaidehi Muley, Megha Kolhekar, Varsha Kesavan, The project focus on generating captions by utilizing content of the image as a source. At the moment, human annotation is required for photos, which makes the process almost impossible for commercial datasets. Using the picture database, the Convolutional Neural Network (CNN) encoder extracts features and subtleties from the image to produce a "thought vector." The objects and features in the picture are then translated by an RNN (Recurrent Neural Network) decoder to create a coherent and sequential caption for an input. In this study, we thoroughly analyze multiple deep neural network-based approaches for producing photo captions and pretrained models, with the goal of determining the most efficient model with fine-tuning. To maximize the model's caption-generating capabilities, they compared models that included and did not include the "attention" concept. To provide a more meaningful comparison, the same dataset is used to train each model.

[5] E K Miller, T J Buschman, In the human visual system, bottom-up signals linked to unexpected, unusual can automatically focus attention, as can top-down signals dictated by the present task. In this work, they use comparable language to designate as “top-down” attention mechanisms those that are handle by nonvisual or task-specific context, and a “bottom-up” attention mechanisms that are solely visual feed-forward. The majority of traditional visual attention processes utilized in VQA and picture captioning are top-down in nature. These methods, which are frequently trained to focus on the output of a convolutional neural network (CNN) layer or layers, use an image's caption or a query about it as their context. Nevertheless, this method pays minimal attention to the process of selecting the visual parts that require attention. It is a time taking process to pre-process the image.

CHAPTER 3
METHODOLOGY

As you can see, each image are presented in the matrix formats, which are made up of rows and columns. The pixel is an image's fundamental building block. A group of pixels make up an image. These are all little squares. We may build the entire image by arranging them side by side. The smallest amount of information that can be present in an image is a single pixel. Every image has pixels with values ranging from 0 to 255.

Each pixel is composed of Three values are R, G, and B, which are the basic colours red, green, and blue. The combination of these three basic colours will create all these colours here in the image so we conclude that a single pixel has three channels, one channel for each one of the basic colours.

3.1 Deep Learning

Deep learning which is a part of machine learning, was first demonstrated by using real neurons from the brain and transforming them into artificial neural networks using learning techniques. No matter how well-optimized, machine learning approaches begin to lose accuracy and performance as data volume increases, but deep learning performs far better in these situations. Many domains, including image classification, speech recognition, recommendation systems, NLP (natural language processing), etc., use deep learning. In light of all of these factors, we have decided to create our project using deep learning techniques. Features are directly learned by Deep Learning from the data. Neural networks automatically find and express pertinent patterns during training, eliminating the need for human feature creation. Deep learning models can generalize well to any variety of complicated datasets thanks to this capability. The primary areas of attention for this project, "Generating Image Captions by using CNN and NLP," are feature extraction and classification. That being said, deep learning was chosen primarily for this reason. The goals of this study are to produce more accurate results for appropriate caption generation based on image attributes utilizing the flickr30K dataset.

3.2 Algorithm Used

3.2.1 Convolutional Neural Networks (CNN)

You can see that each image is represented using a matrix format, which consists of rows and columns. The fundamental unit of an image is the pixel. An image is made up of a set of pixels. Each of these is a little square. By placing them side by side, we can construct the full image. A single pixel is the lowest unit of information that may be found in an image. Pixels in every image have values ranging from 0 to 255.

Each pixel is composed of Three values are R, G, and B, which are the basic colours red, green, and blue. The combination of these three basic colours will create all these colours here in the image so we conclude that a single pixel has three channels, one channel for each one of the basic colours.

Over the past 20 years Deep Learning's capacity to manage massive volumes of data has demonstrated its strength as a technology. Convolutional neural networks, also referred to as CNNs or ConvNets, are the most popular deep neural networks in deep learning, especially for applications in computer vision. The type of deep neural network utilized in deep learning that is most commonly used to evaluate visual data is called a convolutional neural network (CNN). It uses a unique method known as convolution. As it currently understood, When two functions are mathematically operated upon, a third function known as convolution is produced that describes how one's shape is transformed by another.

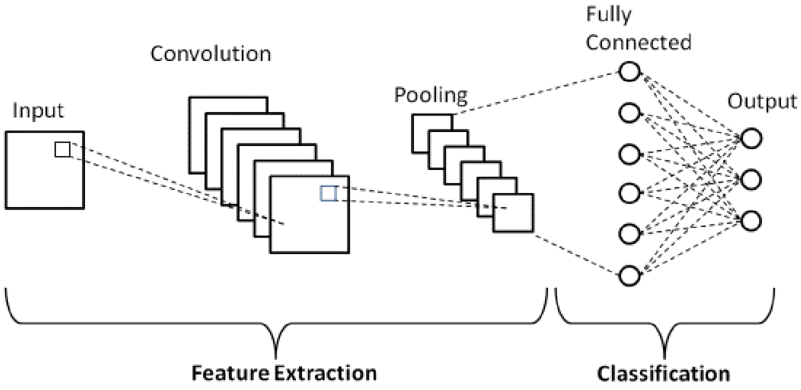


Fig 3.1: Basic Process of CNN Algorithm

Convolutional, fully-connected, and pooling layers are the three layers that make up a CNN. Convolutional, fully-connected and pooling layers are stacked to construct a CNN architecture. Beyond these layers, two more important factors are the dropout layer and the activation function. The RGB (Red, Green, and Blue) color space, which extends from 0 to 255, can be used to extract and identify various visual aspects for analysis (a process known as feature extraction) using a convolution tool. The network used for feature extraction is composed of multiple pairs of convolutional or pooling layers.

3.2.2 Long Short-Term Memory

The Long Short-Term Memory, or LSTM, is an enhanced RNN. For sequence prediction tasks, LSTM performs remarkably well in capturing long-term dependencies. The RNN method has certain drawbacks, which we address by introducing the LSTM algorithm.

It is a difficult task to a network to learn long-term dependencies in a standard RNN since it only has one hidden state that is retained over time. In contrast, LSTMs solve this issue by introducing memory cells, which are containers that can store information for a longer amount of time. Three gates govern a memory cell: Input, Forget and Output gates.

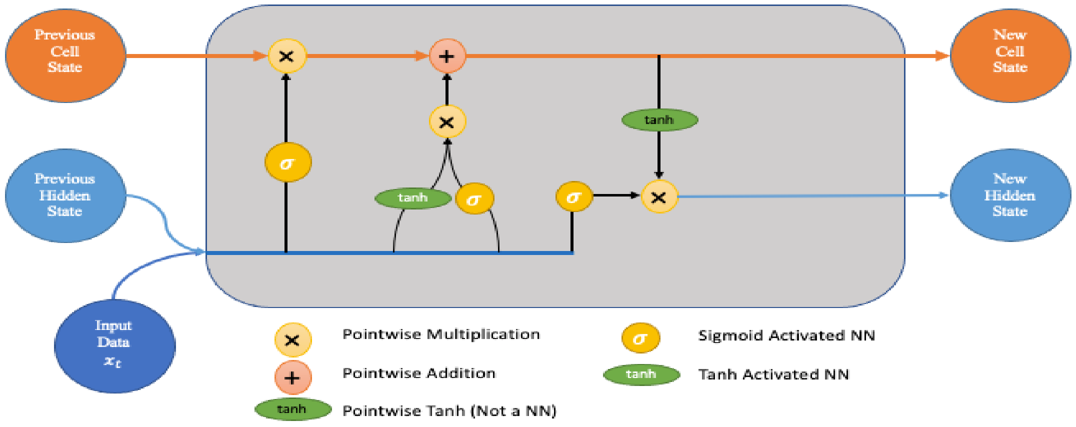


Fig 3.2: Process of LSTM

Input, Forget and Output gates are determine which data should be input into, taken out of, and output from the memory cell.

3.2.3 ResNet 50:

ResNet (Residual Network) was introduced to solve the vanishing gradient problem that arises when deep convolutional neural networks (CNNs) are being trained. ResNet gives the network the ability to learn residual functions by introducing skip connections, also referred to as residual connections. The model can bypass specific levels thanks to these skip connections, which send the image straight to the output layers. This facilitates the training of extremely deep networks by reducing the impact of the vanishing gradient issue.

ResNet can be a key component in feature extraction from photos within an image caption generator. In order to obtain useful features from the images, the encoder portion of an image captioning model usually makes use of a CNN that has already been trained, such as ResNet.

The goal is to extract high-level features from photos by using the information that the pre-trained ResNet model has acquired on a sizable dataset.

- Images are used to obtain features using the ResNet model. For image classification tasks, the model is usually pre-trained on a large dataset. Hierarchical, abstract aspects in photos are captured by the weights that were learned during pre-training.

- The pre-trained ResNet model is used to obtain characteristics from intermediate layers given an input image. The input's high-level visual information is represented by the features.

- The decoder component of the picture captioning model receives the features that were extracted from the image. Based on the input attributes, the decoder—which is frequently implemented as a transformer or recurrent neural network (RNN)—creates a textual description of the image.

CHAPTER 4

SYSTEM REQUIREMENTS SPECIFICATIONS

A software requirements specification (SRS) is a description of a software system to be developed. It lays out functional and nonfunctional requirements, and may include a set of use cases that describe user interactions that the software must provide. It is very important in a SRS to list out the requirements and how to meet them. It helps the team to save upon their time as they are able to comprehend how are going to go about the project. Doing this also enables the team to find out about the limitations and risks early on.

A SRS can also be defined as a detailed description of a software system to be developed with its functional and non-functional requirements. It may include the use cases of how the user is going to interact with the software system. The software requirement specification document is consistent with all necessary requirements required for project development. To develop the software system we should have a clear understanding of Software system. To achieve this we need continuous communication with customers to gather all requirements.

A good SRS defines how the Software System will interact with all internal modules, hardware, and communication with other programs and human user interactions with a wide range of real life scenarios. It is very important that testers must be cleared with every detail specified in this document in order to avoid faults in test cases and its expected results.

4.1 Functional and non-functional requirements

Requirement’s analysis is very critical process that enables the success of a system or software project to be assessed. Requirements are generally split into two types: Functional and non-functional requirements.

Functional Requirements: These are the requirements that the end user specifically demands as basic facilities that the system should offer. All these functionalities need to be necessarily incorporated into the system as a part of the contract. These are represented or stated in the form of input to be given to the system, the operation performed and the output expected. They are basically the requirements stated by the user which one can see directly in the final product, unlike the non-functional requirements.

Benefits of functional requirements:

- Helps you to check whether the application is providing all the functionalities that were mentioned in the functional requirement of that application
- A functional requirement document helps you to define the functionality of a system or one of its subsystems.
- Functional requirements along with requirement analysis help identify missing requirements. They help clearly define the expected system service and behavior.
- Errors caught in the Functional requirement gathering stage are the cheapest to fix.
- Support user goals, tasks, or activities

Examples of functional requirements:

1) Authentication of user whenever he/she logs into the system
2) System shutdown in case of a cyber-attack
3) A verification email is sent to user whenever he/she register for the first time on some software system.

Non-functional requirements: These are basically the quality constraints that the system must satisfy according to the project contract. The priority or extent to which these factors are implemented varies from one project to other. They are also called non-behavioral requirements.

Benefits of Non-Functional Requirements:

- The nonfunctional requirements ensure the software system follows legal and compliance rules.
- They ensure the reliability, availability, and performance of the software system.
- They ensure good user experience and ease of operating the software.
- They help in formulating security policy of the software system.

They basically deal with issues like:

- Portability
- Security
- Maintainability

- Reliability
- Scalability
- Performance
- Reusability
- Flexibility

Examples of non-functional requirements:

1) Emails should be sent with a latency of no greater than 12 hours from such an activity.
2) The processing of each request should be done within 10 seconds
3) The site should load in 3 seconds whenever of simultaneous users are > 10000

Requirements

1. The access permissions for system data may only be changed by the system's data administrator.
2. Passwords shall never be viewable at the point of entry or at any other time.
3. Application should be able to adapt themselves to increased usage or be able to handle more data as time progresses.
4. Application should be responsive to the user Input or to any external interrupt which is of highest priority and return back to the same state.
5. Users should be able to understand the flow of the Application easily i.e. users should be able to use the Application without any guideline or help from experts/manuals.
6. There should be a common plan where the user can access the application to install and look for regular updates to give feedback.
7. The application should be able to render it's layout to different screen sizes. Along with automatic adjustment of Font size and image rendering.
8. The application should run at a speed that is desirable by the users. A slow application can lead to frustration and hence, will not be preferred over other faster applications.
9. The application must be stable. It should never crash or force close in the case of many users using it simultaneously.
10. The application must be easy to maintain.
11. It must be user-friendly. Having a user-friendly application is of key importance for the success of the application.

4.2 Basic Requirements

1. Data collection: The dataset can be collected at the time voter registration in this project. It consists of First name, middle name (optional), last name, generated pin number, Roll number, Email address, phone number, age, States and districts.

2. Data Preprocessing: The purpose of preprocessing is to convert raw data into a form that fits machine learning. Structured and clean data allows a data scientist to get more precise results from an applied machine learning model. The technique includes data formatting, cleaning, and sampling. Here, data pre-processing focuses on finding the attributes with null values or invalid values and finding the relationships between various attributes as well. Data Pre-processing also helps in finding out the impact of each parameter on the target parameter. To preprocess our datasets we used EDA methodology. All the invalid and null values were handled by removing that record or giving the default value of that particular attribute based on its importance.

3. Model training: After a data scientist has preprocessed the collected data and split it into train and test can proceed with a model training. This process entails ―feeding‖ the algorithm with training data. An algorithm will process data and output a model that is able to find a target value (attribute) in new data an answer you want to get a predictive analysis. The purpose of model training is to develop a model. We trained our model using the random forest algorithm. On training the model it predicts the yield on giving the other attributes of the dataset as input.

4.Model evaluation and testing: The goal of this step is to develop the simplest model able to formulate a target value fast and well enough. A data scientist can achieve this goal through model tuning. That's the optimization of model parameters to achieve an algorithm‘s best performance.

4.3 Application Requirements

1. Users must be able to register as a new user.
2. Users should be able to login if they already have registered.
3. User can able to update the details.
4. The admin must check the user login credentials at the time login.

5. The user should be able to choose an image to those they want caption after accepting the user by the admin.

6. All the modules of the application must work in a proper manner.

7. The predictions must be accurate.

8. Users must be able to access the Fertilizers module as well.

9. Generating Image Captions Based on Deep Learning and Natural Language Processing project is helps to generate captions.

10. The user must be able to logout.

HARDWARE REQUIREMENTS

Processor	-I3/Intel Processor
Hard Disk	- 160GB
Monitor	- SVGA
RAM	- 8GB

SOFTWARE REQUIREMENTS:

Operating System	:	Windows 7/8/10
Server Side Script	:	HTML, CSS, Bootstrap & JS
Programming Language	:	Python
Libraries	:	Flask, Pandas, Mysql.connector, Os, Smtplib, Numpy
IDE/Workbench	:	PyCharm
Technology	:	Python 3.6+
Server Deployment	:	Xampp Server
Database	:	MySQL

4.4 Python Libraries:

Normally, a library is a collection of books or is a room or place where many books are stored to be used later. Similarly, in the programming world, a library is a collection of precompiled codes that can be used later on in a program for some specific well-defined operations. Other than pre-compiled codes, a library may contain documentation, configuration data, message templates, classes, and values, etc.

A Python library is a collection of related modules. It contains bundles of code that can be used repeatedly in different programs. It makes Python Programming simpler and convenient for the programmer. As we don’t need to write the same code again and again for different programs. Python libraries play a very vital role in fields of Machine Learning, Data Science, Data Visualization, etc.

Working of Python Library

As is stated above, a Python library is simply a collection of codes or modules of codes that we can use in a program for specific operations. We use libraries so that we don't need to write the code again in our program that is already available. But how it works. Actually, in the MS Windows environment, the library files have a DLL extension (Dynamic Load Libraries). When we link a library with our program and run that program, the linker automatically searches for that library. It extracts the functionalities of that library and interprets the program accordingly. That's how we use the methods of a library in our program. We will see further, how we bring in the libraries in our Python programs.

Python standard library

The Python Standard Library contains the exact syntax, semantics, and tokens of Python. It contains built-in modules that provide access to basic system functionality like I/O and some other core modules. Most of the Python Libraries are written in the C programming language. The Python standard library consists of more than 200 core modules. All these work together to make Python a high-level programming language. Python Standard Library plays a very important role. Without it, the programmers can't have access to the functionalities of Python. But other than this, there are several other libraries in Python that make a programmer's life easier. Let's have a look at some of the commonly used libraries:

1.Pandas: Pandas are an important library for data scientists. It is an open-source machine learning library that provides flexible high-level data structures and a variety of analysis tools. It eases data analysis, data manipulation, and cleaning of data. Pandas support operations like Sorting, Re-indexing, Iteration, Concatenation, Conversion of data, Visualizations, Aggregations, etc.

2. Numpy: The name "Numpy" stands for "Numerical Python". It is the commonly used library. It is a popular machine learning library that supports large matrices and multi-dimensional data. It consists of in-built mathematical functions for easy computations. Even libraries like TensorFlow use Numpy internally to perform several operations on tensors. Array Interface is one of the key features of this library.

3. Flask: Flask is a micro web framework written in Python. It is classified as a micro framework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions. However, Flask supports extensions that can add application features as if they were implemented in Flask itself. Extensions exist for object-relational mappers, form validation, upload handling, various open authentication technologies and several common framework related tools.

4. Mysql.connector: mysql.connector is a Python library used to connect to MySQL databases from Python applications. It provides an interface for interacting with MySQL databases by establishing connections, executing queries, managing transactions, and handling errors. mysql.connector is an official MySQL driver for Python, maintained by the MySQL development team.

5. Operating Systems: An operating system (OS) is software that acts as an intermediary between computer hardware and user applications. It manages computer hardware resources, provides services to applications, and facilitates user interaction with the computer. Operating systems come in various types, including general-purpose OS (e.g., Windows, macOS, Linux), embedded OS (e.g., Android, iOS), real-time OS (RTOS), and server OS (e.g., Windows Server, Linux distributions).

Use of Libraries in Python Program

As we write large-size programs in Python, we want to maintain the code’s modularity. For the easy maintenance of the code, we split the code into different parts and we can use that code later ever we need it. In Python, modules play that part. Instead of using the same code in different programs and making the code complex, we define mostly used functions in modules and we can just simply import them in a program wherever there is a requirement. We don’t need to write that code but still, we can use its functionality by importing its module. Multiple interrelated modules are stored in a library. And whenever we need to use a module, we import it from its library. In Python, it’s a very simple job to do due to its easy syntax. We just need to use **import**.

4.5 Hardware Requirements

The hardware requirements include the requirements specification of the physical computer resources for a system to work efficiently. The hardware requirements may serve as the basis for a contract for the implementation of the system and should therefore be a complete and consistent specification of the whole system. The Hardware Requirements are listed below:

System Processor	:	Intel I3
Hard Disk	:	500 GB
Ram	:	4 GB

1. **Processor:** A processor is an integrated electronic circuit that performs the calculations that run a computer. A processor performs arithmetical, logical, input/output (I/O) and other basic instructions that are passed from an operating

system (OS). Most other processes are dependent on the operations of a processor. A minimum 1 GHz processor should be used, although we would recommend S2GHz or more. A processor includes an arithmetical logic and control unit (CU), which measures capability in terms of the following:

- Ability to process instructions at a given time
- Maximum number of bits/instructions
- Relative clock speed

Fig 4.1: Processor

The proposed system requires a 2.4 GHz processor or higher.

2. Ethernet connection (LAN) OR a wireless adapter (Wi-Fi): Wi-Fi is a family of radio technologies that is commonly used for the wireless local area networking (WLAN) of devices which is based around the IEEE 802.11 family of standards. Devices that can use Wi-Fi technologies include desktops and laptops, smartphones and tablets, TV‘s and printers, digital audio players, digital cameras, cars and drones.

Compatible devices can connect to each other over Crop Yield Prediction and Fertilizer Analysis Using Machine Learning Wi- Fi through a wireless access point as well as to connected Ethernet devices and may use it to access the Internet. Such an access point (or hotspot) has a range of about 20 meters (66 feet) indoors and a greater range outdoors. Hotspot coverage can be as small as a single room with walls that block radio waves, or as large as many square 18ilometers achieved by using multiple overlapping access points.

Fig 4.2: Ethernet Connection

3. Hard Disk: A hard disk is an electro-mechanical data storage device that uses magnetic storage to store and retrieve digital information using one or more rigid rapidly rotating disks, commonly known as platters, coated with magnetic material. The platters are paired with magnetic heads, usually arranged on a moving actuator arm, which reads and writes data to the platter surfaces. Data is accessed in a random-access manner, meaning that individual blocks of data can be stored or retrieved in any order and not only sequentially. HDDs are a type often on volatile storage, retaining stored data even when powered off. 32 GB or higher is recommended for the proposed system.

Fig 4.3: Hard Disk

4. Memory (RAM): Random-access memory (RAM) is a form of computer data storage that stores data and machine code currently being used. A random- access memory device allows data items to be read or written in almost the same amount of time irrespective of the physical location of data inside the memory. In today's technology, random-access memory takes the form of integrated chips. RAM is normally associated with volatile types of memory (such as DRAM modules), where stored information is lost if power is removed, although non- volatile RAM has also been developed. A minimum of 4 GB RAM is recommended for the proposed system.

Fig 4.4: RAM

4.6 Software Requirements

The software requirements are description of features and functionalities of the target system. Requirements convey the expectations of users from the software product. The requirements can be obvious or hidden, known or unknown, expected or unexpected from client‘s point of view.

Operating system	:	Windows OS 7+
Coding Language	:	Python
IDE	:	Pycharm IDE
GUI	:	Flask

1.PyCharm: Py Charm is the most popular IDE for Python, and includes great features such as excellent code completion and inspection with advanced debugger and support for web programming and various frameworks. The intelligent code editor provided by PyCharm enables programmers to write high quality Python code. The editor enables programmers to read code easily through colour schemes, insert indents on new lines automatically, pick the appropriate coding style, and avail context-aware code completion suggestions.

At the same time, the programmers can also use the editor to expand a code block to an expression or logical block, avail code snippets, format the code base, identify errors and misspellings, detect duplicate code, and auto-generate code. PyCharm offers some of the best features to its users and developers in the following aspects

- Code completion and inspection
- Advanced debugging
- Support for web programming and frameworks such as Django and Flask

Fig 4.5: Pycharm image

2. Python: It is an object-oriented, high-level programming language with integrated dynamic semantics primarily for web and app development. It is extremely attractive in the field of Rapid Application Development because it offers dynamic typing and dynamic binding options. Python is relatively simple, so it's easy to learn since it requires a unique syntax that focuses on readability. Developers can read and translate Python code much easier than other languages. In turn, this reduces the cost of program maintenance and development because it allows teams to work collaboratively without significant language and experience barriers. Additionally, Python supports the use of modules and a package, which means that programs can be designed in a modular style and code can be reused across a variety of projects.

Fig 4.6: Python Icon

5.Flask Framework: Web Application Framework or simply Web Framework represents a collection of libraries and modules that enables a web application developer to write applications without having to bother about low-level details such as protocols, thread management etc. Flask is a web application framework written in Python. It is developed by Armin Ronacher, who leads an international group of Python enthusiasts named Pocco. Flask is based on the Werkzeug WSGI toolkit and Jinja2 template engine. Both are Pocco projects. Web Server Gateway Interface (WSGI) has been adopted as a standard for Python web application development. WSGI is a specification for a universal interface between the web server and the web applications.

werkzeug is a WSGI toolkit, which implements requests, response objects, and other utility functions. This enables building a web framework on top of it. The Flask framework uses Werkzeug as one of its bases.

Jinja2 is a popular templating engine for Python. A web templating system combines a template with a certain data source to render dynamic web pages.

Flask is often referred to as a micro framework. It aims to keep the core of an application simple yet extensible. Flask does not have built-in abstraction layer for database handling, nor does it have form a validation support. Instead, Flask supports the extensions to add such functionality to the application.

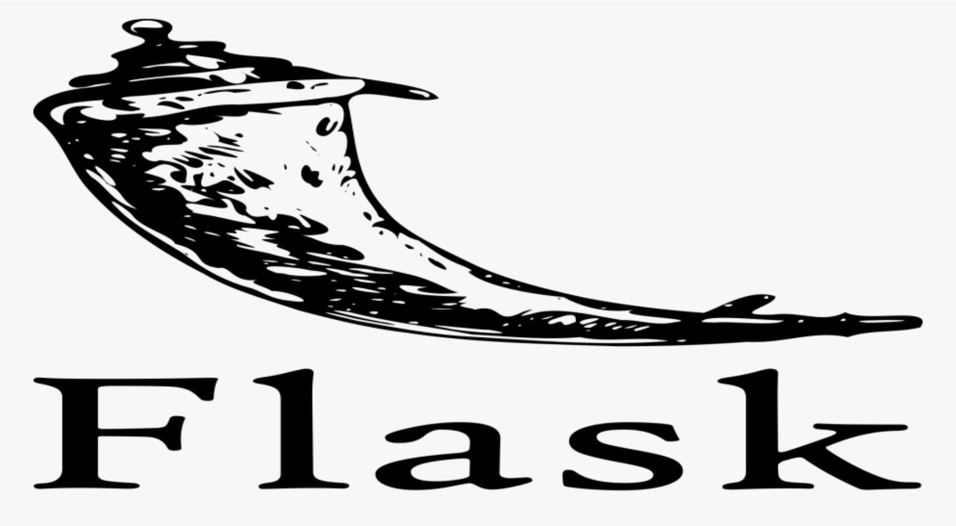


Fig 4.7: Flask Python Logo

CHAPTER 5

SYSTEM ANALYSIS AND DESIGN

Systems development is a systematic process which includes phases such as planning, analysis, design, deployment, and maintenance. System Analysis is a process of collecting and interpreting facts, identifying the problems, and decomposition of a system into its components. System analysis is conducted for the purpose of studying a system or its parts in order to identify its objectives. It is a problem solving technique that improves the system and ensures that all the components of the system work efficiently to accomplish their purpose. Analysis specifies what the system should do.

System Design is a process of planning a new business system or replacing an existing system by defining its components or modules to satisfy the specific requirements. Before planning, you need to understand the old system thoroughly and determine how computers can best be used in order to operate efficiently. System Design focuses on how to accomplish the objective of the system.

5.1 Introduction of Input Design:

In an information system, input is the raw data that is processed to produce output. During the input design, the developers must consider the input devices such as PC, MICR, OMR, etc.

Therefore, the quality of system input determines the quality of system output. Well-designed input forms and screens have following properties −

- It should serve specific purpose effectively such as storing, recording, and retrieving the information.
- It ensures proper completion with accuracy.
- It should be easy to fill and straightforward.
- It should focus on user’s attention, consistency, and simplicity.
- All these objectives are obtained using the knowledge of basic design principles regarding −
    - What are the inputs needed for the system?
    - How end users respond to different elements of forms and screens.

Objectives for Input Design:

The objectives of input design are –

- To design data entry and input procedures
- To reduce input volume
- To design source documents for data capture or devise other data capture methods
- To design input data records, data entry screens, user interface screens, etc.
- To use validation checks and develop effective input controls.

Output Design:

The design of output is the most important task of any system. During output design, developers identify the type of outputs needed, and consider the necessary output controls and prototype report layouts.

Objectives of Output Design:

The objectives of input design are:

- To develop output design that serves the intended purpose and eliminates the production of unwanted output.
- To develop the output design that meets the end user's requirements.
- To deliver the appropriate quantity of output.
- To form the output in appropriate format and direct it to the right person.
- To make the output available on time for making good decisions.

5.2 UML Diagrams:

Use Case Diagram:

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

The goal is for UML to become a regular dialect for making fashions of item arranged PC programming. In its gift frame UML is contained two noteworthy components: a Meta-show and documentation. Later on, a few type of method or system can also likewise be brought to; or related with, UML. The Unified Modeling Language is a popular dialect for indicating, Visualization, Constructing and archiving the curios of programming framework, and for business demonstrating and different non-programming frameworks. The UML speaks to an accumulation of first-rate building practices which have verified fruitful in the showing of full-size and complicated frameworks. The UML is a essential piece of creating gadgets located programming and the product development method. The UML makes use of commonly graphical documentations to specific the plan of programming ventures.

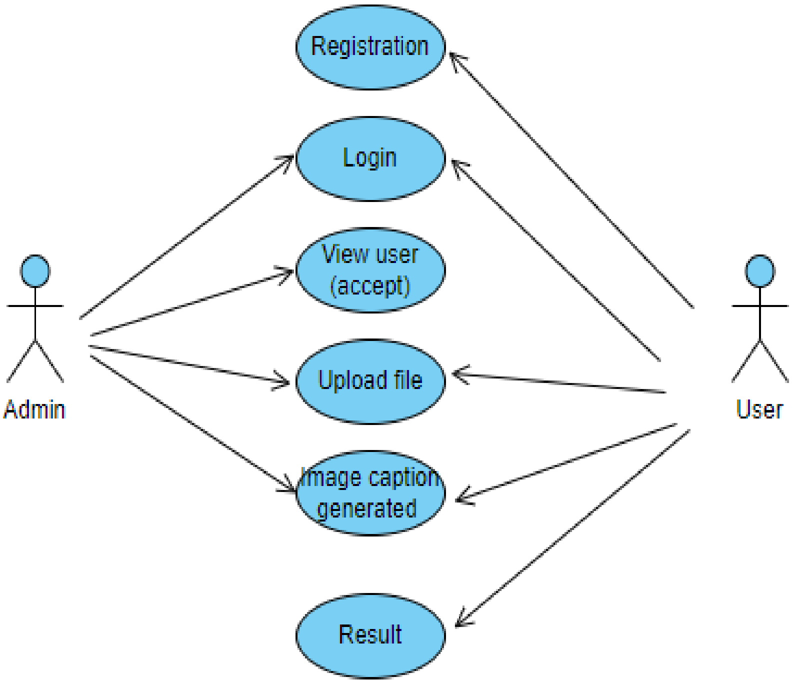


Fig 5.1 Use Case Diagram

CLASS DIAGRAM

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.

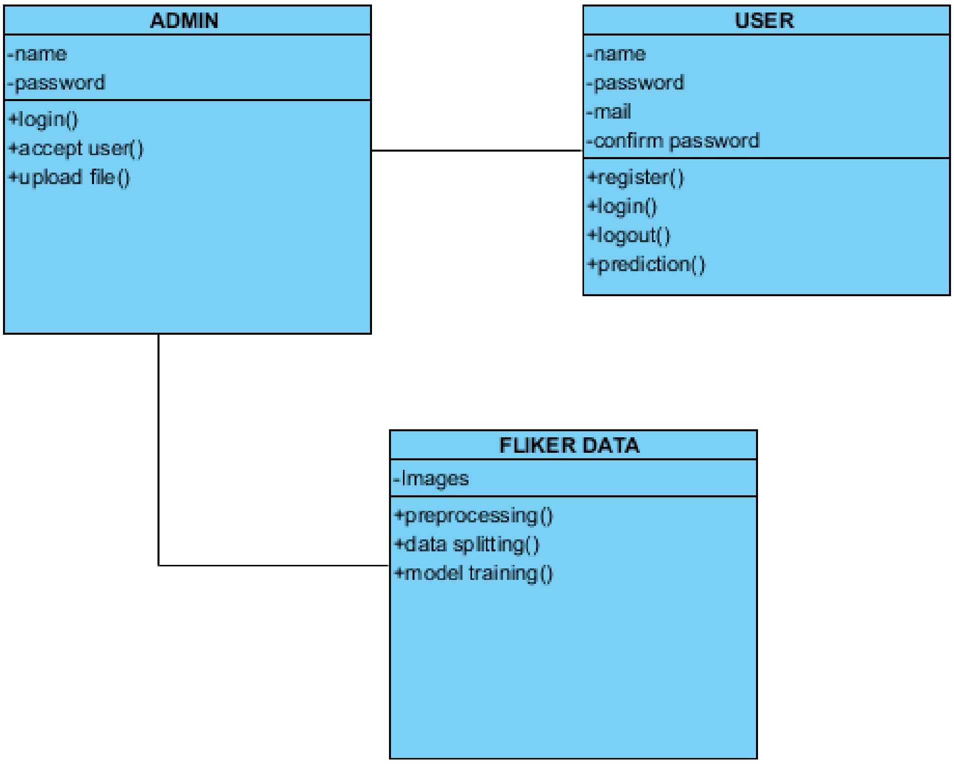


Fig 5.2: Class Diagram

SEQUENCE DIAGRAM:

A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.

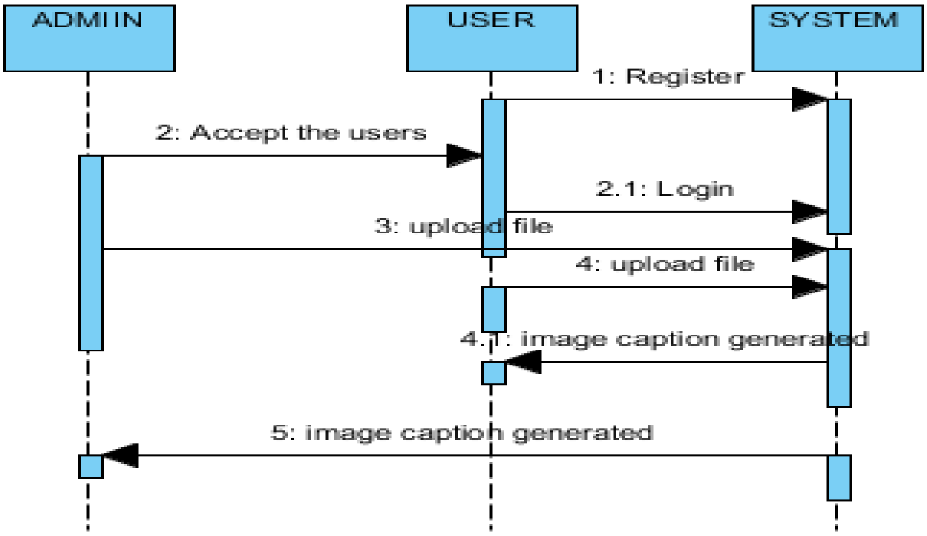


Fig 5.3: Sequence Diagram

COLLABORATION DIAGRAM:

In collaboration diagram the method call sequence is indicated by some numbering technique as shown below. The number indicates how the methods are called one after another. We have taken the same order management system to describe the collaboration diagram. The method calls are similar to that of a sequence diagram. But the difference is that the sequence diagram does not describe the object organization whereas the collaboration diagram shows the object organization.

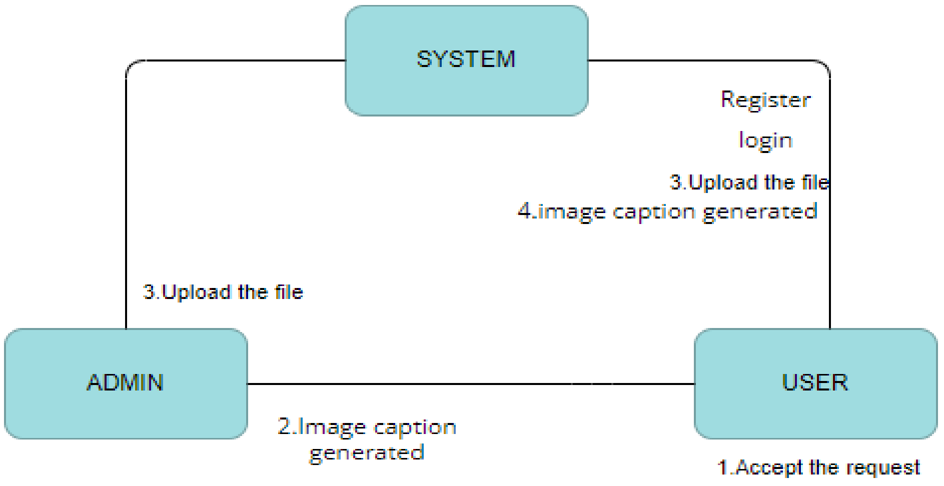


Fig 5.4: Collaboration Diagram

DEPLOYEMENT DIAGRAM:

Deployment diagram represents the deployment view of a system. It is related to the component diagram. Because the components are deployed using the deployment diagrams. A deployment diagram consists of nodes. Nodes are nothing but physical hardware's used to deploy the application.

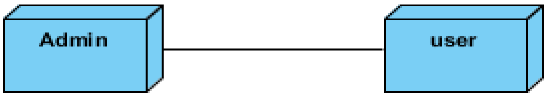


Fig 5.5: Deployement diagram

ACTIVITY DIAGRAM:

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.

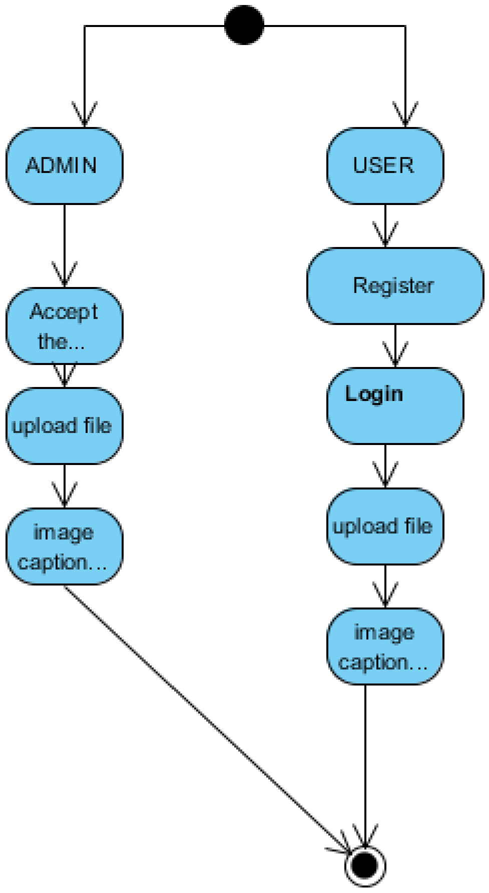


Fig 5.6: Activity Diagram

COMPONENT DIAGRAM:

A component diagram, also known as a UML component diagram, describes the organization and wiring of the physical **c**omponents in a system. Component diagrams are often drawn to help model implementation details and double-check that every aspect of the system's required functions is covered by planned development.

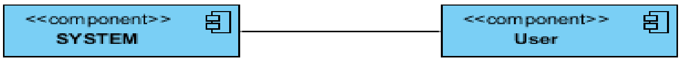


Fig 5.7: Component Diagram

ER DIAGRAM:

An Entity–relationship model (ER model) describes the structure of a database with the help of a diagram, which is known as Entity Relationship Diagram (ER Diagram). An ER model is a design or blueprint of a database that can later be implemented as a database. The main components of E-R model are: entity set and relationship set.

An ER diagram shows the relationship among entity sets. An entity set is a group of similar entities and these entities can have attributes. In terms of DBMS, an entity is a table or attribute of a table in database, so by showing relationship among tables and their attributes, ER diagram shows the complete logical structure of a database. Let's have a look at a simple ER diagram to understand this concept.

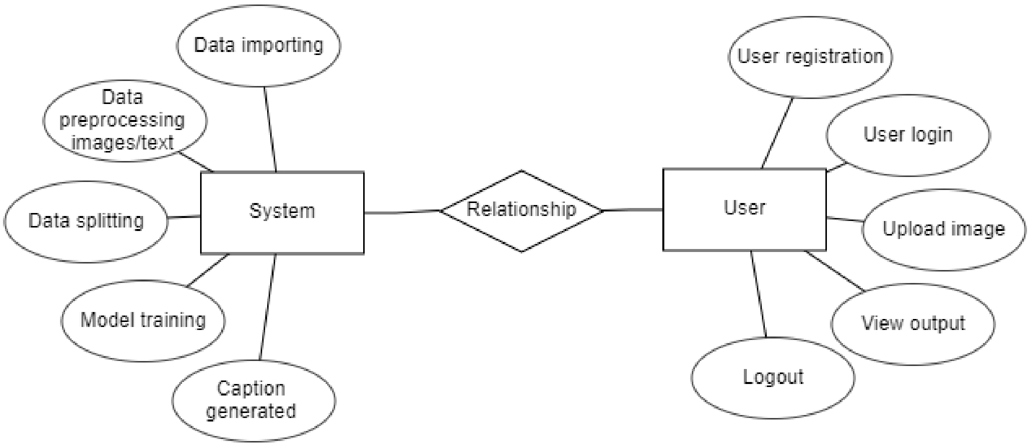


Fig 5.8: ER Diagram

Usage of UML in Project

As the strategic value of software increases for many companies, the industry looks for techniques to automate the production of software and to improve quality and reduce cost and time to the market. These techniques include component technology, visual programming, patterns and frameworks. Additionally, the development for the World Wide Web, while making somethings simpler, has exacerbated these architectural problems. The UML was designed to respond to these needs. Simply, systems design refers to the process of defining the architecture, components, modules, interfaces and data for a system to satisfy specified requirements which can bed one easily through UML diagrams.

5.3 System Architecture

Architecture diagrams can help system designers and developers visualize the high-level, overall structure of their system or application for the purpose of ensuring the system meets their users' needs. They can also be used to describe patterns that are used throughout the design. It's somewhat like a blueprint that can be used as a guide for the convenience of discussing, improving, and following among a team.

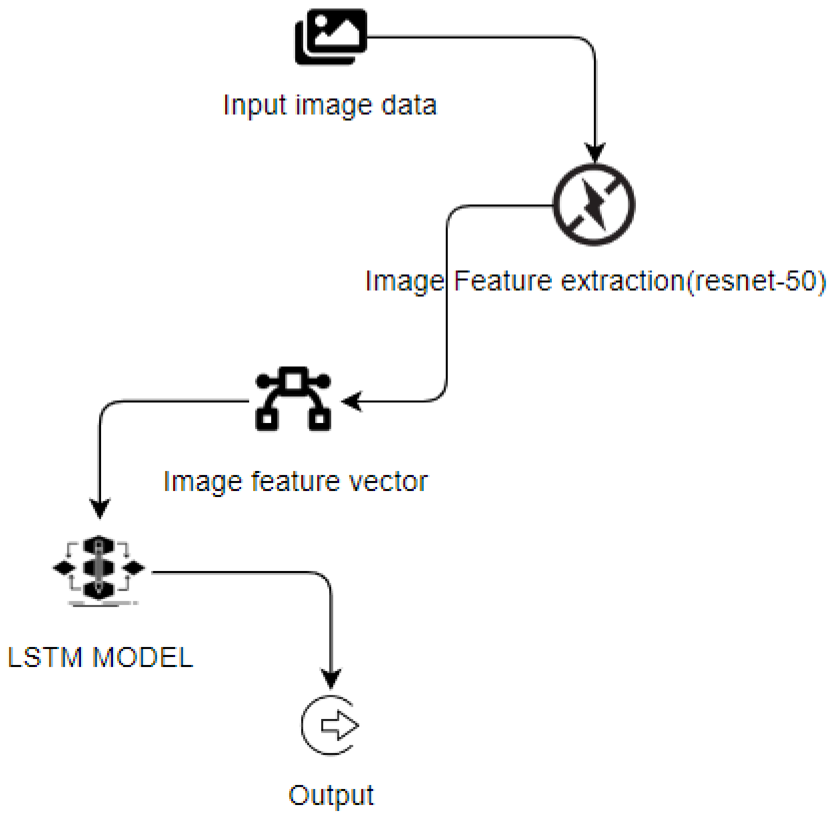


Fig 5.9: System Architecture

5.4 Data Flowchart Diagram (DFD):

A Data Flow Diagram (DFD) is a traditional way to visualize the information flows within a system. A neat and clear DFD can depict a good amount of the system requirements graphically. It can be manual, automated, or a combination of both. It shows how information enters and leaves the system, what changes the information and where information is stored. The purpose of a DFD is to show the scope and boundaries of a system as a whole. It may be used as a communications tool between a systems analyst and any person who plays a part in the system that acts as the starting point for redesigning a system.

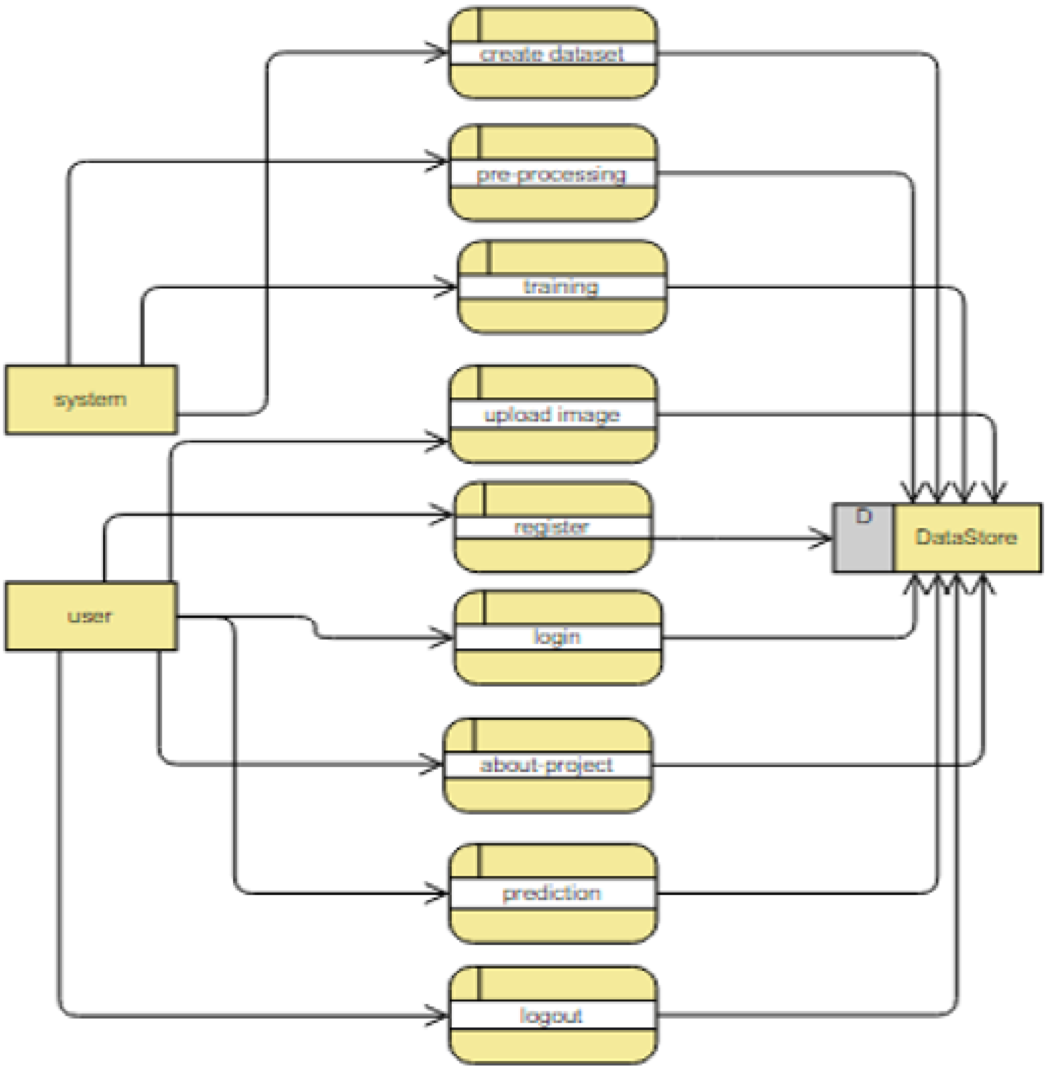


Fig 5.10: Data Flow Diagram

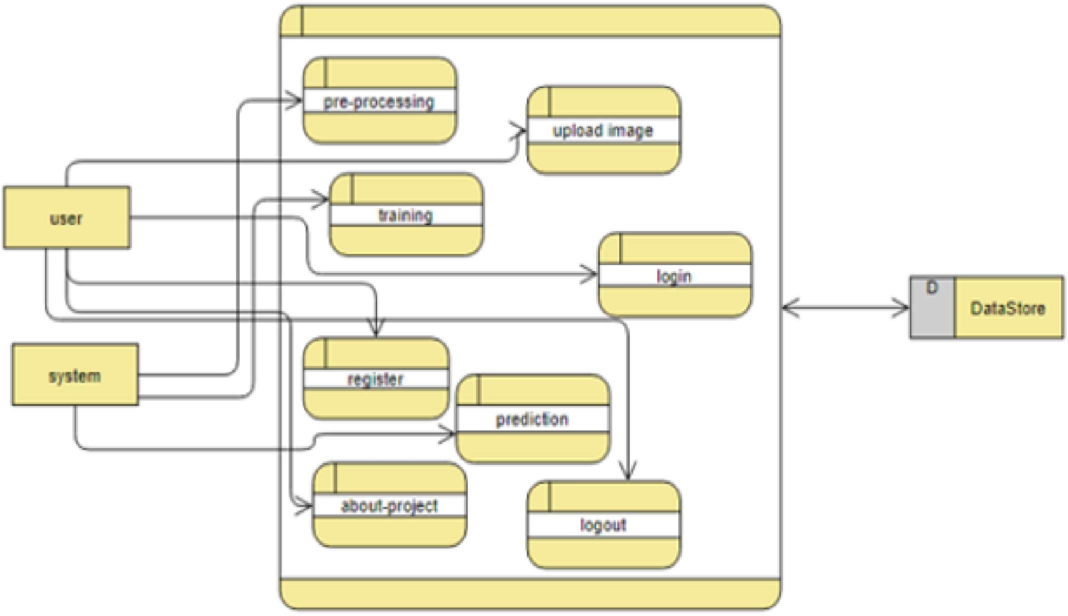


Fig 5.11: Data Flow Diagram(2)

CHAPTER 6

IMPLEMENTATION

The two modules that make up the proposed system are the admin login, user registration and login, and user application.

1.System:

1.1 Create Dataset:

The dataset containing images and text data of the desired objects to be captioned is split into training and testing dataset with the test size of 20-30%.

1.2 Pre-processing:

Resizing and reshaping the images into appropriate format to train our model.

1.3Training:

Use the pre-processed training dataset is used to train our model using RESNET-50 and LSTM algorithm

2.User:

2.1 Register

The user needs to register and the data stored database**.**

2.2 Admin login

Admin logs in into the administrator login and views the user registered list, once he accepts the user data only then the user will be allowed to login.

2.3 Login

A registered user can login using the valid credentials to the website to use a application.

2.4 About-Project

In this application, we have successfully created an application which takes to classify the images.

2.5 Upload Image

The user has to upload an image which needs to be Captioned the images.

2.6 Prediction

The results of our model will display the caption of image we have assigned to it.

2.7 Logout

Once the prediction is over, the user can logout of the application.

Working Flow of the System

i). Each new user must first register for retrieve captions by providing their information. Registration is therefore the first thing we do.

ii). Before choose an image from the dataset the registered user must be verified by the admin like email verification can be held to verify details and admin allow the user for the further proceedings. After authenticating successfully, system allow user. Now the registration of the user is successfully completed.

iii). At the time of registration, we will use two levels of security first one is user email id and second one is password verification.

iv). System will check username and password entered by the user is correct or not.

v). If username and password is correct then system allow to choose an image from the dataset to retrieve the caption and the audio for that particular image.

vi). If the user name and password will not match it will display a message like the login credentials are incorrect please check.

vii). On the users page, It will display all the users list which is connected with the mysql connector.

viii). The user is allowed to this website for multiple times with the same login credentials.

6.1 Datasets

Flickr30k is a popular dataset widely used in computer vision and natural language processing (NLP) research, specifically for tasks such as image captioning and image-text alignment.

The dataset contains around 31,000 images, each paired with five descriptive captions, resulting in approximately 150,000 caption-image pairs. This substantial size makes it suitable for training and evaluating models for various image-related tasks.

Machine learning techniques are used to train the dataset images.

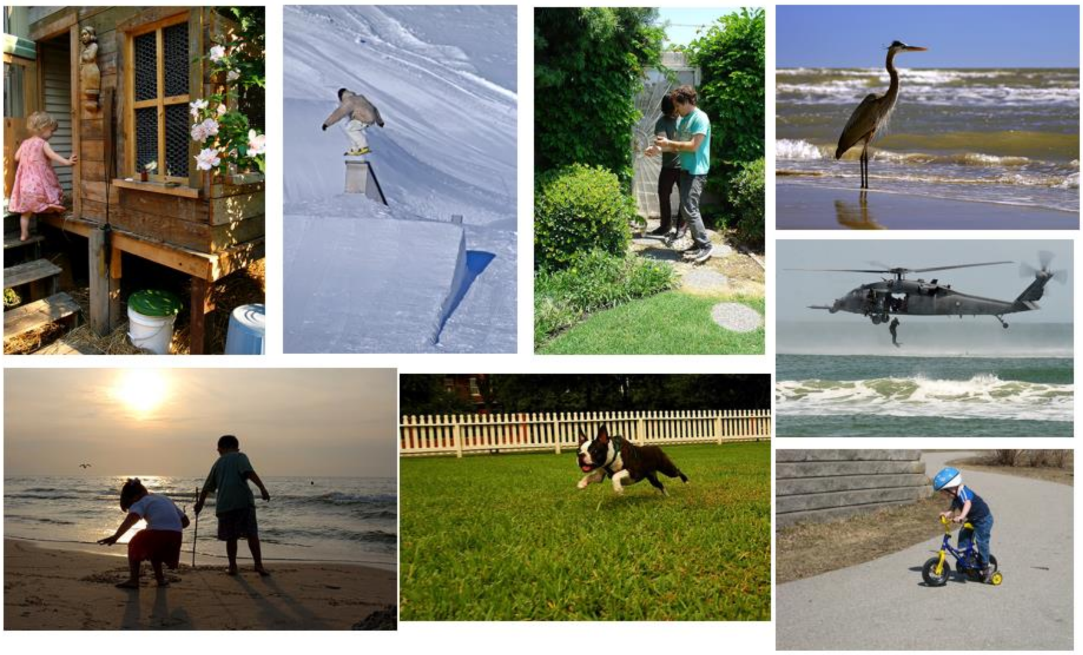


Fig 6.1: Dataset Collection

6.2 Data Pre-Processing

Data Pre-Processing is a Data Mining method that entails converting raw data into a format that can be understood. Real-world data is frequently inadequate, inconsistent, and/or lacking in specific activities or trends, as well as including numerous inaccuracies. This might result in low-quality data collection and, as a result, low-quality models based on that data. Preprocessing data is a method of resolving such problems. Machines do not comprehend free text, image, or video data; instead, they comprehend 1s and 0s. So putting on a slideshow of all our photographs and expecting our machine learning model to learn from it is probably not going to be adequate. Data Pre-processing is the step in any Machine Learning process in which the data is changed, or encoded, to make it easier for the machine to parse it. In other words, the algorithm can now easily interpret the data's features. Data Pre-processing can be done in four different ways. Data cleaning/cleaning, data integration, data transformation, and data reduction are the four categories.

6.2.1 Data Cleaning:

Data in the real world is frequently incomplete, noisy, and inconsistent. Many bits of the data may be irrelevant or missing. Data cleaning is carried out to handle this aspect. Data cleaning methods aim to fill in missing values, smooth out noise while identifying outliers, and fix data discrepancies. Unclean data can confuse data and the model. Therefore, running the data through various Data Cleaning/Cleansing methods is an important Data Pre-processing step.

6.2.2 Data Integration:

It is involved in a data analysis task that combines data from multiple sources into a coherent data store. These sources may include multiple databases. Do you think how data can be matched up?? For a data analyst in one database, he finds Customer_ID and in another he finds cust_id, how can he sure about them and say these two belong to the same entity.

6.2.3 Data Reduction:

Because data mining is a methodology for dealing with large amounts of data. When dealing with large amounts of data, analysis becomes more difficult. We employ a data reduction technique to get rid of this. Its goal is to improve storage efficiency while lowering data storage and analysis expenses.

Dimensionality Reduction

A huge number of features may be found in most real-world datasets. Consider an image processing problem: there could be hundreds of features, also known as dimensions, to deal with. As the name suggests, dimensionality reduction seeks to minimize the number of features but not just by selecting a sample of features from the feature set, which is something else entirely Feature Subset Selection or feature selection.

CHAPTER 7

TESTING

Software testing is an investigation conducted to provide stakeholders with information about the quality of the software product or service under test. Software testing can also provide an objective, independent view of the software to allow the business to appreciate and understand the risks of software implementation. Test techniques include the process of executing a program or application with the intent of finding software bugs (errors or other defects), and verifying that the software product is fit for use.

Software testing involves the execution of a software component or system component to evaluate one or more properties of interest. In general, these properties indicate the extent to which the component or system under test:

- Meets the requirements that guided its design and development,
- Responds correctly to all kinds of inputs,
- Performs its functions within an acceptable time,
- It is sufficiently usable,
- Can be installed and run in its intended environments, and
- Achieves the general result its stakeholder's desire.

7.1 Feasibility Study

The feasibility of the project is analysed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

- ♦ Economical feasibility
- ♦ Technical feasibility
- ♦ Social feasibility

Economical Feasibility

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

Technical Feasibility

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

Social Feasibility

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

System Testing

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub-assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the

Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

7.2 Types of Testing:

7.3 Functionality Testing

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

Valid Input : identified classes of valid input must be accepted.

Invalid Input : identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.

Output : identified classes of application outputs must be exercised.

Systems/Procedures: interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

7.4 Usability Testing

- The application enables smooth navigation, hence gives a user-friendly experience.
- The inputs taken from the user are via dropdown hence correct inputs are provided to the system.
- Wrong inputs given by the system are handled effectively.
- The content provided by the application is verified and is taken by the trusted sources.
- The datasets trained for prediction of the crop yield are accurate and balanced.

7.5 Interface Testing

- The application connects correctly with the server. In case of failure an appropriate message is displayed.

- Interruptions by the server or by the user are handled efficiently.
- If the user enters wrong credentials or invalid email id, the application handles it efficiently by displaying appropriate messages.
- The interaction with the user is smooth and easy.

7.6 Performance Testing

- It works fine with moderate internet speed.
- The connection is secured and user details are stored in a secured manner.
- The switch from one screen to another is quick and smooth.
- The inputs from users are taken correctly and response is recorded quickly.

7.7 Unit Testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

7.8 Integration Testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components. Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects. The task of the integration test is to check that components or software applications, e.g. components in a software system or – one step up – software applications at the company level – interact without error.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

Acceptance Testing:

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

7.9 System Testing

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

7.10 White Box Testing

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level.

7.11 Black Box Testing:

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot “see” into it. The test provides inputs and responds to outputs without considering how the software works.

Test objectives

- All field entries must work properly.
- Pages must be activated from the identified link.

- The entry screen, messages and responses must not be delayed.

Features to be tested

- Verify that the entries are of the correct format
- No duplicate entries should be allowed

All links should take the user to the correct page.

CHAPTER 8

RESULTS

In the final implementation of the application the first screen the user can view is the Home page.

Home Page: It contain functionalities User Registration, User Login.

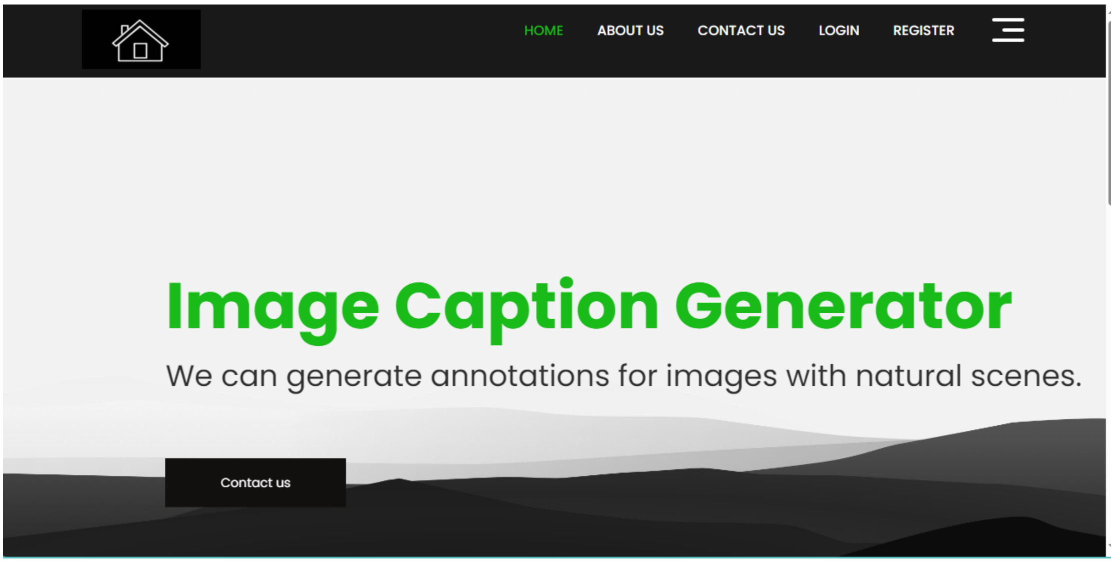


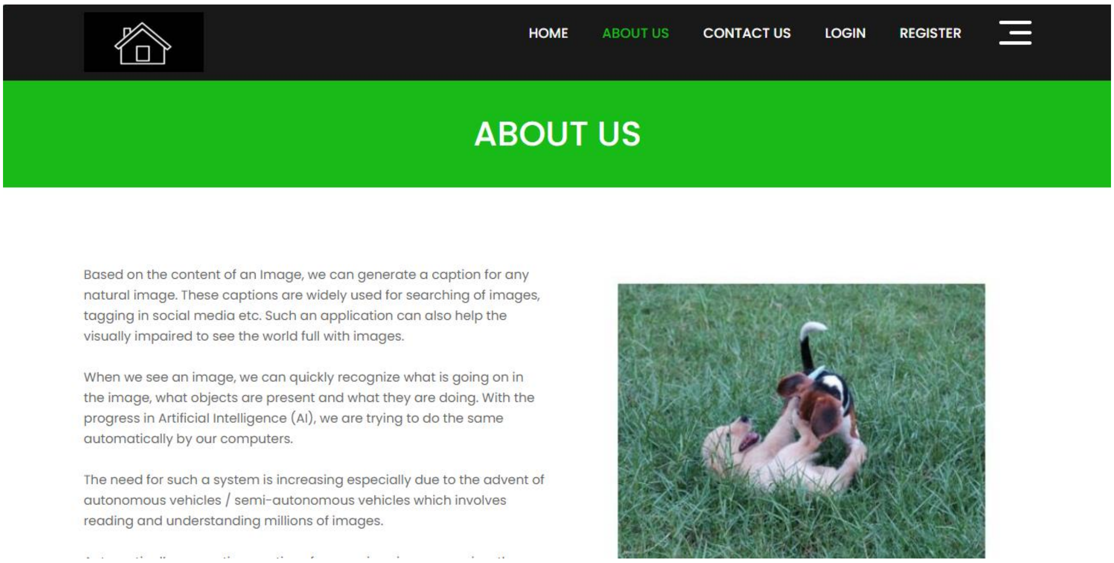
Image Caption Generator

We can generate annotations for images with natural scenes.

Fig 8.1: Home Page

The above figure represents home page where we land after clicking on the link.

About Page: here we have a slight description about the project



ABOUT US

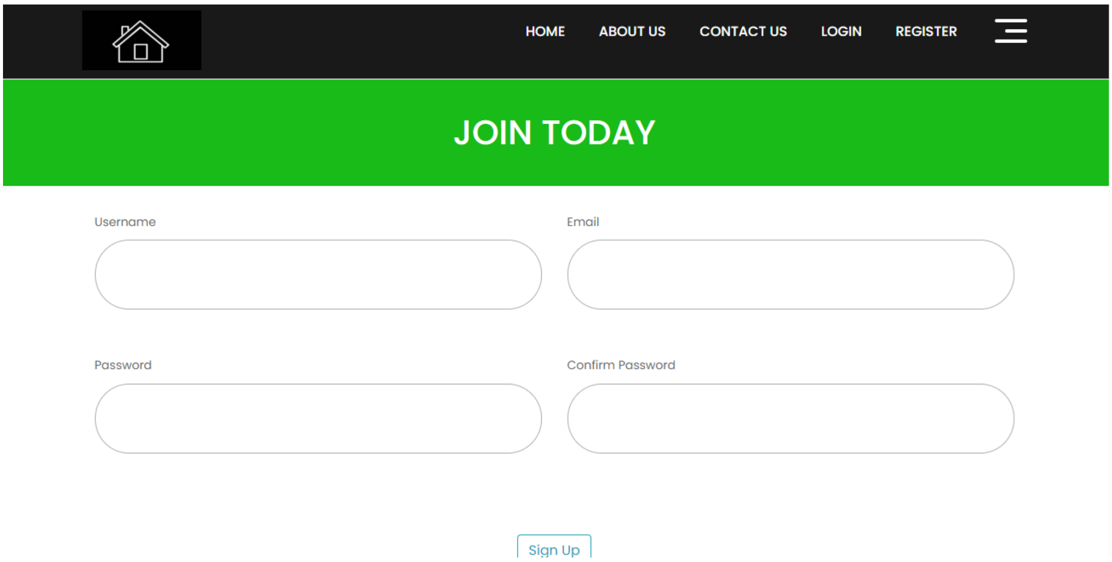
Based on the content of an Image, we can generate a caption for any natural image. These captions are widely used for searching of images, tagging in social media etc. Such an application can also help the visually impaired to see the world full with images.

When we see an image, we can quickly recognize what is going on in the image, what objects are present and what they are doing. With the progress in Artificial Intelligence (AI), we are trying to do the same automatically by our computers.

The need for such a system is increasing especially due to the advent of autonomous vehicles / semi-autonomous vehicles which involves reading and understanding millions of images.

Fig 8.2 : About Page

Register page: Here User registers himself/herself



Username

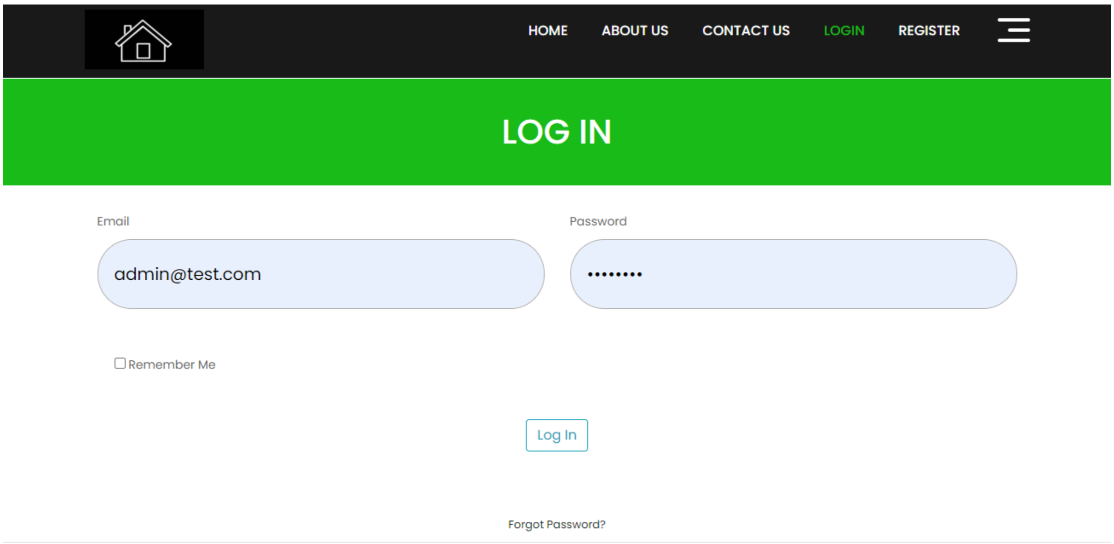
Email

Password

Confirm Password

Fig 8.3: Register Page

Login Page: Here user logs in with the credentials they registered with



Email: admin@test.com

Password: ••••••••

□ Remember Me

Fig 8.4: Login Page

Upload page: Here user uploads the image and caption is generated

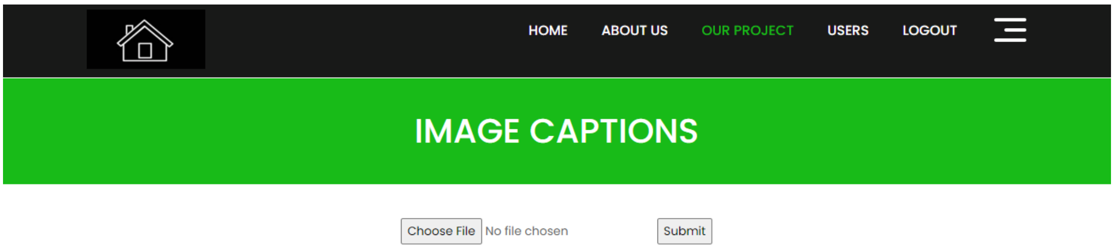
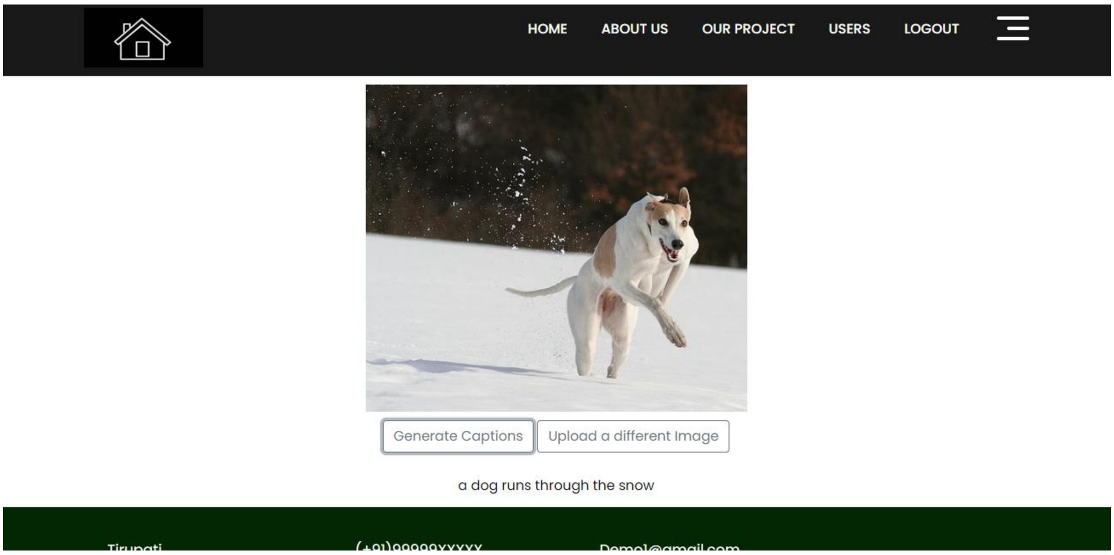


Fig 8.5: Upload Page

Result: Here we get the results classified



a dog runs through the snow

Fig 8.6: Results

CONCLUSION

The issue of creating meaningful captions for images has been found to be powerfully and effectively solved by the image caption generator that combines Long Short-Term Memory (LSTM) networks with Convolutional Neural Networks (CNNs). Using CNN layers to obtain relevant characteristics and capture spatial information, the CNN-LSTM model showed how to generate contextually relevant captions by efficiently utilizing LSTM layers. Visual perception and sequential data processing work together to address the problems of picture understanding and natural language synthesis through the integration of these two architectures. This work emphasizes the need of merging all the layers to produce better results from challenging tasks like image captioning.

Future research can build upon and enhance the CNN and LSTM picture caption generator in a number of ways. First, by investigating more complex architectures like attention processes, transformer models, or language models that have already undergone training, like BERT, the model may be able to better capture the complex interactions between textual and visual data.

Moreover, enlarging and diversifying the dataset inside the training set might enhance the generalization of the model, enabling it to precisely depict a greater range of images. Optimizing the model for certain domains or tasks could also be useful, allowing the generator to focus on areas such as medical imaging or satellite photography. Furthermore, investigating ways to make the model more interpretable and controllable could aid in enhancing understanding and providing direction for the captioning process. Finally, but just as importantly, putting the model to use in real situations and gathering user feedback would reveal its practicality and highlight areas for development.

REFERENCES

[1] M Sailaja, K Harika, B Sridhar. Rajan Singh, "Image Caption Generator using Deep Learning", 2022 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC).

[2] C S Kanimozhiselvi, Karthika V, Kalavani S P, Krithika S, "Image Captioning Using Deep Learning", 2022 International Conference on Computer Communication and Informatics (ICCCI).

[3] Chetan Amritkar, Vaishali Jabade, "Image Caption Generation Using Deep Learning Technique", 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA).

[4] Varsha Kesavan, Vaidehi Muley, Megha Kolhekar, "Deep Learning based Automatic Image Caption Generation", 2019 Global Conference for Advancement in Technology (GCAT).

[5] Lakshmi narasimhan Srinivasan, Dinesh Sreekanthan and A.L Amutha, "Image captioning - A Deep Learning Approach", International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 9 (2018) pp.

[6] T. J. Buschman and E. K. Miller. "Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices. Science", 315(5820):1860–1862, 2007.

[7] William Fedus, Ian Goodfellow, Andrew M Dai. Maskgan, "Better text generation", 1801.07736, 47, 2018.

[8] D. Elliott, F. Keller, "Image Description using Visual Dependency Representations", Conference on Empirical Methods in Natural Language Processing.

[9] N K Kumar, D Vigneswari, A Mohan, K Laxman, J Yuvaraj, "Detection and Recognition of Objects in Image Caption Generator System: A Deep Learning Approach", IEEE – 2019..

[10]C. Amritkar, V. Jabade, "Image Caption Generation using Deep Learning Technique", IEEE Access, 2018.

[11] Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. "Baby talk: Understanding and generating image descriptions", IEEE Transactions on Pattern Analysis and Machine Intelligence, 35:2891–2903, June 2013.

[12] "Bottom-up and top-down attention for image captioning and visual question answering" by Anderson, Peter, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. (2018)

[13] "Meshed-memory transformer for image captioning" by Cornia, Marcella, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara (2020).

[14] "Unified vision-language pre-training for image captioning" and vqa by Zhou, Luowei, et al (2020).

[15] Xiaodan Liang, Zhiting Hu, Hao Zhang, Chuang Ggan ,and Eric P Xing.2017. "Recurrent Topic-Transition for Visual Paragraph Generation."

Generating Image Captions Based on Deep Learning and Natural Language Processing

Rohini. P,[1, a)] Bhargavi M, Jasmin G, Manjusha P, Anulekha Sai A[2,3,4,5 b)]

Affiliations Author

[1]Associate Professor Srinivasa Ramanujan Institute of Technology Anantapur, Andhra Pradesh, 515001, India

[2,3,4,5]Srinivasa Ramanujan Institute of Technology Anantapur, Andhra Pradesh, 515001, India

Author Emails

[1)]rohini.cse@srit.ac.in,
[2)]204g1a0521@srit.ac.in
[3)]204g1a0542@srit.ac.in
[4)]204g1a0552@srit.ac.in
[5)]214g5a0504@srit.ac.in

Abstract: Humans and computers are attempting to communicate because everything in today's society depends on systems like computers, mobile phones, etc. This is how our project is visualized. Our undertaking People with visual impairments can benefit from the creation of image captions. Computers are unable to distinguish objects, things, or activities with the same ease as humans. To recognize them, they require some training. The suggested method is used to identify activities or similar items. We offer several deep neural network-based models for creating captions for images, with a particular emphasis on CNNs (Convolutional Neural Networks) that extract characteristics from the image. Using LSTM (Long Short-Term Memory) techniques, RNNs (Recurrent Neural Networks) create captions based on the image's attributes. and examining how they affect the construction of sentences. Here, encoder-decoders are used to create a link between descriptions from natural language processing and visual information such as image features. The process of generating a caption's sequence is handled by the decoder, while the encoder extracts features. In order to determine which feature extraction and encoder model produces the best results and accuracy, we have also created captions for sample photos and compared them with one another. We also introduce Deep Voice, a text-to-speech system of production quality that uses only deep neural networks to generate captions based on visual attributes. The evaluation of our project will be conducted utilizing several machine learning methods and Python.

Keywords: CNN, RNN, LSTM, Encoder - Decoder.

INTRODUCTION

It is relatively easy for humans to describe the environments they are living. It is normal for a human to be able to quickly describe a vast amount of information about an image. This is a fundamental human ability. The ability to identify objects and describe images is facilitated by the human brain. Artificial Intelligence introduces numerous algorithms that are based on the architecture of the brain. Here, human beings are employing these algorithms to mimic human visual world interpretation on computers. Despite significant advancements in computer vision fields including object identification, picture classification, attribute classification, and scene recognition. To allow a computer the automatically explains an image, which has been forwarded to it using a language that resembles a person, is a relatively new undertaking. The image captioning is a process that automatically produces a description for the given image using a computer. It is a difficult task. Image captioning necessitates both a high-level comprehension of an image's semantic contents and the ability to convey the information in a sentence that sounds human because it connects computer vision with natural language processing. Giving computers access to the world visually, this will bring a wide range of applications, including the creation of human-robot interactions, early childhood education, information retrieval, and support for those who are visually impaired. It is a significantly a new endeavor to make a computer automatically explain an image that is forwarded to it in a language that sounds human.

The mechanism of automatically generating a natural language caption for a picture with a computer known as image captioning. It's a challenging task. Because picture captioning combines computer vision and natural language processing, that requires both a high level of knowledge of an image's semantic contents and the ability to transmit the information in a human-sounding sentence. Full of reasonable applications, will become possible if computers are given access to the visual world.

Convolutional neural networks (CNN) algorithm is mainly used to extracting features from the picture, while recurrent neural networks (RNN) are used to generate sequence of words in a meaningful way which may easily understandable to human beings based on the image. During first stage, we have taken a novel technique to extracting features from a picture, which will provide us with details on even the smallest change between two comparable photos, instead of just detecting the objects present in the image. The 16 convolutional layer model VGG-16 (Visual Geometry Group) has been employed by us for object recognition.

Additionally, the IoT framework introduces an analytical layer, endowing the central unit with the ability to process incoming data and offer insights into trends, patterns, and potential issues. This analytical capability empowers users to make informed decisions and take preventive actions to maintain an ideal indoor plant environment. The incorporation of IoT elevates the indoor plant monitoring system by enabling real-time data exchange, remote management, and automated control. This interconnected ecosystem enhances efficiency, promoting optimized care for indoor plants and ensuring their sustained development and overall health.

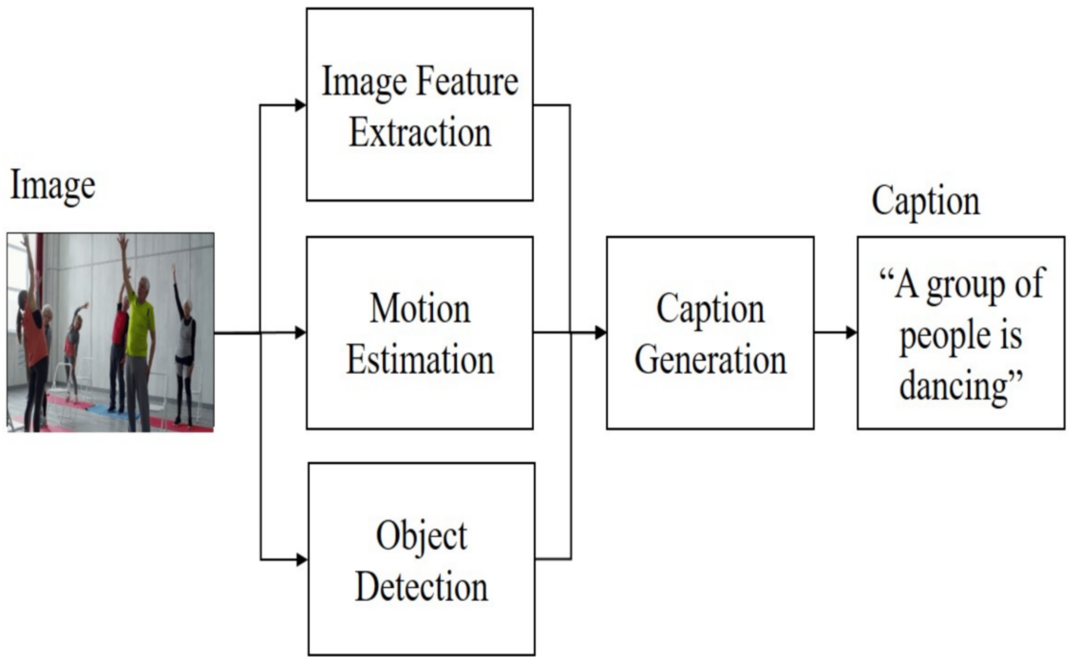


Fig1.: Block Diagram

LITERATURE SURVEY

The low accuracy of the current models is what led to the creation of this research. However, since everything is digital and has displays these days, this concept will have practical implications in the modern world. Numerous investigators devised multiple methods to guarantee precision. However, some of them fall short, while some succeed to the best of their abilities.

K Harika, Rajan Singh, M Sailaja, B Sridhar. [1] In recent years, deep neural networks have enabled the captioning of images. Based on the dataset, the photo caption generator assigns a suitable title to an applied input image. The current study suggests a deep learning-based model and applies it to produce a caption for the input image. The model frames a statement associated with an input image by utilizing CNN and LSTM algorithms. This CNN model recognizes the objects in the picture, and the Long Short-Term Memory (LSTM) algorithm generates text as well as caption that fits the project. Thus, the primary goals of the suggested model are object recognition and title generation for the input images.

Kalaivani S P, Karthika S, Karthika V, C S Kanimozhiselvi. [2] Writing a description for an image is known as image captioning. It is among the most important and contemporary research problems at the moment. Every day, new methods for solving the problem are developed. Even with so many possibilities available, more work needs to be done in order to get more accurate and superior results. We thus considered developing an image captioning model that combines multiple Long Short-Term Memory and Convolutional Neural Network architecture configurations in order to attain superior results. Three CNN and LSTM combinations were combined to create the model. The proposed model is trained using three Convolutional Neural Network architectures: Xception, ResNet50, and Inception-v3. These networks are used to use LSTM to extract features from the image and generate pertinent captions. The three CNN and LSTM combinations that perform the best are selected based on the accuracy of the model. The model is trained on Flicker8k dataset.

Vaishali Jabade, Amritkar Chetan. [3] Artificial Intelligence automatically synthesizes the information of images using machine vision and natural language processing (NLP). A neuronal regeneration model is developed. Both machine translation and vision are involved. This process results in the creation of organic sentences that ultimately describe the visual. Recurrent neural networks (RNN) and convolutional neural networks (CNN) make up this paradigm. RNNs are used to construct sentences, whereas CNN is used to extract features from images. When given an input image, the model is trained to provide captions that, in essence, explain it. Various datasets are used to assess the language proficiency and accuracy of the model learned from picture descriptions. The results of these tests demonstrate that the model often provides precise descriptions of the input image.

Vaidehi Muley[1], Megha Kolhekar[2], Varsha Kesavan[3] [4] The project focus on generating captions by utilizing content of the image as a source. At the moment, human annotation is required for photos, which makes the process almost impossible for commercial datasets. Using the picture database, the Convolutional Neural Network (CNN) encoder extracts features and subtleties from the image to produce a "thought vector." The objects and features in the picture are then translated by an RNN (Recurrent Neural Network) decoder to create a coherent and sequential caption for an input. In this study, we thoroughly analyze multiple deep neural network-based approaches for producing photo captions and pretrained models, with the goal of determining the most efficient model with fine-tuning. To maximize the model's caption-generating capabilities, they compared models that included and did not include the "attention" concept. To provide a more meaningful comparison, the same dataset is used to train each model.

Lakshmi Narasimhan Srinivasan, Dinesh Sreekanthan, A L Amutha. [5] The keras framework's TensorFlow backend has been utilized in this study's model evaluation. Utilizing assessment measures that were appropriate for the problem's nature allowed for an understanding of how The model has made correct predictions. This paper presents the results of mathematical computations performed on the confusion matrix.

E K Miller, T J Buschman. [6] In the human visual system, bottom-up signals linked to unexpected, unusual can automatically focus attention, as can top-down signals dictated by the present task. In this work, they use comparable language to designate as “top-down” attention mechanisms those that are handle by nonvisual or task-specific context, and a “bottom-up” attention mechanisms that are solely visual feed-forward. The majority of traditional visual attention processes utilized in VQA and picture captioning are top-down in nature. These methods, which are frequently trained to focus on the output of a convolutional neural network (CNN) layer or layers, use an image's caption or a query about it as their context. Nevertheless, this method pays minimal attention to the process of selecting the visual parts that require attention. It is a time taking process to pre-process the image

Farhadi et al. [7] The three primary categories used in this paper are picture captioning techniques are covered

in this section i.e., template-based image captioning, novel caption creation and retrieval-based image captioning. In these approaches, captions are generated using preset templates that contain blank spaces. These systems fill in the blanks in the templates after first identifying the various objects, actions, and characteristics. To generate image captions, for instance, complete the template slots with three distinct scene pieces.

D Elliott, F Keller [8]. The main difficulties in this research include identifying the objects in an image and their characteristics, which are challenging computer vision problems, as well as figuring out how the objects interact and what relationships exist between them. Automatic image description is not without its difficulties. To improve the model's performance, the authors trained it over several layers (or levels) using CNN.

D Vigneswari, N K Kumar, K Laxman, A Mohan, J Yuvraj [9]. This work uses deep learning to discover, recognize, and produce meaningful captions for a given image. For object identification, recognition, and caption generation, Regional Object Detector (ROD) is utilized. The proposed method makes use of deep learning to further improve the existing methods. Python is utilized to conduct experiments on the Flickr 8k dataset in order to illustrate the suggested approach. We are using the best and large dataset i.e., Flickr 30k.

OBJECTIVE

The primary objective is develop a deep learning-based image captioning model that surpasses the current state-of-the-art performance in terms of captioning accuracy through train the dataset model efficiently. The project is designed to convert the generated image caption into audio by using machine learning techniques.

PROPOSED SYSTEM

The project's suggested system is an advanced and adaptable picture captioning solution that uses deep learning techniques in conjunction with natural language processing (NLP) to generate accurate, linguistically coherent, and contextually appropriate captions for photos that have audio accompanying them. CNN is given an image to evaluate and create a feature vector from. This feature vector enhances the user's perception of the image by providing an auditory context that corresponds with the visual content. It is used as input for sigmoid functions and RELU functions in the GRU and LSTM. It also provides image descriptions and is associated with pertinent sounds for the image.

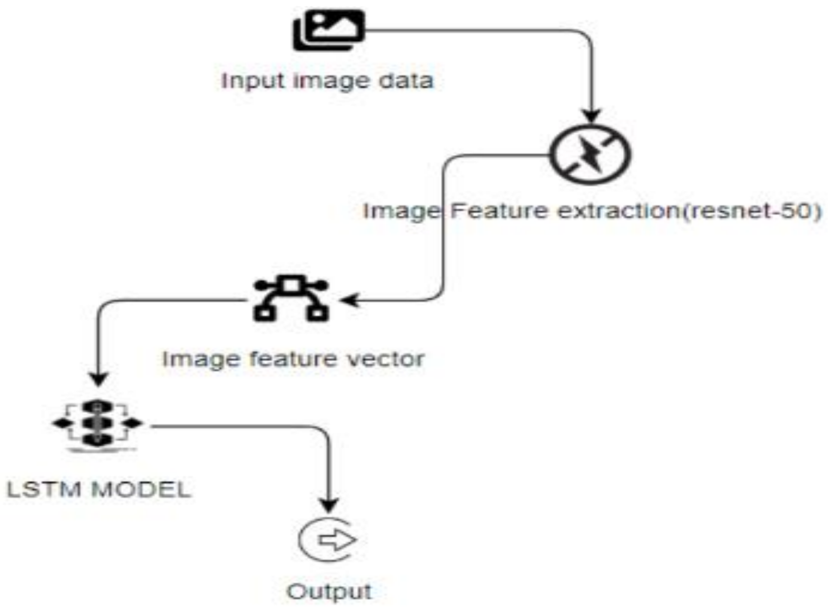


Fig2.: Architecture

METHODOLOGY

Create Dataset: The assess the performance of the model, the target item dataset which consists of both text and picture
data is divided into training and testing datasets.

***Pre-processing**:* Prior to being input into a machine learning model, images must be enhanced and prepared. To train our model, we resize and reshape the photos into the proper format

User

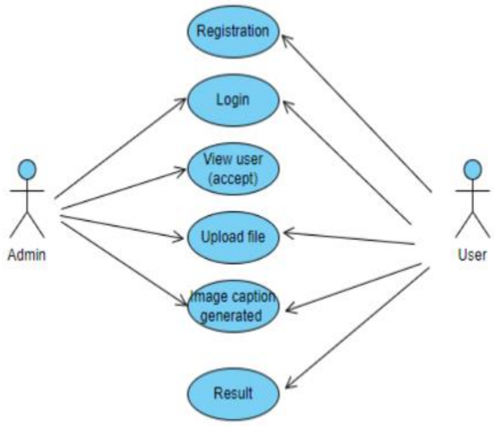


Fig3.: Use Case Diagram

Register: The user must register. The information is kept in a database, and the administrator will review it and grant access if the administrator determines that the registered information is accurate.

Admin Login: The administrator logs in and examines the list of users who have registered. Only the user will be able to log in when he has approved the user data. The administrator has complete control over who may read, edit and manage the data.

Login: Users can use this technique to authenticate themselves with the system and access the application by entering their credentials, which consist of their username and password.

Upload image*:* The user must choose an image from the dataset and upload it into the program. The image must have a caption

Prediction: It produces a caption as an output, and the image caption that we have given to it will be displayed as a result of our model's operation.

Logout: The user has the option to exit the application when the result has been generated**.**

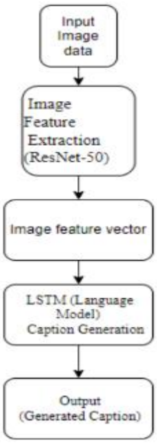

Input Image data
Image Feature Extraction (ResNet-50)
Image feature vector
LSTM (Language Model) Caption Generation
Output (Generated Caption)

Fig4.: Use Case Diagram

IMPLEMENTATION

Over the past 20 years, Deep Learning has shown to be a very powerful technology due to its ability to handle large amounts of data. Convolutional neural networks, also referred to as CNNs or ConvNets, are the most popular deep neural networks in deep learning, especially for applications in computer vision. The type of deep neural network utilized in deep learning that is most commonly used to evaluate visual data is called a convolutional neural network (CNN). It uses a unique method known as convolution. As it currently understood, When two functions are mathematically operated upon, a third function known as convolution is produced that describes how one's shape is transformed by another.

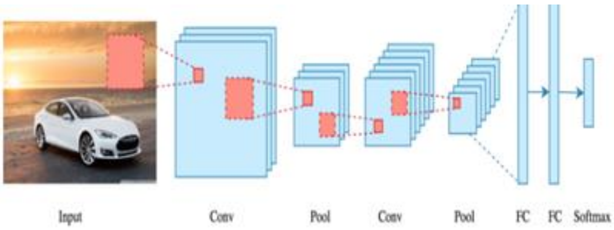


Fig5.: Working of CNN

Convolutional, pooling, and fully-connected (FC) layers are the three layers that make up a CNN. These layers are stacked to construct a CNN architecture. Beyond these layers, two more important factors are the dropout layer and the activation function. The RGB (Red, Green, and Blue) color space, which extends from 0 to 255, can be used to extract and identify various visual aspects for analysis (a process known as feature extraction) using a convolution tool. The network used for feature extraction is composed of multiple pairs of convolutional or pooling layers.

ResNet 50 a kind of deep neural network architecture called ResNet (Residual Network) was introduced to solve the vanishing gradient problem that arises when deep convolutional neural networks (CNNs) are being trained. ResNet gives the network the ability to learn residual functions by introducing skip connections, also referred to as residual connections. The model can bypass specific levels thanks to these skip connections, which send the image straight to the output layers. This facilitates the training of extremely deep networks by reducing the impact of the vanishing gradient issue. ResNet can be a key component in feature extraction from photos within an image caption generator. In order to obtain useful features from the images, the encoder portion of an image captioning model usually makes use of a CNN that has already been trained, such as ResNet. The goal is to extract high-level features from photos by using the information that the pre-trained ResNet model has acquired on a sizable dataset.

Images are used to obtain features using the ResNet model. For image classification tasks, the model is usually pre-trained on a large dataset (e.g., ImageNet). Hierarchical and abstract aspects in photos are captured by the weights that were learned during pre-training. The pre-trained ResNet model is used to obtain characteristics from intermediate layers given an input image. The input's high-level visual information is represented by the features. The decoder component of the picture captioning model receives the features that were extracted from the image. Based on the input attributes, the decoder—which is frequently implemented as a transformer or recurrent neural network (RNN)—creates a textual description of the image.

RNN

A Recurrent Neural Network (RNN), which takes as its input the output of the previous stage. Conventional neural networks have inputs and outputs that are independent of each other. On the other hand, the preceding words must be retained in cases when guessing the next word in a sentence requires knowledge of the preceding words. RNN was developed as a result, and it approached this issue by using a Hidden Layer. Its main and most important characteristic is the Hidden state of an RNN, which holds some information about a sequence. It is sometimes referred to as Memory State since it keeps track of previous inputs to the network.

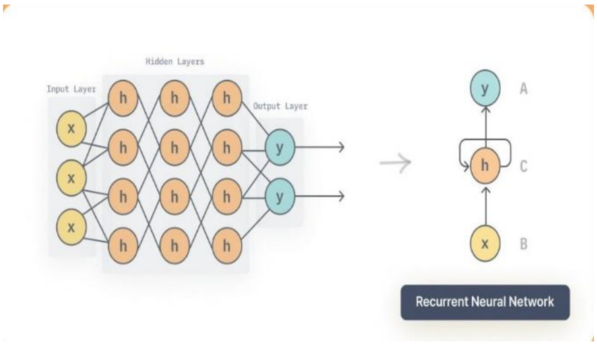


Fig6.: Single RNN

Similar work is done on all the hidden layers to produce the output, using the same settings for every hidden layer. Unlike other neural networks, this lowers the parameter complexity. Although it isn't referred to as a "recurrent neuron," a recurrent unit is the basic processing unit of a recurrent neural network (RNN). Because of its special capacity to preserve a hidden state, this unit enables the network to recognize sequential relationships by processing and remembering prior inputs. RNNs have the same input and output architecture as other deep neural architectures. However, the information flow differs depending on the input and output. It generates the output by doing the same operation on all inputs or hidden layers, using the same parameters for each input. It differs from other neural networks in that it has less parameter complexity. A recurrent unit, although not referred to as a "recurrent neuron," is the central processing unit of a recurrent neural network (RNN). This unit's unique ability to maintain a concealed state allows the network to process and retain previous inputs, which in turn helps the network understand sequential relationships. Like other deep neural architectures, RNNs have the same input and output architecture. Yet, there are differences in the way data is transferred from input to output.

The formula for calculating the current state

$$ht = f(ht-1,xt)$$

Where, ht -> current state

ht-1 -> previous state

Formula for applying Activation function(tanh)

$$h=tanh(Whhht-1+Wxhxt)$$

Where, Whh -> weight at recurrent neuron

Wxh -> weight at input neuron

The formula for calculating output:

Yt=Whyht

Where, Yt -> output

Why -> weight at output layer

LSTM

The Long Short-Term Memory, or LSTM, is an enhanced RNN. For sequence prediction tasks, LSTM performs remarkably well in capturing long-term dependencies. The RNN method has certain drawbacks, which we address by introducing the LSTM algorithm.

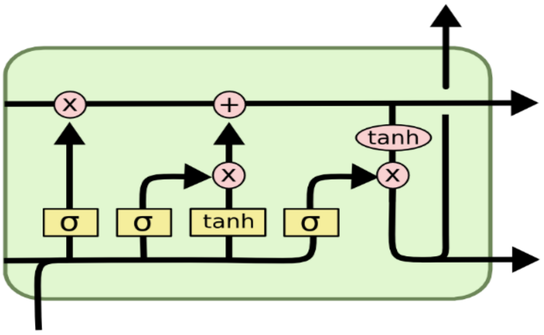


Fig7.: Architecture of LSTM

It is a difficult task to a network to learn long-term dependencies in a standard RNN since it only has one hidden state that is retained over time. In contrast, LSTMs solve this issue by introducing memory cells, which are containers that can store information for a longer amount of time.
Three gates govern a memory cell:
i. Input gate
ii. Forget gate
iii. Output gate
The gates determine which data should be input into, taken out of, and output from the memory cell. •   Unlike RNNs, which lack memory units, LSTMs have a unique memory unit that enables them to recognize long-term dependencies in sequential data. While RNN is also meant to process sequential data, its memory capacity is constrained. In contrast, LSTM is well suited for handling sequential data. Compared to RNN, the training process of LSTM is slower because of its increased complexity. Because of its more straightforward architecture, RNNs train a little bit faster. Long

sequences are more effective for LSTM, but RNN finds it difficult to store information.

RESULT

This is the home page where we land after clicking on the link.

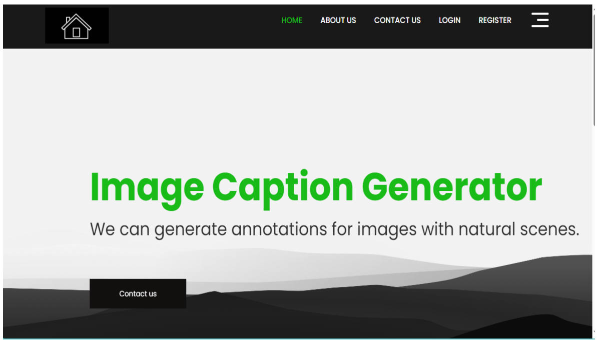
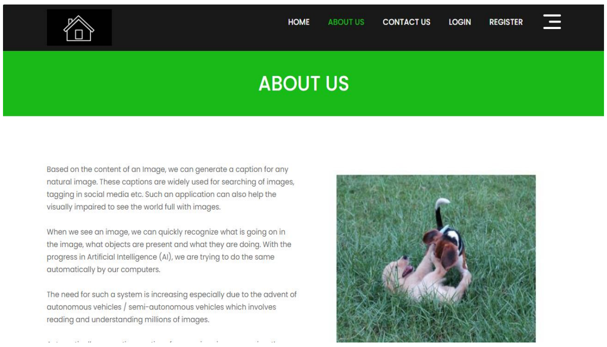


Image Caption Generator

We can generate annotations for images with natural scenes.

Fig 8.: Home Page



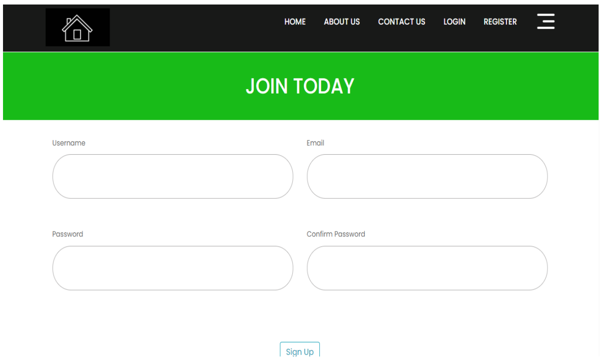
ABOUT US

Based on the content of an Image, we can generate a caption for any natural image. These captions are widely used for searching of images, tagging in social media etc. Such an application can also help the visually impaired to see the world full with images.

When we see an image, we can quickly recognize what is going on in the image, what objects are present and what they are doing. With the progress in Artificial Intelligence (AI), we are trying to do the same automatically by our computers.

The need for such a system is increasing especially due to the advent of autonomous vehicles / semi-autonomous vehicles which involves reading and understanding millions of images.

Fig 9.: About Page



JOIN TODAY

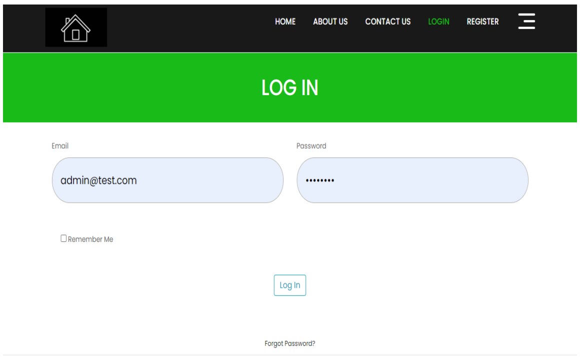
Username

Email

Password

Confirm Password

Fig 10.: Register Page



Email
admin@test.com

Password
•••••••••

Fig 11.: Login Page

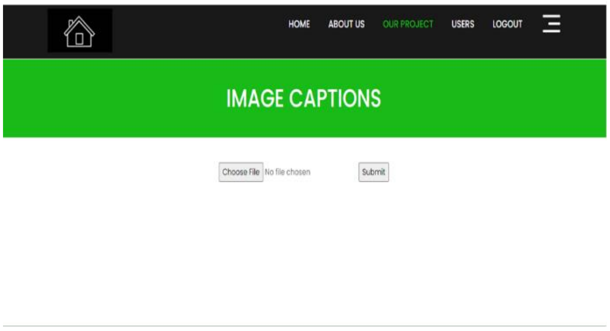
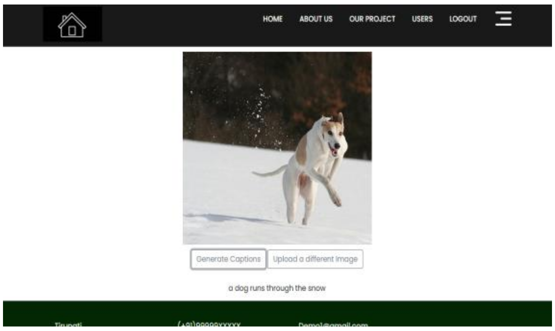


Fig 12.: Upload Page



a dog runs through the snow

Fig 13.: Caption Generated

CONCLUSION

The issue of creating meaningful captions for images has been found to be powerfully and effectively solved by the image caption generator that combines Long Short-Term Memory (LSTM) networks with Convolutional Neural Networks (CNNs). Using CNN layers to obtain relevant characteristics and capture spatial information, the CNN-LSTM model showed how to generate contextually relevant captions by efficiently utilizing LSTM layers. Visual perception and sequential data processing work together to address the problems of picture understanding and natural language synthesis through the integration of these two architectures. This work emphasizes the need of merging all the layers to produce better results from challenging tasks like image captioning.

Future research can build upon and enhance the CNN and LSTM picture caption generator in a number of ways. First, by investigating more complex architectures like attention processes, transformer models, or language models that have already undergone training, like BERT, the model may be able to better capture the complex interactions between textual and visual data.

Moreover, enlarging and diversifying the dataset inside the training set might enhance the generalization of the model, enabling it to precisely depict a greater range of images. Optimizing the model for certain domains or tasks could also be useful, allowing the generator to focus on areas such as medical imaging or satellite photography. Furthermore, investigating ways to make the model more interpretable and controllable could aid in enhancing understanding and providing direction for the captioning process. Finally, but just as importantly, putting the model to use in real situations and gathering user feedback would reveal its practicality and highlight areas for development.

REFERENCES

1. M Sailaja, K Harika, B Sridhar. Rajan Singh, "Image Caption Generator using Deep Learning", 2022 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC).
2. C S Kanimozhiselvi, Karthika V, Kalavani S P, Krithika S, "Image Captioning Using Deep Learning", 2022 International Conference on Computer Communication and Informatics (ICCCI).
3. Chetan Amritkar, Vaishali Jabade, "Image Caption Generation Using Deep Learning Technique", 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA).
4. Varsha Kesavan, Vaidehi Muley, Megha Kolhekar, "Deep Learning based Automatic Image Caption Generation", 2019 Global Conference for Advancement in Technology (GCAT).
5. Lakshmi narasimhan Srinivasan, Dinesh Sreekanthan and A.L Amutha, "Image captioning - A Deep Learning Approach", International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 9 (2018) pp.
6. T. J. Buschman and E. K. Miller. "Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices. Science", 315(5820):1860–1862, 2007.
7. William Fedus, Ian Goodfellow, Andrew M Dai. Maskgan, "Better text generation", 1801.07736, 47, 2018.
8. D. Elliott, F. Keller, "Image Description using Visual Dependency Representations", Conference on Empirical Methods in Natural Language Processing.
9. N K Kumar, D Vigneswari, A Mohan, K Laxman, J Yuvaraj, "Detection and Recognition of Objects in Image Caption Generator System: A Deep Learning Approach", IEEE – 2019..
10. C. Amritkar, V. Jabade, "Image Caption Generation using Deep Learning Technique", IEEE Access, 2018.

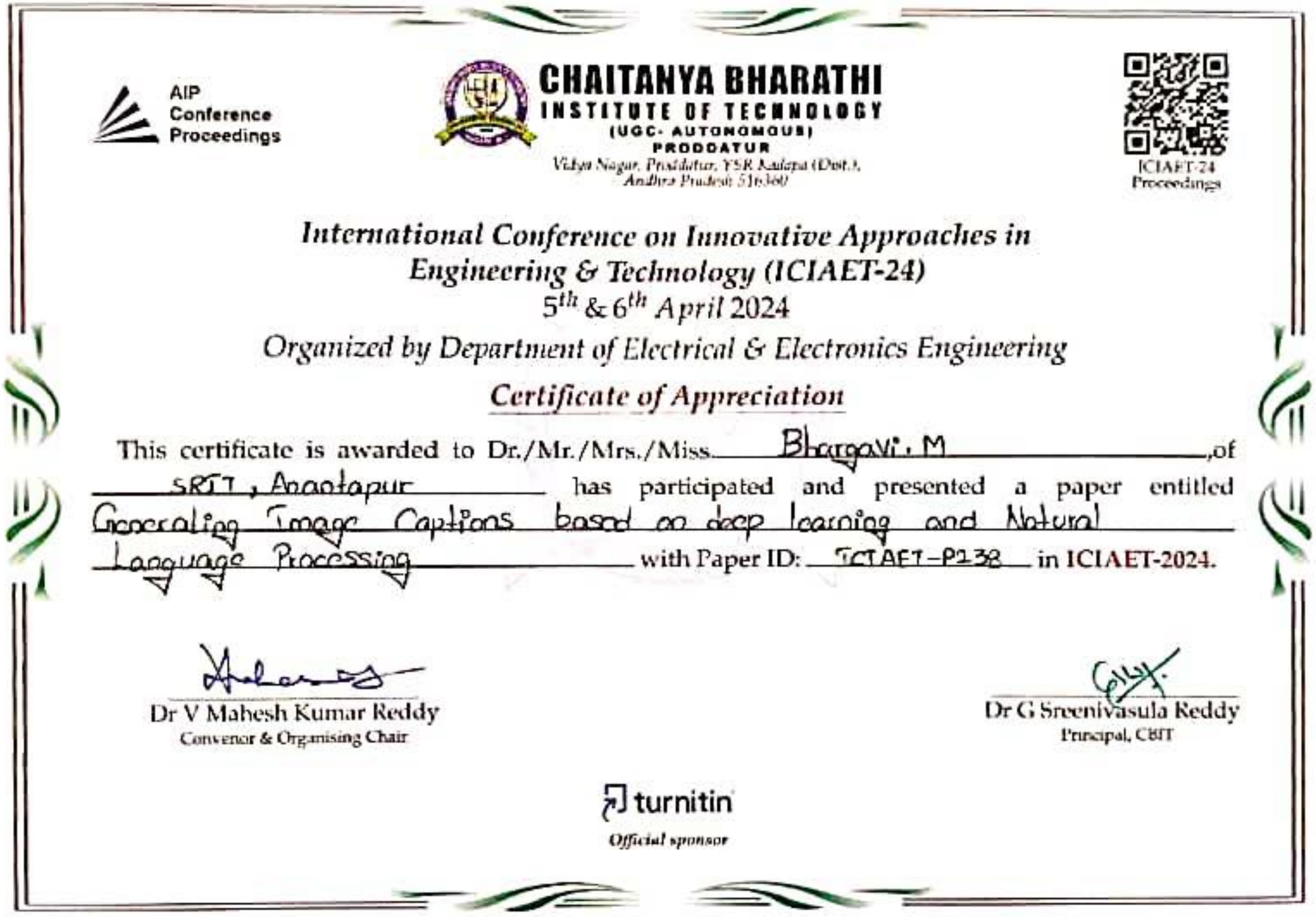
11. Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. "Baby talk: Understanding and generating image descriptions", IEEE Transactions on Pattern Analysis and Machine Intelligence, 35:2891–2903, June 2013.

12. "Bottom-up and top-down attention for image captioning and visual question answering" by Anderson, Peter, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. (2018)

13. "Meshed-memory transformer for image captioning" by Cornia, Marcella, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara (2020).

14. "Unified vision-language pre-training for image captioning" and vqa by Zhou, Luowei, et al (2020).

15. Xiaodan Liang, Zhiting Hu, Hao Zhang, Chuang Ggan ,and Eric P Xing.2017. "Recurrent Topic-Transition for Visual Paragraph Generation."



AIP
Conference
Proceedings

CHAITANYA BHARATHI
INSTITUTE OF TECHNOLOGY
(UGC- AUTONOMOUS)
PRODDATUR
Vidya Nagar, Proddatur, YSR Kadapa (Dist.),
Andhra Pradesh 516360

ICIAET-24
Proceedings

International Conference on Innovative Approaches in Engineering & Technology (ICIAET-24)
5th & 6th April 2024

Organized by Department of Electrical & Electronics Engineering

Certificate of Appreciation

This certificate is awarded to Dr./Mr./Mrs./Miss. Bhargavi. M ,of SRIT, Anantapur has participated and presented a paper entitled Generating Image Captions based on deep learning and Natural Language Processing with Paper ID: ICIAET-P138 in **ICIAET-2024.**

Dr V Mahesh Kumar Reddy
Convenor & Organising Chair

Dr G Sreenivasula Reddy
Principal, CBIT

turnitin
Official sponsor