

IMAGE CAPTIONING

A PROJECT REPORT

Submitted by

DEEP SUREJA (20BECE30255)

JASH UMRETIYA(20BECE30270)

DHRUVIN VACHHANI (20BECE30271)

HITARTH UPADHYAY (20BECE30286)

In fulfillment for the award of the degree

of

BACHELOR OF ENGINEERING

in

Computer Engineering



LDRP Institute of Technology and Research, Gandhinagar

Kadi Sarva Vishwavidyalaya

April,2023

LDRP INSTITUTE OF TECHNOLOGY AND RESEARCH

GANDHINAGAR

CE-IT Department



CERTIFICATE

This is to certify that the Project Work entitled **“Image Captioning”** has been carried out by **Deep Sureja (20BECE30255)** under my guidance in fulfilment of the degree of Bachelor of Engineering in Computer Engineering Semester-6 of Kadi Sarva Vishwavidyalaya University during the academic year 2020-24.

Dr. Lokesh Gagnani

Internal Guide

LDRP ITR

Dr. Sandip Modha

Head of the Department

LDRP ITR

LDRP INSTITUTE OF TECHNOLOGY AND RESEARCH

GANDHINAGAR

CE-IT Department



CERTIFICATE

This is to certify that the Project Work entitled **“Image Captioning”** has been carried out by **Jash Umretiya (20BECE30270)** under my guidance in fulfilment of the degree of Bachelor of Engineering in Computer Engineering Semester-6 of Kadi Sarva Vishwavidyalaya University during the academic year **2020-24**.

Dr. Lokesh Gagnani

Internal Guide

LDRP ITR

Dr. Sandip Modha

Head of the Department

LDRP ITR

LDRP INSTITUTE OF TECHNOLOGY AND RESEARCH

GANDHINAGAR

CE-IT Department



CERTIFICATE

This is to certify that the Project Work entitled **“Image Captioning”** has been carried out by **Dhruvin Vachhani (20BECE30271)** under my guidance in fulfilment of the degree of Bachelor of Engineering in Computer Engineering Semester-6 of Kadi Sarva Vishwavidyalaya University during the academic year **2020-24**.

Dr. Lokesh Gagnani

Internal Guide

LDRP ITR

Dr. Sandip Modha

Head of the Department

LDRP ITR

LDRP INSTITUTE OF TECHNOLOGY AND RESEARCH

GANDHINAGAR

CE-IT Department



CERTIFICATE

This is to certify that the Project Work entitled **“Image Captioning”** has been carried out by **Hitarth Upadhyay (20BECE30286)** under my guidance in fulfilment of the degree of Bachelor of Engineering in Computer Engineering Semester-6 of Kadi Sarva Vishwavidyalaya University during the academic year **2020-24**.

Dr. Lokesh Gagnani

Internal Guide

LDRP ITR

Dr. Sandip Modha

Head of the Department

LDRP ITR

Presentation-I for Project-I

1. Name & Signature of Internal Guide	
2. Comments from Panel Members	
3. Name & Signature of Panel Members	

ACKNOWLEDGEMENT

I take this opportunity to express my profound gratitude and deep regards to my guide for her exemplary guidance, monitoring and constant encouragement throughout the course of this thesis. The blessing, help and guidance given by her time to time shall carry me a long way in the journey of life on which I am about to embark. I also take this opportunity to express a deep sense of gratitude to University for cordial support, valuable information and guidance, which helped me in completing this task through various stages.

Deep Sureja (20BECE30255)

Jash Umretiya (20BECE30270)

Dhruvin Vachhani (20BECE30271)

Hitarth Upadhyay (20BECE30286)

ABSTRACT

With the development of deep learning, the combination of computer vision and natural language process has aroused great attention in the past few years. Image captioning is a representative of this field, which makes the computer learn to use one or more sentences to understand the visual content of an image. The meaningful description generation process of high level image semantics requires not only the recognition of the object and the scene, but the ability of analyzing the state, the attributes and the relationship among these objects.

In this project, we have tried to develop a model which can take an image as an input and output a sentence that can describe the things in that picture. After that it also converts the sentences into speech. The model uses the Flickr8 dataset for the training purpose. The components of the method we used are: Convolutional Neural Network (CNN), and sentence generation. The image is captioned by Recognizing the objects that appear in the input image, using automatic feature engineering. Image captioning has a lot of important usage these days as the motivation behind this project comes from several real-life scenarios like- Self driving cars, Aid to the blind etc.

TABLE OF CONTENTS:

Acknowledgement	i
Abstract	ii
Table of Contents	iii
List of Figures	v
1 Introduction	1
1.1 Introduction	1
1.2 Scope	2
1.3 Project Summary	3
1.4 Project Purpose	4
2 Technology and Literature Review	5
2.1 Tools and Technology	5
2.1.1 Front-end technologies	5
2.1.2 Backend technologies	5
2.2 Methodologies	10
2.2.1 Data source	10
2.2.2 Data preprocessing (image)	10
2.2.3 Data preprocessing (caption)	10
3 System Requirements Study	11
3.1 User Characteristics	11
3.2 Hardware and Software Requirements	11
3.2.1 Software requirements	11
3.2.2 Hardware requirements	11
3.3 Non-Functional and Functional Requirements	12
3.4 Assumptions and Dependencies	13
3.4.1 Assumptions	13
3.4.2 Dependencies	13
4 System Diagrams	14
4.1 Model	14
4.1.1 ResNet-50 model	14
4.1.2 LSTM model	15
4.2 Prototype	16
4.3 System Diagram	17
5 Data Dictionary	23
5.1 Probability Distribution	23
5.2 Description Relationship of Data	24
5.3 Performance of Dataset Flickr8k-SAU	24
5.4 Representation of Image Data	25
6 System Testing	26
6.1 Test Plan	26
6.2 Testing Strategy	26
6.3 Testing Methods	27
7 Result, Discussion and Conclusion	29
7.1 Result	29
7.2 Discussion	30
7.2.1 Benefits	30

7.2.2 Intelligent monitoring	30
7.2.3 Image and video annotation	30
7.2.4 Inconsistent objects during training and testing	30
7.2.5 cross-language text description of images	30
7.3 Conclusion	31
8 References	32

List of Figures

Fg No.	Figure Name	Page No.
4.1.1	ResNet-50 Model	14
4.2.1	Prototype	16
4.2.2	ResNet-50 Ex.	16
4.3.1	Overall Architecture	17
4.3.2	Flow Chart	18
4.3.3	Block Diagram	19
4.3.4	Block Diagram of Multimodal Space	19
4.3.4	Sequence Diagram	20
4.3.5	State Diagram	21
5	Example	25

1. INTRODUCTION

1.1 Introduction

Image caption generator is a task that involves computer vision and natural language Processing concepts to recognize the context of an image and describe them in a natural language like English. For a machine to be able to automatically describe objects in an image with its relationships or the work being done using a learned language model is a daunting task, but plays an essential role in many areas. For example, it can help visually impaired people with better under-stand visual input, thereby acting as a facilitator or a guide . The generated image caption should not only contain the image object names but also their properties, relationships and functions. In addition, the generated caption must be expressed through a natural language such as English.

The model uses the Flickr8k dataset for training purpose. The components of the method we use: Convolutional Neural Network(CNN),Recurrent Neural Network (RNN) and SentenceGeneration. In this paper, CNN is used to create a dense feature vector. This dense vector is Also Called an embedding. For an image caption model, it becomes a dense representation of the embedding image and is used as the initial state of the LSTM. At each time-step, the LSTM considers the previous cell position and outputs a prediction for the most likely next value in the sequence.

With the rapid development of digitalization, there are a huge amount of images, Accompanied With a lot of related texts. Automatic image captioning has recently Attracted much research interest. The objective of automatic image captioning is to generate properly formed English sentences to describe the content of an image automatically , which is of great impact in various domains such as virtual assistants, image indexing, recommendation in editing applications, and the help of the disabled. Although it is an easy task for a human to describe an image, it becomes very difficult for A machine to perform such a task. Image captioning does not only need to detect the objects contained in an image but also capture how these objects related to each other and their attributes as well as the activities involved in.

1.2 Scope

Deep learning is advanced up to now exact caption generation is not possible due to many reasons like hardware requirements problem, no proper programming logic or model to generate the exact captions because machines cannot think or make decisions as accurately as humans do. So in future with the advancement of hardware and deep learning models we hope to generate captions with higher accuracy. It is also thought to extend this model and build complete Image - Speech conversion by converting captions of images to speech. This is very much helpful for blind people.

Certain images are not well recognized and we found out that there is, still some scope of improvement. There are certain points which can be incorporated into this model to make it even better like larger dataset, using Beam Search to generate captions, BLEU score can be implemented for performance measurement, text-to-speech technology.

Key features of the System :-

- **Home page:** The application contains only one page that is home page which accepts any image from user and generates appropriate caption for that and converts that caption into speech.

1.3 Project summary

Image caption generator is a task that involves computer vision and natural language processing concept to recognize the context of an image and describe them in a natural language like English. For a machine to be able to automatically describe objects in an image with its relationships or the work being done using a learned language model is a daunting task, but plays an essential role in many areas. For example, it can help visually impaired people with better understanding of visual input, thereby acting as a facilitator guide. The generated image caption should not only contain the image object names, but also their properties, relationships and functions. In addition, the generated caption must be expressed through a natural language such as English. The model uses the Flickr8 dataset for training purpose. The components of the method we use are: Convolutional Neural Network(CNN), Recurrent Neural Network(RNN) and Sentence Generation.

CNN is used to create a dense feature vector. This dense vector is also called an embedding. For an image caption model, it becomes a dense representation of the embedding image and is used as the initial state of the LSTM. At each time-step, the LSTM considers the previous cell position and outputs a prediction for the most likely next value in the sequence.

1.4 Project purpose

This project is specifically design for blind people. People who are enable to see anything can also visualize an image a nd can understand what is happening in particular image.

2. Technology and Literature Review

2.1 Tools and Technology :

2.1.1 Front-end technologies

- **Flask**
 - A Web Application Framework or a simply a Web Framework represents a collection of libraries and modules that enable web application developers to write applications without worrying about low-level details such as protocol, thread management, and so on.
- **HTML**
 - HTML is Hyper Text Markup Language used to design the architecture of anywebsite.
 - It is a simple markup language with easy using tags and powerful features in HTML5.
 - HTML is used in this project to design GUI portion of website for easy access and good elegance.
- **CSS**
 - Cascading Style Sheet is used for good elegance and appearance of website.
 - It is used for designing web pages with different features and functionalities such as scroll pane, floating effect etc.

2.1.2 Backend technologies

- **Python**
 - Python is an interpreted high-level general-purpose programming language. Its design philosophy emphasizes code readability with its use of significant indentation. Its language constructs as well as its object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects.

- Python is dynamically-typed and garbage-collected. It supports multiple programming paradigms, including structured (particularly, procedural), object-oriented and functional programming. It is often described as a "batteries included" language due to its comprehensive standard library.
- In this Project python is used as main coding language as python supports many featured libraries. From them such libraries are used here, and they are: numpy, pandas, glob, matplotlib etc.
- NumPy is used to create numpy arrays for less memory consumption fast execution of program.
- Glob library is used to return all file paths for images and their captions files.
- Matplotlib library is used to plot graph for visualization of model.
- **Open CV**
 - OpenCV (Open Source Computer Vision Library) is an open source computer vision and machine learning software library. OpenCV was built to provide a common infrastructure for computer vision applications and to accelerate the use of machine perception in the commercial products.
 - Computer Vision can be defined as a discipline that explains how to reconstruct, interpret, and understand a 3D scene from its 2D images, in terms of the properties of the structure present in the scene. It deals with modeling and replicating human vision using computer software and hardware.
 - Features of OpenCV Library -:
 - Capture and save videos
 - Process images (filter, transform)
 - Perform feature detection
 - Detect specific objects such as faces, eyes, cars, in the videos or images.
 - Analyze the video, i.e., estimate the motion in it, subtract the background, and track objects in it.
 - Here open cv is to process images and also for image visualization.
 - OpenCV is a great tool for image processing and performing computer vision tasks.

- **Deep Learning**

- Deep learning is a machine learning technique that teaches computers to do what comes naturally to humans. Deep learning is a subset of machine learning where artificial neural networks, algorithms inspired by the human brain, learn from large amounts of data.
- Deep learning allows machines to solve complex problems even when using a data set that is very diverse, unstructured and inter-connected.
- Deep learning is a branch of machine learning which is completely based on artificial neural networks. Deep learning is an artificial intelligence function that imitates the workings of the human brain in processing data and creating patterns for use in decision making.
- Deep learning is a subset of machine learning in artificial intelligence (AI) that has networks capable of learning unsupervised from data that is unstructured or unlabelled. It has a greater number of hidden layers and known as deep neural learning or deep neural network.
- Deep learning has evolved hand-in-hand with the digital era, which has brought about an explosion of data in all forms and from every region of the world. This data, known simply as big data, is drawn from sources like social media, internet search engines, ecommerce platforms, and online cinemas, among others.
- This enormous amount of data is readily accessible and can be shared through fintech applications like cloud computing. However, the data, which normally is unstructured, is so vast that it could take decades for humans to comprehend it and extract relevant information.
- Deep learning utilizes a hierarchical level of artificial neural networks to carry out the process of machine learning. The artificial neural networks are built like the human brain, with neuron nodes connected like a web. While traditional programs build analysis with data in a linear way, the hierarchical function of deep learning systems enables machines to process data with a nonlinear approach.
- Companies realize the incredible potential that can result from unravelling this wealth of information and are increasingly adapting to AI systems for automated support.

- **Machine Learning**

- Machine Learning is an idea to learn from examples and experience, without being explicitly programmed. Instead of writing code, you feed data to the generic algorithm, and it builds logic based on the data given.
- Machine Learning is a field which is raised out of Artificial Intelligence(AI). Applying AI, we wanted to build better and intelligent machines.
- But except for few mere tasks such as finding the shortest path between point A and B, we were unable to program more complex and constantly evolving challenges.
- There was a realisation that the only way to be able to achieve this task was to let machine learn from itself. This sounds similar to a child learning from its self. So machine learning was developed as a new capability for computers.
- And now machine learning is present in so many segments of technology, that didn't even realise it while using it.

- **Supervised Learning**

- Supervised learning is the most popular paradigm for machine learning. It is the easiest to understand and the simplest to implement. It is the task of learning a function that maps an input to an output based on example input-output pairs. It infers a function from labelled training data consisting of a set of training examples.
- In supervised learning, each example is a pair consisting of an input object (typically a vector) and a desired output value (also called the supervisory signal). A supervised learning algorithm analyses the training data and produces an inferred function, which can be used for mapping new examples.
- Over time, the algorithm will learn to approximate the exact nature of the relationship between examples and their labels. When fully trained, the supervised learning algorithm will be able to observe a new, never-before-seen example and predict a good label for it.

- **Unsupervised Learning**

- Unsupervised Learning is a machine learning technique, where you do not need to supervise the model. Instead, you need to allow the model to work on its own to discover information.
- It mainly deals with the unlabelled data and looks for previously undetected patterns in a data set with no pre-existing labels and with a minimum of human supervision.
- In contrast to supervised learning that usually makes use of human-labelled data, unsupervised learning, also known as self-organization, allows for modelling of probability densities over inputs. Unsupervised machine learning algorithms infer patterns from a dataset without reference to known, or labelled outcomes.
- It is the training of machine using information that is neither classified nor labelled and allowing the algorithm to act on that information without guidance.
- Here the task of machine is to group unsorted information according to similarities, patterns, and differences without any prior training of data.
- Unlike supervised learning, no teacher is provided that means no training will be given to the machine. Therefore, machine is restricted to find the hidden structure in unlabelled data by our- self.

- **Reinforcement Learning**

- Reinforcement Learning (RL) is a type of machine learning technique that enables an agent to learn in an interactive environment by trial and error using feedback from its own actions and experiences.
- Machine mainly learns from past experiences and tries to perform best possible solution to a certain problem. It is the training of machine learning models to make a sequence of decisions.
- Though both supervised and reinforcement learning use mapping between input and output, unlike supervised learning where the feedback provided to the agent is correct set of actions for performing a task, reinforcement learning uses rewards and punishments as signals for positive and negative behaviour.
- Reinforcement learning is currently the most effective way to hint machine's creativity.

2.2 Methodologies :

2.2.1 Data Source

- In our project, we have used the Flickr8k image dataset to train the model for understanding how to discover the relation between images and words for generating captions.
- It contains 8000 images in JPEG format with different shapes and sizes and each image has 5 different captions. The images are chosen from 6 different Flickr groups, and do not contain any well-known people or locations. These were manually selected to depict a variety of scenes and situations.
- The images are bifurcated as follows in the code :
 - Training Set — 6000 images
 - Dev Set — 1000 images
 - Test Set — 1000 images

2.2.2 Data Preprocessing (image)

- The feature extractor needs an image 224x224x3 size. The model uses ResNet50 pre trained on Image Net dataset where the features of the image are extracted just before the last layer of classification. Another dense layer is added and converted to get a vector of length 2048.

2.2.3 Data Preprocessing (caption)

- To define the vocabulary, 8253 unique words are tokenized from the training dataset. As computers do not understand English words, we have represented them with numbers and mapped each word of the vocabulary with a unique index value and we encoded each word into a fixed sized vector and represented each word as a number.
- Also, we maintain a list for each caption that stores the next word at each sub- iteration. Further, one hot encoding is applied on the list that contains the next word. Further, both partial sequence and one hot encoded next word are converted into arrays.

3. SYSTEM REQUIREMENTS STUDY

3.1 User Characteristics

- Analyzing user characteristics is an important aspect of any project. It allows us to clearly define and focus on who the end users are for the project. Also, it allows checking the progress of the project to ensure that we are still developing the system for the end users.
- The user must have following characteristics :
 - User must have basic knowledge of Computers.
 - User should understand the use of application.
 - User can easily interact and provide input.

3.2 Software and Hardware Requirements

- Software and Hardware Requirements are used to describe the minimum hardware and software requirements to run the Software. These requirements are described below.

3.2.1 Software Requirements :

- Programming language: Python
- Operating system: Windows
- Tools: Anaconda Navigator /TensorFlow

3.2.2 Hardware Requirements :

- Processor: Intel Multicore processor (i3 or i5)
- RAM: 4GB or above
- Hard disk: 100 GB or above

3.3 Non-Functional and Functional Requirements

3.3.1 Non-Functional Requirements:

Following is a list of nonfunctional requirements:

◆ **Reliability:**

It can be accessed by the end users 24*7 as needed hence is highly reliable for end users.

◆ **Availability:**

Internet connection for end users with the database servers sure and hence the application will be available any time for access.

◆ **Portability:**

The developed web application is portable as it can be accessed from many operation systems regardless Windows, Mac, Linux provided they have a browser to access Internet.

◆ **Consistency:**

The modification in server data with respect to user's response must be server right the time and it must be displayed at every machine having this application.

3.4 Assumptions and Dependencies

3.4.1 Assumptions

- User is the person having enough knowledge for the traversing operation.
- We will provide a user friendly interface so that any user can easily navigate through the system, but he/she should be capable of providing valid credentials for successful login.

3.4.2 Dependencies

- This application depends on technologies and downloaded software requirements.

4. System Diagrams

4.1 Model :

4.1.1 ResNet-50 Model

- Deep residual networks like the popular ResNet-50 model is a convolutional neural network (CNN) that is 50 layers deep.
- A Residual Neural Network (ResNet) is an Artificial Neural Network (ANN) of a kind that stacks residual blocks on top of each other to form a network.
- ResNet50 is a 50 layer deep convolutional neural network. It is trained on millions of images from the ImageNet dataset. ResNet is highly used for various computer vision tasks. It has the ability to classify images into 1000 categories.

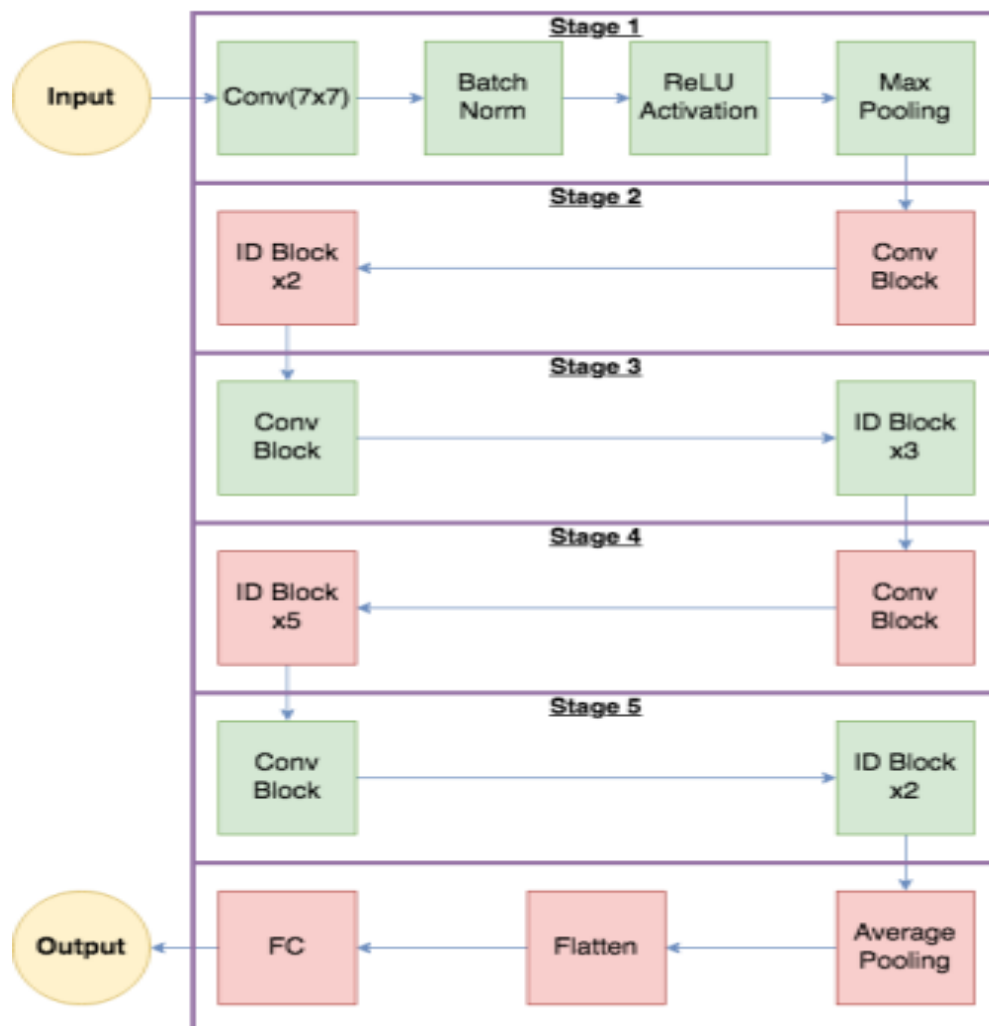


Figure 4.1.1 ResNet-50 Model

4.1.2 LSTM Model

- LSTM is a type of Recurrent Neural Network or RNN in short. LSTM stands for Long Short Term Memory and they are specifically designed to remember information for a long period of time by default.
- They are a special kind of RNN as they can retain both short-term and long-term memory. LSTMs have three main components in their cell, each for forgetting, remembering and updating data.
- As stated before, in LSTM the previous layers inputs and outputs play a part in determining the output of the present layer.
- This is taken to a next level by checking the input of the following layer as well. For example, in the sentences, “The train arrived early at the station” and “Muhammad Ali asked his coach to train him in martial arts as well”, the highlighted word “train” has different meanings.
- Therefore to predict a particular word, the context of the sentence is to be determined first and to predict the context, one has to look at the words that come before and after the said word.
- LSTM uses something called Encoder and Decoder for input and output sequence. Using these LSTM transforms one sequence into a completely different sequence. To understand this, consider an example of an image captioning model.
- The encoder and decoder have their own imaginary language to communicate with each other, so this way, the encoder understands what is present in the input image as well as the imaginary language and the decoder understands the output and the imaginary language. So when the input is an image a CNN encoder generates a hidden state.
- This hidden state is decoded using an LSTM model which acts as our decoder, and a caption is generated word by word. But with all this, the context is also important. This happens just like a human sees a picture and understands what is happening in the picture.
- While looking at a picture a human would also keep in mind all the aspects of the image just to remember the context of the current word.

4.2 Prototype

- To analyze the Choice of the user, we have to use some transform learning algorithms which is the part of deep learning. We selected the resNet-50 algorithm to design a model.

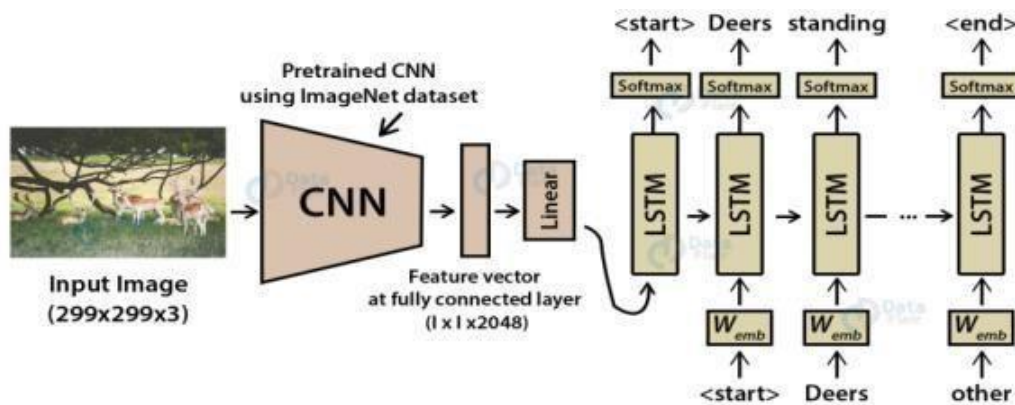


Figure 4.2.1 Prototype

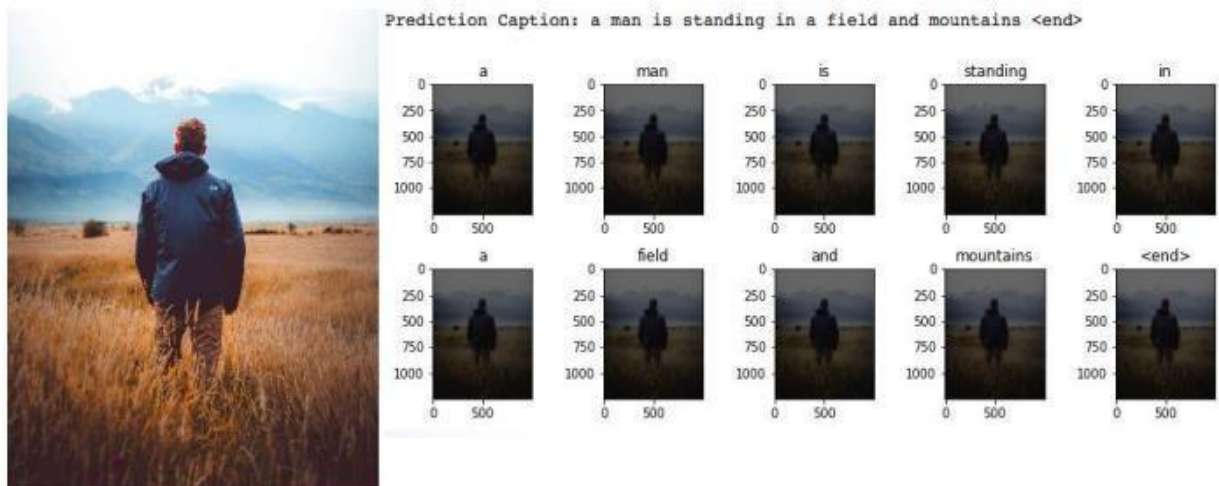


Figure 4.2.2 ResNet-50 Ex.

4.3 System Diagrams

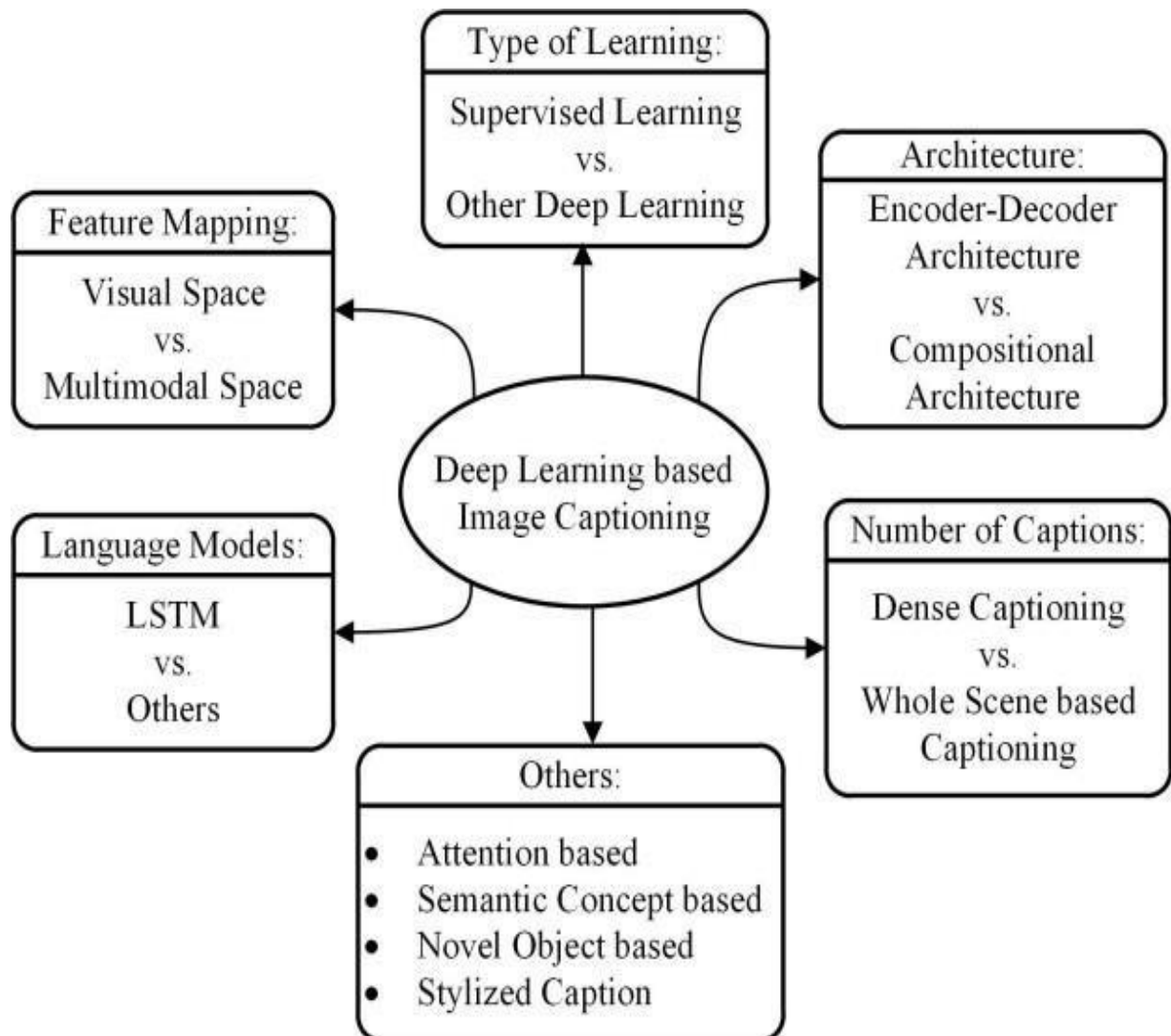


Figure 4.3.1 Overall Architecture

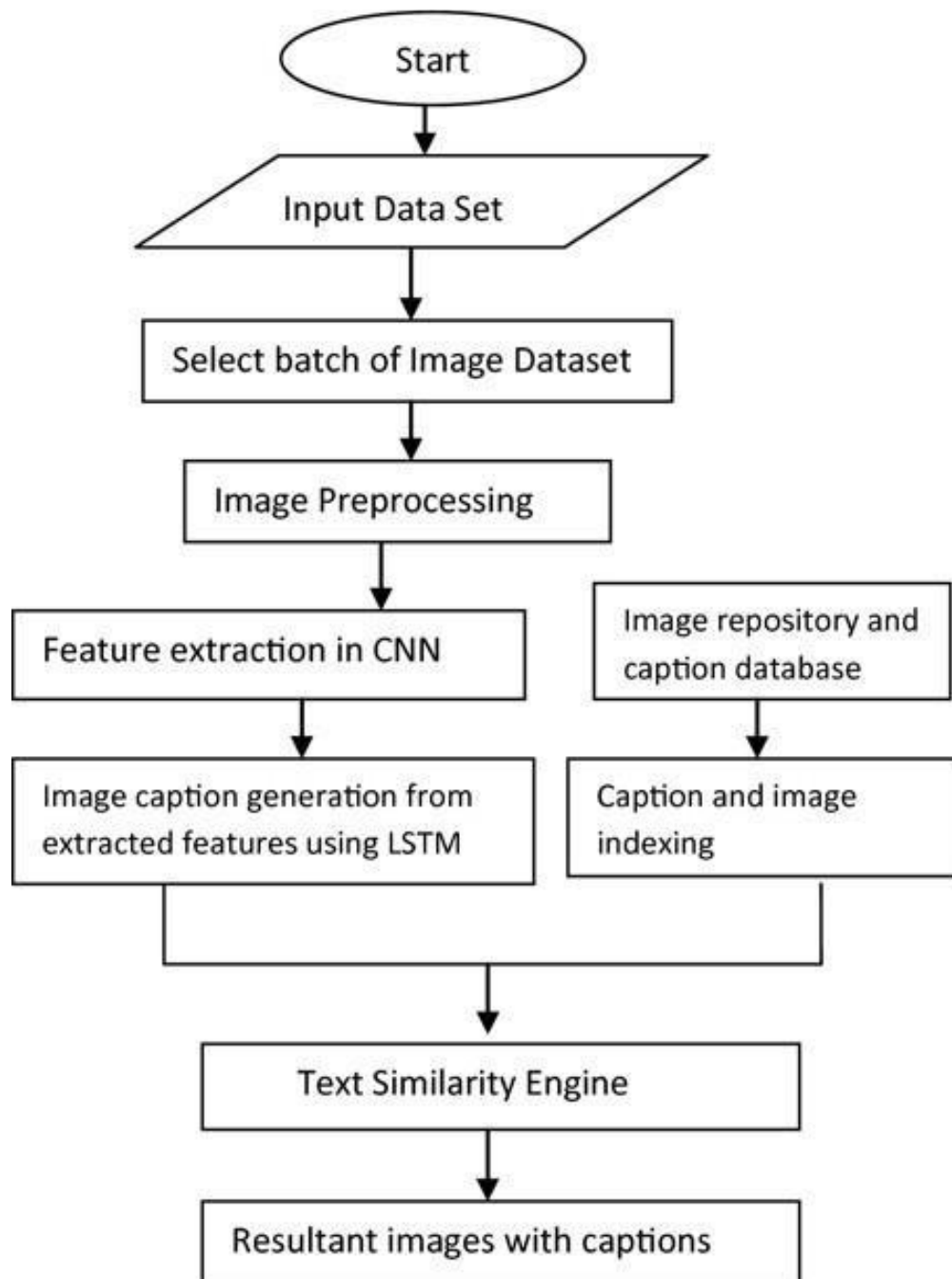


Figure 4.3.2 Flow Chart

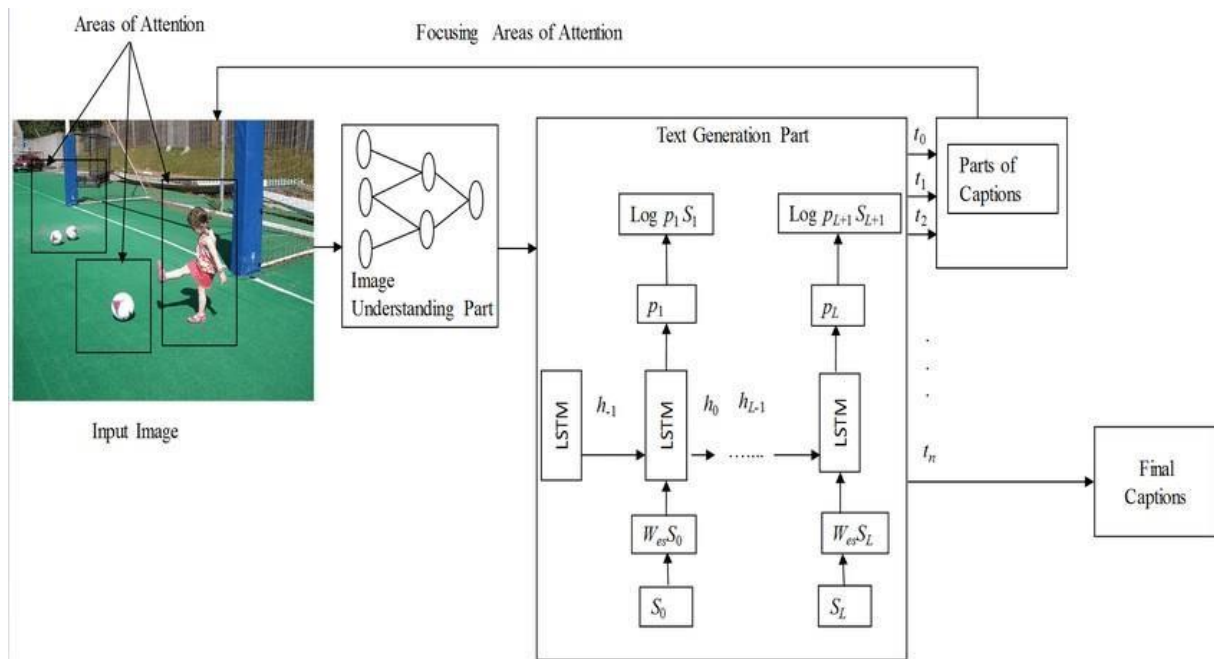


Figure 4.3.3 Block Diagram

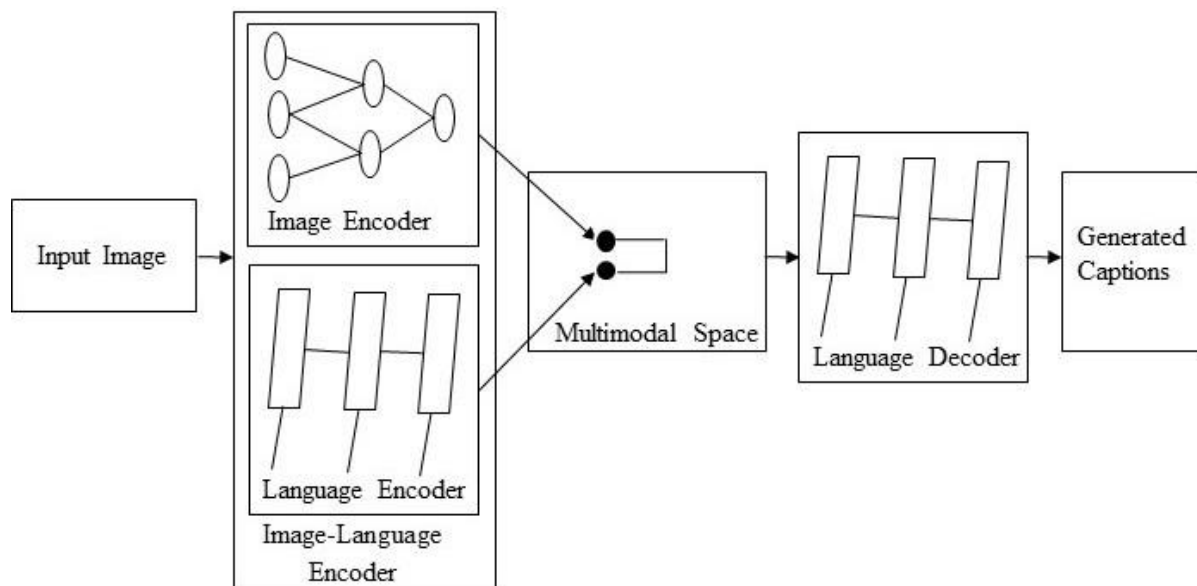
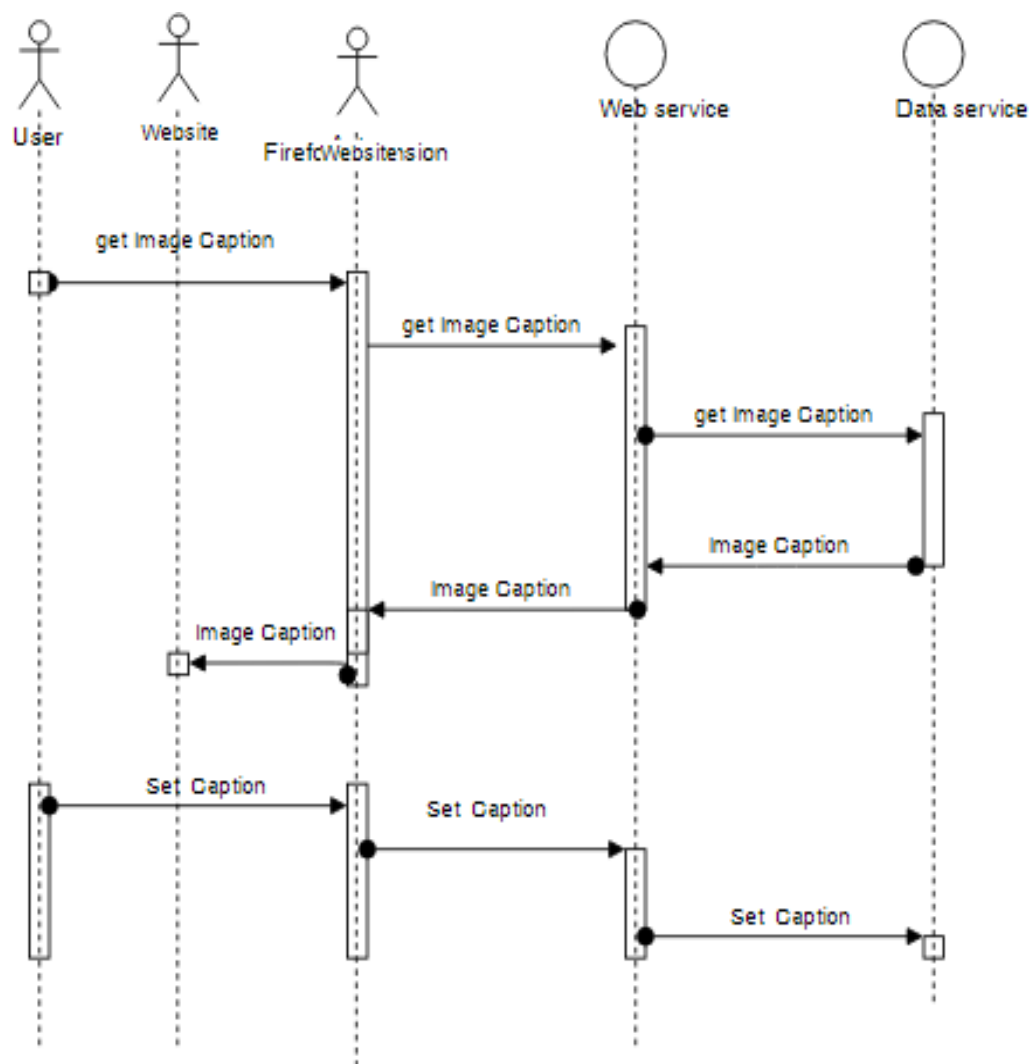


Figure 4.3.4 Block Diagram of Multimodal Space

**Figure 4.3.4 Sequence Diagram**

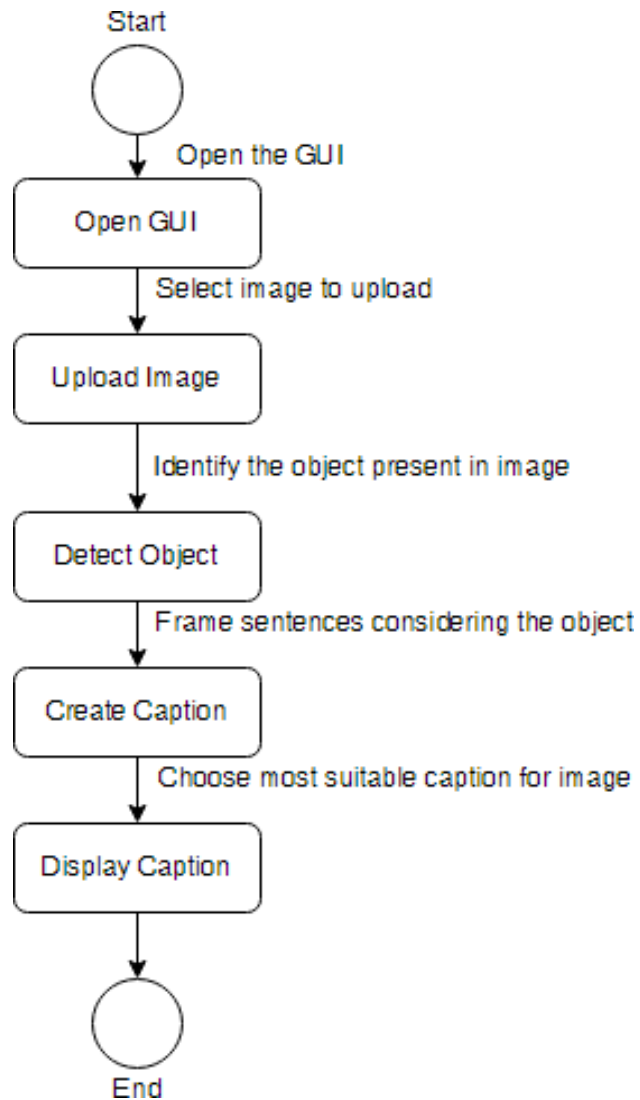


Figure 4.3.5 State Diagram



Figure 4.3.5 DFD Diagram Level 0

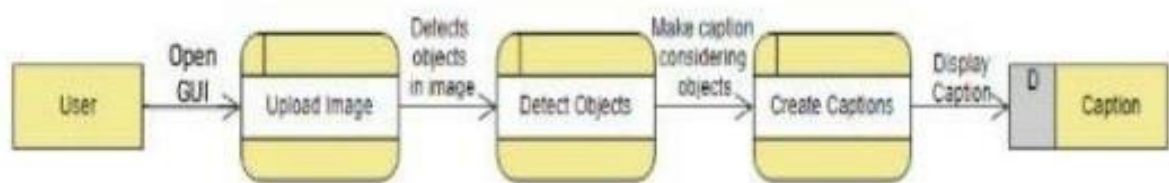


Figure 4.3.6 DFD Diagram Level 1

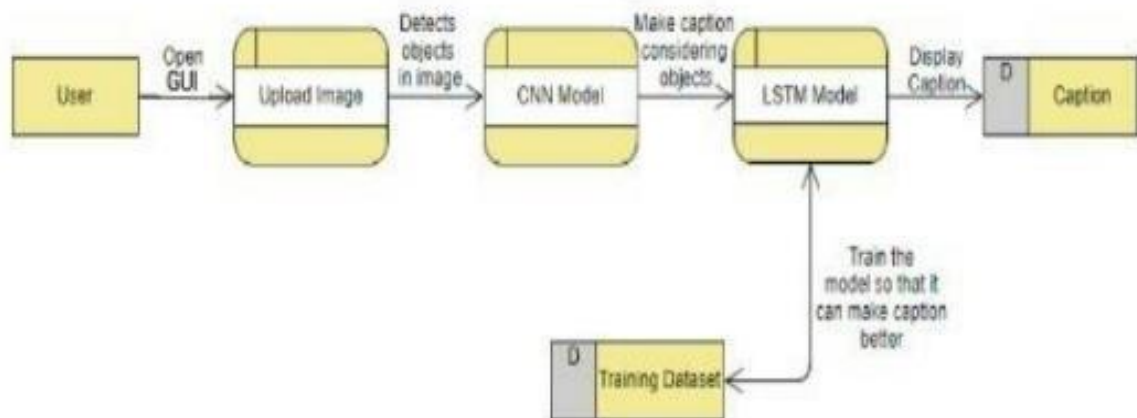
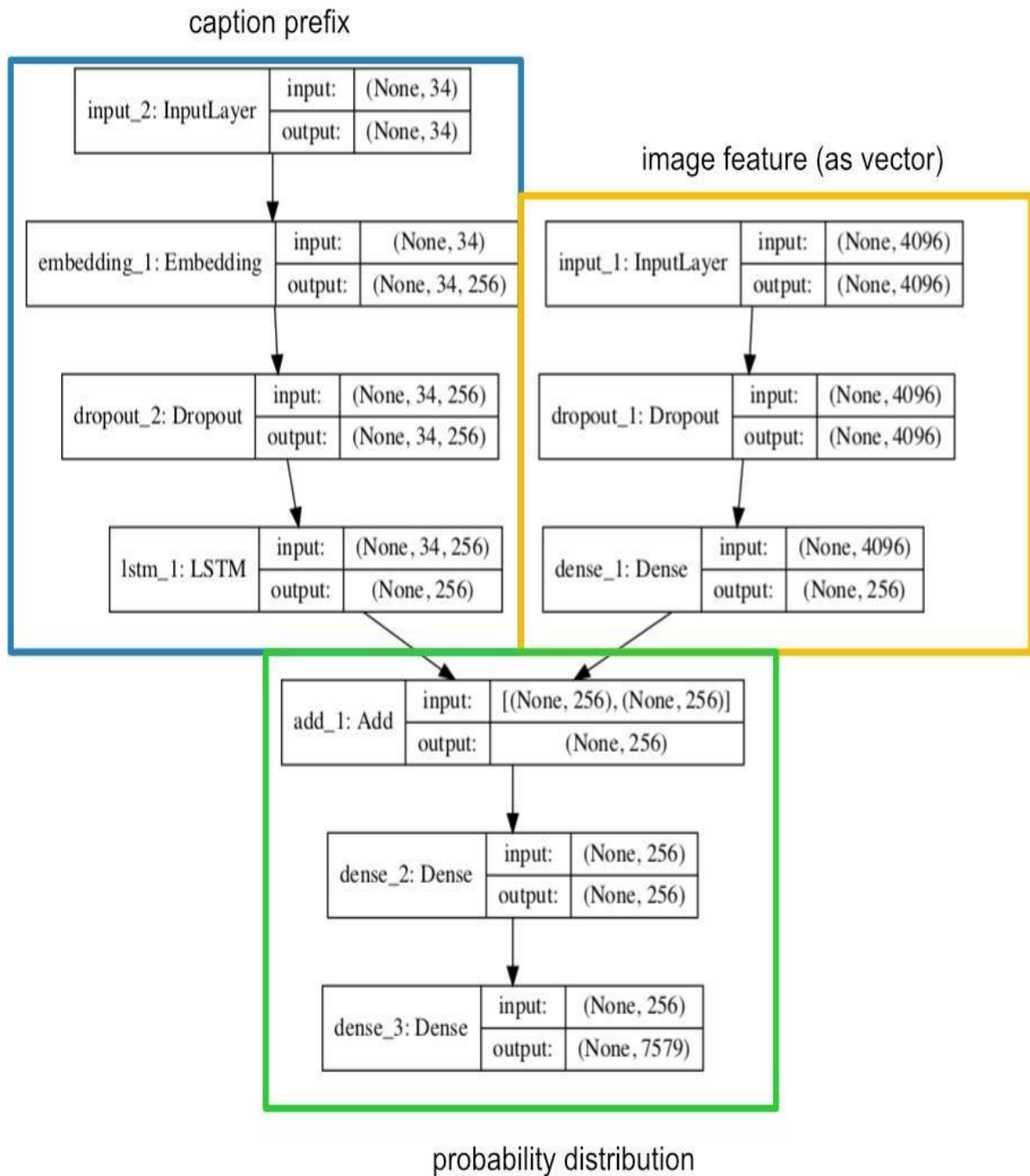


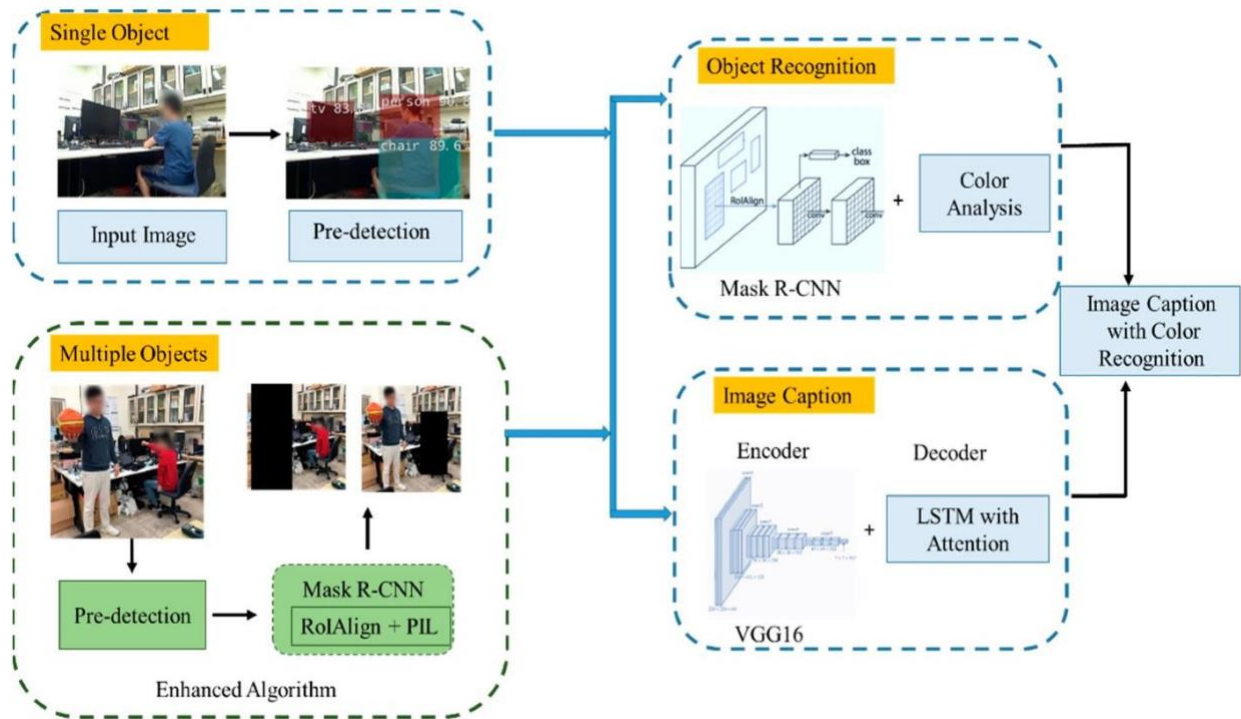
Figure 4.3.7 DFD Diagram Level 2

5. Data Dictionary

5.1 Probability Distribution






5.2 Description Relationship of Data



5.3 Performance of Dataset Flickr8k-SAUI

Dataset	Model	BLEU1	BLEU4	ROUGE	METEOR	CIDER
Flickr8k	Baseline	35.8	4.5	32.0	12.1	6.6
	Attention	62.6	18.2	48.3	18.4	38.5
	Attention + Beam	67.6	24.6	52.4	21.4	60.1
	Reference	67	21.3	-	20.3	-
Flickr30k	Baseline	36.8	5.3	33.3	13.4	7.4
	Attention	61.2	16.6	48.3	16.8	25.6
	Attention + Beam	68.4	23.4	50.0	19.3	44.3
	Reference	66.7	19.1	-	18.4	-
MSCOCO	Baseline	41.4	11.0	43.7	19.6	35.3
	Attention	70.8	24.0	52.9	22.7	48.3
	Attention + Beam	72.7	30.3	54.4	24.1	87.6
	Reference	70.7	24.3	-	23.9	-

5.4 Representation of Image Data

Input Image			
Ground Truth	<ol style="list-style-type: none"> 1. The dogs are in the snow in front of a fence 2. The dogs play on the snow 3. Two brown dogs playfully fight in the snow 4. Two brown dogs playfully fight in the snow 5. Two dogs playing in the snow 	<ol style="list-style-type: none"> 1. A brown and white dog swimming towards some in the pool 2. A dog in a swimming pool swims toward somebody we cannot see 3. A dog swims in a pool near a person 4. Small dog is paddling through the water in a pool 5. The small brown and white dog is in the pool 	<ol style="list-style-type: none"> 1. A man and a woman in festive costumes dancing 2. A man and a woman with feathers on her head dance 3. A man and a woman wearing decorative costumes and dancing in a crowd of onlookers 4. One performer wearing a feathered headdress dancing with another performer in the streets 5. Two people are dancing with drums on the right and a crowd behind them
BLSTM+PMFO	There are two brown dogs running in the snow	A black and white dog swimming in a pool	A group of people playing in a parade
BLSTM+MFO[40]	A dog running through the snow	A dog in a black black blue dog in a pool	A group of people in a red red shirt in a red red and red and red and red shirt
BLSTM[39]	A brown dog is running through the snow	A white dog is jumping into a pool	'A group of people are standing in a parade
LSTM[38]	A brown and brown dog is running in the snow	A black dog is playing in the water	A group of people are walking on a red shirt
c-RNN[2]	A dog is running in the snow	A dog in a white dog in a water	A man in a red shirt is in a red shirt and a red shirt and a red shirt and a red shirt and a red shirt and a red shirt

6. System Testing

6.1 Test Plan

- A test plan documents the strategy that will be used to verify and ensure that a product or system meets its design specifications and other requirements. A test plan is usually prepared by or with significant input from test engineers.
- Software testing has a dual function; it is used to establish the presence of defects in program and it is used to help/judge whether or not the program is usable in practice. The software testing is used for validation and verification, which ensures that software, conforms to its specification and meets the need of the software customer. It is used to check errors and validations. We have started with first testing for requirements gathering and designing. Design errors are costly to repair once the system has started operating. So it is very important to fix the designing errors first. Then we move onto implementation error.

6.2 Testing Strategy

- System testing is an expensive but critical process that can take as much as 50% of the budget of the program development. So testing is most important part in the system. The common view of testing held by users is that it is performed to prove that there are no errors in the program but the most useful approach is the explicit intention of finding the error i.e. making the program fail. Thus any system that is developed is exhaustively tested before it is finally implemented.

1) Code Testing: Code testing is done to examine the logic of the problem. Analyst tests every path through the program. A path is specific combination of condition that is handled by the program. Code testing seems to be a method for testing software. In this project the expert programmer tests the code.

2) System Testing: Entire system is tested as per the requirements. Black-box type testing that is based on overall requirements specifications, covers all combined parts of a system.


3) Storage Testing: I specify a capacity for the system when it is designed and constructed. Capacities are measured in terms of the number of records that a disk will handle or file can contain.


4) Alpha Testing: The above different testing process described takes place in different stages of development as per the requirements and needs. But a final testing is always made after a full finished product that is before it released to end users and this is called as alpha testing. The alpha testing involves both the white box testing and black box testing thus making alpha testing to be carried out in two phases.

5) Beta Testing: This process of testing is carried out to have more validity of the software developed. This takes place after the alpha testing. After the alpha phase also, generally, the release is not made fully available to all end users. The server is released to a set of people and feedback is taken from them to ensure the validity of the server. So here normally the testing is being done by group of end users and therefore this beta testing phase covers black box testing or functionality testing only.

6.3 Testing Methods

- Software testing methods are traditionally divided into black box testing and white box testing. These two approaches are used to describe the point of view that a test engineer takes when designing test cases.

 **Black Box Testing:** It takes an external perspective of the test object to derive test cases. These tests can be functional or non-functional, though usually functional. The test designer selects valid and invalid input and determines the correct output. There is no knowledge of the test object's internal structure. Black Box Testing is testing without knowledge of the internal workings of the item being tested. For example, when black box testing is applied to software engineering, the tester would only know the "legal" inputs and what the expected outputs should be, but not how the program actually arrives at those outputs. It is because of this that black box testing can be considered testing with respect to the specifications, no other knowledge of the program is necessary. For this reason, the tester and the programmer can be independent of one another, avoiding programmer bias toward his own work. Due to the nature of black box testing, the test planning can begin as soon as the specifications are written.

 **White Box Testing:** The opposite of black box testing would be glass box testing, where test data are derived from direct examination of the code to be tested. For glass box testing, the test cases cannot be determined until the code has actually been written. Both of these testing techniques have advantages and disadvantages, but when combined, they help to ensure thorough testing of the product. Software testing approaches that examine the program structure and derive test data from the program logic. Structural testing is sometimes referred to as clear-box testing since white boxes are considered opaque and do not really permit visibility into the code.

Unit Testing: Unit testing is the process of test verification on the smallest unit of software design software module. It is used to uncover errors within the boundary of the module.

Integrating Testing: The Integration Testing is associated with Unit testing. Here the modules of unit testing are put together and checked whether they work properly, when they are integrated, or not. It contains different strategies for same. They are as follows. I performed Integrating Testing 34 by merging all the Modules and Testing as whole Application.

Validation Testing: In integration testing, the software is assembled as a package. Validation Testing is completely associated with requirement satisfaction of customers. This testing checks whether all functional requirements of customer are satisfied or not.

7. Result, Discussion and Conclusion

7.1 Result

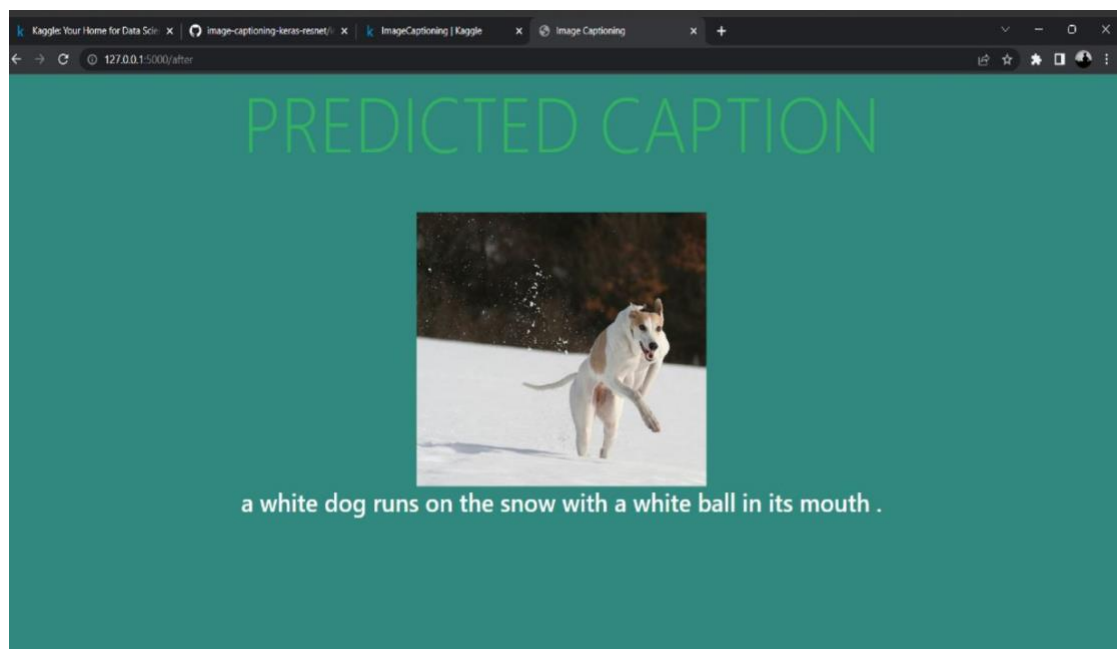
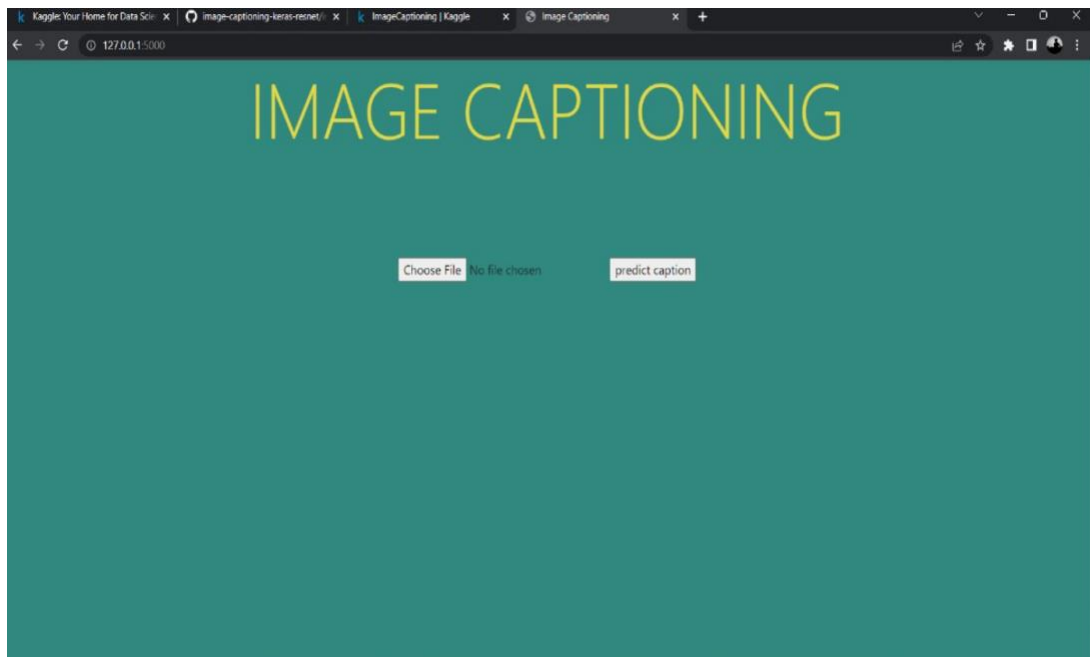
- ❖ Data is encoded to numbers via a dictionary of words. After the data is consumed by the pipeline, the output will also be in the encoded format which needs to be reversed back into English words in order to make sense to human.
- ❖ The other important thing is the output from RNN network is a series of likelihoods of words (likelihoods). Picking highest likelihood word at each decode step in RNN tend to yield a sub-optimal result.
- ❖ Instead, we have implemented Beam Search introduced in Section 2.7 which is a popular solution for finding optimal path for decoding natural language sentences. Some generated captions on test images are presented below.
- ❖ Apparently, the network generated captions are not all perfect, some of which miss important information in the image and others have misidentified visual features.



Figure 5. An example of an incorrectly generated caption

- ❖ For example, the top left image in Figure 5 is "A little girl in a white dress is playing a hunki". The ladder in the image is recognized by the CNN network but the RNN failed to generate the word "ladder". The use of hunki suggests that training images has few number of images with ladders or the word "ladder" is very rare in training captions.

❖ Output –



7.2 Discussion

7.2.1 Benefits

- ❖ If we are able to perform automatic image annotations, then this can have both practical and theoretical benefits. In the current social development process, the most important thing is the massive data that exists on the Internet. Most of these data are different from traditional data, and media data occupies a large proportion.
- ❖ The machine will be able to better assist human beings to use these media data to do more things.

7.2.2 Intelligent monitoring

- ❖ Intelligent monitoring enables the machine to identify and determine the behaviour of people or vehicles in the captured scene and generate alarms under appropriate conditions to prompt the user to react to emergencies and prevent unnecessary accidents.

7.2.3 Image and Video annotation

- ❖ When a user uploads a picture, the picture needs to be illustrated and annotated which can be easily found by the other users. The traditional method is to retrieve the most similar picture in the database for annotation, but this method often results in incorrectly annotated images.

7.2.4 Inconsistent objects during training and testing

- ❖ From the current study, during the training process, the input to the network at each time step is a real word vector or a mixture of real words and images, and the output of the network is the predicted word.

7.2.5 Cross-language text description of images

- ❖ The existing image captioning method based on deep learning or machine learning requires a lot of marked training samples. In practical applications, it is required that a text description of a plurality of languages can be provided for the image to meet the needs of different native language users.

7.3 Conclusion

- ❖ Image captioning is a very exciting exercise and raises tough competition among researchers. There are more and more scientists who are deciding to explore this study field, so the amount of information is constantly increasing. It was noticed that the results are usually compared with quite old articles, although there are dozens of new ones, with even higher results and new ideas for improvements.
- ❖ The comparison with older articles gives a misunderstanding of the real view of result increase—usually there have been much higher results already achieved, however not included in the paper. New ideas can also very easily become lost if they are not looked for carefully.
- ❖ In order to prevent good ideas been lost and to increase fair competition among the new models created, this systematic literature review summarizes all the newest articles and their results in one place.
- ❖ Moreover, it is still not clear if MS COCO and Flickr30k datasets are enough for model evaluation and if they serve sufficiently well when having in mind diverse environments. The amount of data will never stop increasing and new information will keep appearing, so future studies should consider if static models are good enough when thinking of long term application or if lifelong learning should be increasingly thought of.

8. References

- 1) Jonathan Hui Blog. <https://jhui.github.io/2017/03/15/Soft-and-hard-attention/>
- 2) Flick, Carlos. "ROUGE: A Package for Automatic Evaluation of summaries." The Workshop on Text Summarization Branches Out 2004:10. (2014).
- 3) You Tube : <https://www.youtube.com/>
- 4) Kaggle : <https://www.kaggle.com/>
- 5) Github : github.com