**IMAGE CAPTION GENERATOR**

***Report Submitted in partial fulfillment of requirements* for *the***

***B.Tech****.* ***degree in Computer Science and Engineering***

BY

**NAME OF STUDENT ROLL NUMBER**

Paranjay Singh 2020UCO1678

Piyush Singh 2020UCO1700

Satender 2020UCO1655

**Under the Supervision**

**of**

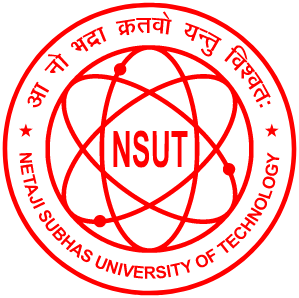
**(Dr Anand Gupta)**



**Department of Computer Science and Engineering**

**Netaji Subhas University of Technology (NSUT)**

**CERTIFICATE**



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

This is to certify that the work embodied in project thesis titled, **"Image Caption**

**Generator”** by *Paranjay Singh(2020UCO1678)*, *Piyush Singh(2020UCO1700)*

and *Satender (2020UCO1655)* is the bonafide work of the group submitted to

Netaji Subhas University of Technology (NSUT) for consideration in 7th

Semester B.Tech. Project Evaluation.

The original Research work was carried out by the team under my/our guidance

and supervision in the academic year 2023-2024. This work has not been

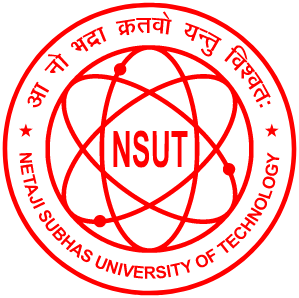
submitted for any other diploma or degree of any university. On the basis of

declaration made by the group, we recommend the project report for evaluation.

DR. Anand Gupta

Department of Computer science & engineering

**CANDIDATE(S) DECLARATION**



**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

I/We, *Paranjay Singh (2020UCO1678), Piyush Singh(2020UCO1700)* and *Satendar* *(2020UCO1655*) of B. Tech. Department of Computer Science & Engineering, hereby declare that the Project-Thesis titled “**Image Caption Generator**” which is submitted by me/us to the Department of Computer Science & Engineering, Netaji Subhas University of Technology (NSUT) Dwarka, New Delhi in partial fulfillment of the requirement for the award of the degree of Bachelor of Technology is original and not copied from the source without proper citation. The manuscript has been subjected to plagiarism checks by Turnitin software. This work has not previously formed the basis for the award of any degree.

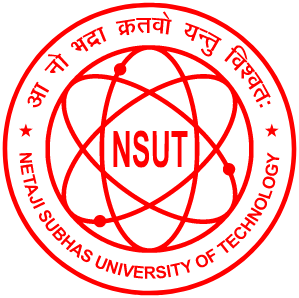
Place :

Date :

Paranjay Singh Piyush Singh Satender

(2020UCO1678) (2020UCO1700) (2020UCO1655)

**CERTIFICATION OF DECLARATION**



**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

This is to certify that the Project-Thesis titled "Image Caption Generator" which is being submitted by Paranjay Singh (2020UCO1678), Piyush Singh (2020UCO1700) and Satender (2020UCO1655) to the Department of Computer Science & Engineering, Netaji Subhas University of Technology (NSUT) Dwarka, New Delhi in partial fulfillment of the requirement for the award of the degree of Bachelor of Technology, is a record of the thesis work carried out by the students under my supervision and guidance. The content of this thesis, in full or in parts, has not been submitted for any other degree or diploma.

Place:

Date :

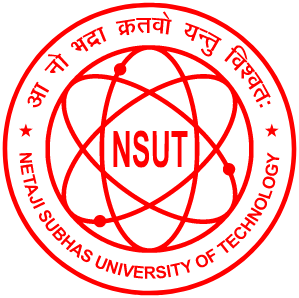
**Dr.Anand Gupta**

(Assistant Professor)

Department of Computer science and engineering

Netaji Subhas University of Technology

**ACKNOWLEDGEMENT**



**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

We would like to express our gratitude and appreciation to all those who make it possible to complete this project. Special thanks to our project supervisor **Dr. Anand Gupta** whose help, stimulating suggestions and encouragement helped us in writing this report. We also sincerely thank our colleagues for the time spent proofreading and correcting our mistakes.

We would also like to acknowledge with much appreciation the crucial role of the staff in the Computer Science & Engineering department, who gave us permission to use the lab and the systems and giving permission to use all necessary things related to project.

Paranjay Singh Piyush Singh Satender

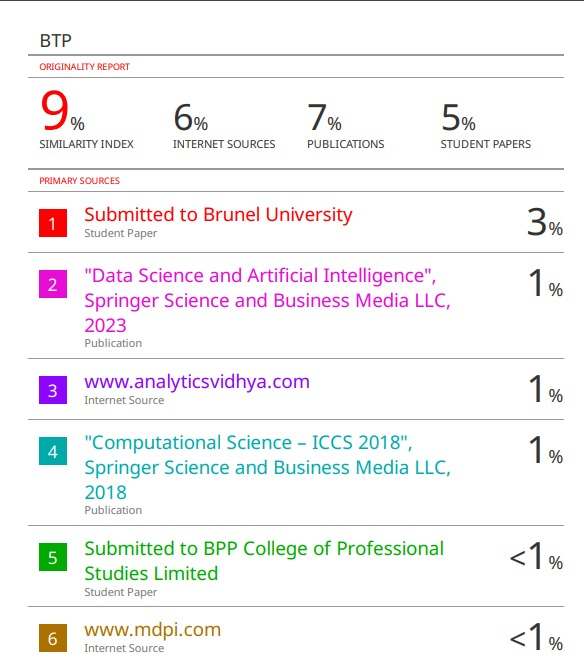
(2020UCO1678) (2020UCO1700) (2020UCO1655)

**PLAGARISM REPORT**

We have conducted a thorough analysis of our Project Thesis, Image Super Resolution,

using Turnitin to ensure the Originality and Integrity of the Content. Here is the Report

of Plagiarism for your reference:



**ABSTRACT**

In this project, we use CNN and LSTM to identify the caption of the image. As the deep learning techniques are growing, huge datasets and computer power are helpful to build models that can generate captions for an image. This is what we are going to implement in this Python based project where we will use deep learning techniques like CNN and RNN. Image caption generator is a process which involves natural language processing and computer vision concepts to recognize the context of an image and present it in English. In this survey paper, we carefully follow some of the core concepts of image captioning and its common approaches. We discuss Keras library, numpy and jupyter notebooks for the making of this project.We also discuss about flickr\_8K dataset and CNN used for image classification.

**KEYWORDS**: generate captions, deep learning techniques, concepts of image captioning.

**Table of Contents**

[**CERTIFICATE**](about:blank) **2**

[**CANDIDATE(S) DECLARATION**](about:blank) **3**

[**CERTIFICATION OF DECLARATION**](about:blank) **4**

[**ACKNOWLEDGEMENT**](about:blank) **5**

[**ABSTRACT**](about:blank) **6**

**CONTENTS**

[**1. INTRODUCTION**](about:blank) **09**

[1.1 Motivation](about:blank) 09

[1.2 Image Captioning](about:blank) 10

[**2. LITERATURE REVIEW**](about:blank) **12**

[2.1 Image Caption Methods](about:blank) 12

[2.1.1 Template-Based Approaches](about:blank) 12

[2.1.2 Retrieval-Based Approaches](about:blank) 12

[2.1.3 Novel Caption Generation](about:blank) 13

[2.2 Deep Learning Based Image Captioning](about:blank) 13

[2.3 Supervised Learning](about:blank) 14

[2.4 Dense Captioning](about:blank) 14

[2.5 Encoder-Decoder Architecture-Based Image captioning](about:blank) 15

[2.5.1 Encoder-Decoder Architecture-Based Image Captioning](about:blank) 15

[2.5.2 Compositional Architecture-Based Image Captioning](about:blank) 15

[2.6 LSRM VS Others](about:blank) 16

[**3. PROBLEM AND PROPOSED WORK**](about:blank) **18**

[3.1 Physically Identification](about:blank) 18

[3.2 The Vanishing Gradient Problem](about:blank) 18

[3.2.1 How to Overcome Vanishing Gradient](about:blank) 18

[3.3 Proposed Work](about:blank) 19

[3.4 Convolutional Neural Network](about:blank) 19

[3.5 Long Short Term Memory(LSTM)](about:blank) 20

[**4 SYSTEM DESIGN AND IMPLEMENTATION**](about:blank) **22**

[4.1 Flicker 8k Dataset](about:blank) 22

[4.2 Image Data Preparation](about:blank) 22

[4.3 Caption Data Preparation](about:blank) 23

[4.4 High Level Design](about:blank) 24

[4.5 Workflow](about:blank) 25

[4.6 Implementation](about:blank) 26

[4.6.1 Pre-requisites](about:blank) 26

[4.6.2 Project File Structure](about:blank) 26

[4.6.3 Building The Python Based Project](about:blank) 27

[4.7 Extracting The Feature Vector From All Images](about:blank) 28

[4.8 Tokenizing The Vocabulary](about:blank) 28

[4.9 Create Data Generator](about:blank) 29

[4.10 Training](about:blank) 29

[4.11 Testing](about:blank) 31

[**5. CONCLUSION**](about:blank) **32**

[5.1 Conclusion](about:blank) 32

[5.2 Limitations](about:blank) 32

[5.3 Future Work](about:blank) 33

[**7. REFERENCES**](about:blank) **34**

**LIST OF FIGURE**

**FIGURE No. Page No**

Figure 1…………………………………….……………………..12

Figure 2………………………………………..…………….. …..15

Figure 3………………………………………………………..….16

Figure 4……………………………………………………..…….20

Figure 5……………………………………………………..…….21

Figure 6…………………………………………………..……….22

Figure 7 …………………………………………………..………24

Figure 8. .………………………. ……………………….……… 25

Figure 9………………………….……...……………….………..26

Figure 10…………………………….………………..………. ....28

Figure 11………………………………………………………….28

Figure 12………………………………………………………….31

Figure 13………………………………………………………….31

Figure 14………………………………………………………….32

Figure 15………………………………………………………….32

**CHAPTER -1**

**INTRODUCTION**

We come across a lot of images every day from different sources like the internet, advertisements, news articles, and document diagrams. The images in these sources are left up to the viewers' interpretation. The majority of photos lack a description, but the Without their extensive captions, humans can understand them quite well. But the device must if people require automatically generated image captions, they should interpret some kind of image caption.Image captioning is crucial for a variety of factors. Every image on the internet can have a caption. result in image searches and indexing that are quicker and more descriptively accurate.

Ever since scientists began studying object recognition in photos, it has become evident that a comprehensive human-like description—rather than just the names of the objects identified—leads to a more favorable impression. Natural language descriptions will always be a difficult problem to solve as long as machines cannot think, speak, or act like humans.

Image captioning has various applications in various fields such as biomedicine, commerce, web searching and military etc. Social media like Instagram , Facebook etc can generate captions automatically from images.

**1.1 Motivation-**

A crucial task in the fields of computer vision and natural language processing is creating captions for images. A machine that can mimic a human's ability to describe an image is a remarkable advancement in artificial intelligence. Capturing the relationships between objects in the image and expressing them in a natural language (like English) is the main challenge of this task.Computer systems have traditionally generated text descriptions for images by using predefined templates. Nevertheless, this method falls short of offering the necessary diversity to produce lexically rich text descriptions. Due to neural networks' increased efficiency, this flaw has been mitigated.Neural networks are widely used in state-of-the-art models to generate captions by using an image as input and forecasting the next lexical unit in the output sentence.

**1.2 Image Captioning-**

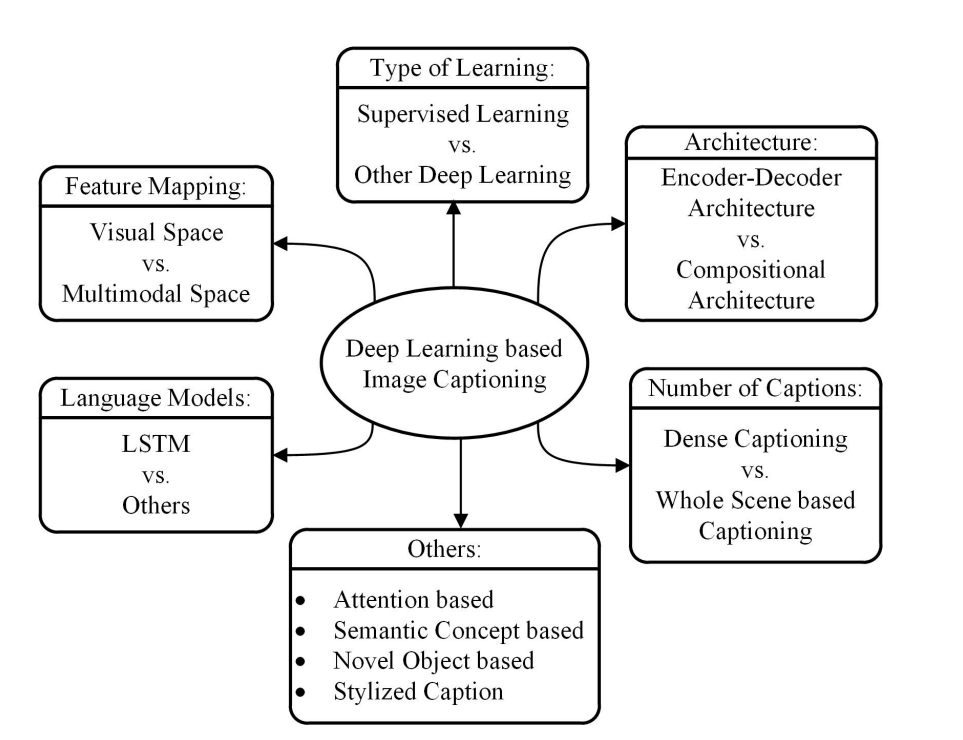
**Process :-**Image Captioning involves generating a textual description of an image, leveraging both Natural Language Processing and Computer Vision. This area of Artificial Intelligence focuses on comprehending images and providing language-based descriptions. The process involves detecting and recognizing objects, understanding scene details, object properties, and their interactions. Generating coherent sentences requires a grasp of both syntactic and semantic aspects of language. Successful image understanding relies on obtaining relevant image features, which can be used for tasks like automatic image indexing. Image indexing plays a crucial role in Content-Based Image Retrieval (CBIR) and finds applications in various fields such as biomedicine, commerce, military, education, digital libraries, and web searching. Social media platforms like Facebook and Twitter utilize image captioning to automatically generate descriptions, encompassing details like location, attire, and activities**.**

**Techniques :-** The techniques used for this purpose can be broadly divided into two categories:

(1) Traditional machine learning based techniques.

(2) Deep machine learning based techniques.

In traditional machine learning, hand crafted features such as Local Binary Patterns (LBP), Scale-Invariant Feature Transform (SIFT), the Histogram of Oriented Gradients (HOG) , and a combination of such features are widely used. In these techniques, features are extracted from input data. They are then passed to a classifier such as Support Vector Machines (SVM) in order to classify an object. Since hand crafted features are task specific, extracting features from a large and diverse set of data is not feasible. Moreover, real world data such as images and video are complex and have different semantic interpretations. On the other hand, in deep machine learning based techniques, features are learned automatically from training data and they can handle a large and diverse set of images and videos. For example, Convolutional Neural Networks (CNN) are widely used for feature learning, and a classifier such as Softmax is used for classification. CNN is generally followed by Recurrent Neural Networks (RNN) or Long Short-Term Memory Networks (LSTM) in order to generate 10 captions.Deep learning algorithms can handle complexities and challenges of image captioning quite well.



**FIGURE 1 Taxonomy of deep learning-based image captioning systems.**

**CHAPTER-2**

**LITERATURE REVIEW**

Image captioning has gained significant attention, particularly within the realm of natural language processing. The demand for context-based natural language descriptions of images has become more pronounced. Although this might have seemed ambitious in the past, recent advancements in neural networks, computer vision, and natural language processing have paved the way for accurately portraying the visually grounded meaning of images. Our approach involves utilizing cutting-edge techniques such as Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN), along with curated datasets containing images and their human-perceived descriptions. Through our alignment model, we showcase compelling results in retrieval experiments conducted on various datasets.

We draw an overall taxonomy in Figure 1 for deep learning-based image captioning methods. We discuss their similarities and dissimilarities by grouping them into visual space vs. multimodal space, dense captioning vs. captions for the whole scene, Supervised learning vs. Other deep learning, Encoder-Decoder architecture vs. Compositional architecture, and one „Others‟ group that contains Attention-Based, Semantic Concept-Based, Stylized captions, and Novel Object-Based captioning. We also create a category named LSTM vs. Others.

**2.1 Image Captioning Methods**

There are different Image Caption generator Techniques some are seldom utilized in current time yet it is important to take an outline of those technologies prior to continuing ahead. We bunch the current picture subtitle generator strategy into three fundamental classifications

**2.1.1 TEMPLATE-BASED APPROACHES**

Template-based approaches have fixed templates with a number of blank slots to generate captions. In these approaches, different objects, attributes, actions are detected first and then the blank spaces in the templates are filled. For example a triplet of scene elements to fill the template slots for generating image captions or extract the phrases related to detected objects, attributes and their relationships for this purpose.

**2.1.2 RETRIEVAL-BASED APPROACHES**

Captions can be retrieved from visual space and multimodal space. In retrieval-based approaches, captions are retrieved from a set of existing captions. Retrieval based methods first find the visually similar images with their captions from the training data set. These captions are called candidate captions. The captions for the query image are selected from these captions pool. These methods produce general and syntactically correct captions. However, they cannot generate image specific and semantically correct captions.

**2.1.3 NOVEL CAPTION GENERATION**

Instead of matching to an existing caption, novel image captions are created by the model using a combination of the image features and a language model. Creating original captions for images resolves the two issues with utilizing pre-existing captions, making it a far more intriguing and practical problem.It is possible to generate new captions from both multimodal and visual spaces. This category's standard procedure is to use a language model to generate image captions from the visual content of the image after first analyzing its visual content. Compared to earlier techniques, these methods are able to produce new captions for every image that are more semantically accurate. The majority of cutting edge methods for creating captions rely on deep machine learning techniques.

**2.2 DEEP LEARNING BASED IMAGE CAPTIONING METHODS**

Deep learning-based image captioning methods use neural networks to automatically generate descriptive captions for images. These methods typically consist of two main components: a convolutional neural network (CNN) for image feature extraction and a recurrent neural network (RNN) for generating captions.

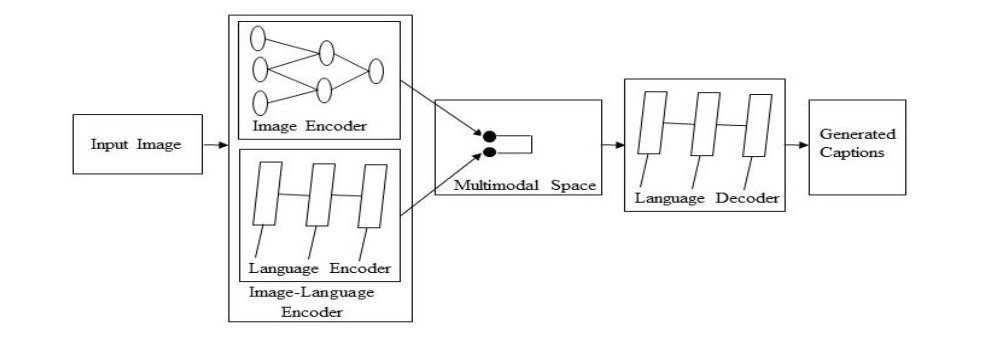
Both visual and multimodal space can be used to generate captions for images using deep learning-based techniques. It makes sense that text captions correspond to the images in captioning datasets. The image features and their corresponding captions are sent to the language decoder separately in the visual space-based methods. On the other hand, a shared multimodal space is discovered from the images and the associated caption-text in a multimodal space case.

**VISUAL SPACE**

Bulk of the image captioning methods use visual space for generating captions. In the visual space-based methods, the image features and the corresponding captions are independently passed to the language decoder.

**MULTIMODAL SPACE**

The architecture of a typical multimodal space-based method contains a language Encoder part, a vision part, a multimodal space part, and a language decoder part. The vision part uses a deep convolutional neural network as a feature extractor to extract the image features. The language encoder part extracts the word features and learns a dense feature embedding for each word. It then forwards the semantic temporal context to the recurrent layers. The multimodal space part maps the image features into a common space with the word features.



**Figure 2**

**2.3 SUPERVISED LEARNING**

Supervised learning is a type of machine learning where the algorithm is trained on a labeled dataset, which means that each input data point is associated with a corresponding output label. The goal of supervised learning is for the algorithm to learn a mapping from inputs to outputs, allowing it to make predictions or classifications on new, unseen data.

Supervised learning-based networks have successfully been used for many years in image classification , object detection and attribute learning . This progress makes researchers interested in using them in automatic image captioning .In this paper, we have identified a large number of supervised learning-based image captioning methods. We classify them into different categories: (i) Encoder-Decoder Architecture, (ii) Compositional Architecture,(iii)Attention Based.

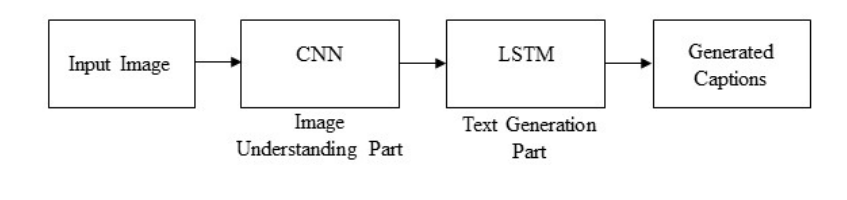
**2.4 DENSE CAPTIONING**

In dense captioning, captions are generated for each region of the scene. Other methods generate captions for the whole scene.The previous image captioning methods can generate only one caption for the whole image. They use different regions of the image to obtain information of various objects. However, these methods do not generate region wise captions.A typical method of this category has the following steps:

(1) Region proposals are generated for the different regions of the given image.

(2) CNN is used to obtain the region-based image features.

(3) The outputs of Step 2 are used by a language model to generate captions for every region. A block diagram of a typical dense captioning method is given in Figure 3.



**Figure 3**

**2.5 ENCODER-DECODER ARCHITECTURE VS. COMPOSITIONAL ARCHITECTURE**

Some methods use just simple vanilla encoder and decoder to generate captions. However, other methods use multiple networks for it.

**2.5.1 ENCODER-DECODER ARCHITECTURE-BASED IMAGE CAPTIONING**

Neural network-based image captioning techniques operate in an end-to-end fashion.

These techniques bear a lot of similarities to neural machine translation based on the encoder-decoder framework. In this network, global image features are taken out of the CNN's hidden activations and fed into an LSTM to produce a word sequence. The general steps of an average method in this category are as follows:

(1) A classical CNN is used to obtain the scene type, to detect the objects and their relationships

(2) The output of Step 1 is used by a language model to convert them into words, combined phrases that produce an image captions.

**2.5.2 COMPOSITIONAL ARCHITECTURE-BASED IMAGE**

**CAPTIONING**

Compositional architecture-based methods composed of several independent functional building blocks: First, a CNN is used to extract the semantic concepts from the image. Then a language model is used to generate a set of candidate captions. In generating the final caption, these candidate captions are re-ranked using a deep multimodal similarity model. A typical method of this category maintains the following steps:

(1) Image features are obtained using a CNN

(2) Visual concepts (e.g. attributes) are obtained from visual features

(3) Multiple captions are generated by language model using the information of Step 1 & Step 2

(4) The generated captions are re-ranked using a deep multimodal similarity model to select high quality image captions.

**2.6 LSTM VS. OTHERS**

Natural language processing (NLP) and computer vision research are intersected by image captioning. Generally speaking, NLP tasks can be designed as sequence to sequence learning. For learning sequence-to-sequence tasks, a number of neural language models have been proposed, including recurrent neural networks (RNNs), skip-gram models, log-bilinear models, and neural probabilistic language model. RNNs are frequently utilized in a variety of sequence learning applications. Traditional RNNs, however, are unable to effectively handle long-term temporal dependencies due to their vanishing and exploding gradient issues.

One kind of RNN with additional special units to the standard units are called LSTM networks. A memory cell that can store information in memory for extended periods of time is used by LSTM units. Sequence to sequence learning tasks have seen a significant increase in the use of LSTM based models in recent years. Although it uses fewer gates to regulate information flow and does not employ separate memory cells, the Gated Recurrent Unit (GRU) network is structurally similar to the Long Short-Term Memory (LSTM).But LSTMs don't take into account a sentence's underlying hierarchical structure. Long-term dependencies via a memory cell necessitate significant storage as well. On the other hand, CNNs process information faster than LSTMs and are able to understand the internal hierarchical structure of the sentences.As a result, convolutional architectures have been applied recently to other sequence-to-sequence tasks, such as machine translation and conditional image generation. Gu proposed an image captioning technique based on CNN language models, which was motivated by the above-mentioned success of CNNs in sequence learning tasks. For statistical language modeling, this approach makes use of a language-CNN. But employing a language-CNN alone is insufficient to capture the dynamic temporal behavior of the language model. In order to accurately represent the temporal dependencies, it combines a recurrent network with the language CNN.

**CHAPTER -3**

**PROBLEM AND PROPOSED WORK**

**3.1 PROBLEM IDENTIFICATION**

Despite the successes of many systems based on the Recurrent Neural Networks (RNN) many issues remain to be addressed. Among those issues the following two are prominent for most systems.

1. The Vanishing Gradient Problem.

2.Training an RNN is a very difficult task.

Recurrent neural networks are deep learning algorithms that are used to solve a wide range of challenging computer tasks, including speech recognition and object classification. Recurrent neural networks (RNNs) are built to process a series of consecutive events, with each event's comprehension predicated on data from earlier events.

Ideally, we would prefer to have the deepest RNNs so they could have a longer memory period and better capabilities. These could be applied for many real-world use-cases such as stock prediction and enhanced speech detection. However, while they sound promising, RNNs are rarely used for real-world scenarios because of the vanishing gradient problem.

**3.2 THE VANISHING GRADIENT PROBLEM**

One of the biggest obstacles to RNN performance is the problem of vanishing gradient. In actuality, RNNs' long-term memory is constrained by their architecture, which can only retain a small number of sequences at once. As a result, RNNs' memory is only effective for shorter sequences and shorter times.

Vanishing Gradient problem arises while training an Artificial Neural Network. This mainly occurs when the network parameters and hyperparameters are not properly set. The vanishing gradient problem restricts the memory capabilities of traditional RNNs—adding too many time-steps increases the chance of facing a gradient problem and losing information when you use backpropagation.

**3.2.1 How to overcome vanishing gradient**

1.Replace sigmoid or tanh activations with activation functions that do not saturate as easily, such as Rectified Linear Unit (ReLU) or variants like Leaky ReLU.

2.Limit the magnitude of gradients during training to prevent exploding gradients. This can be particularly useful in recurrent neural networks.

3.Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) cells are designed to alleviate the vanishing gradient problem in recurrent networks by introducing gating mechanisms

**3.3 PROPOSED WORK**

The main aim of this project is to get a little bit of knowledge of deep learning techniques. We use two techniques mainly CNN and LSTM for image classification.

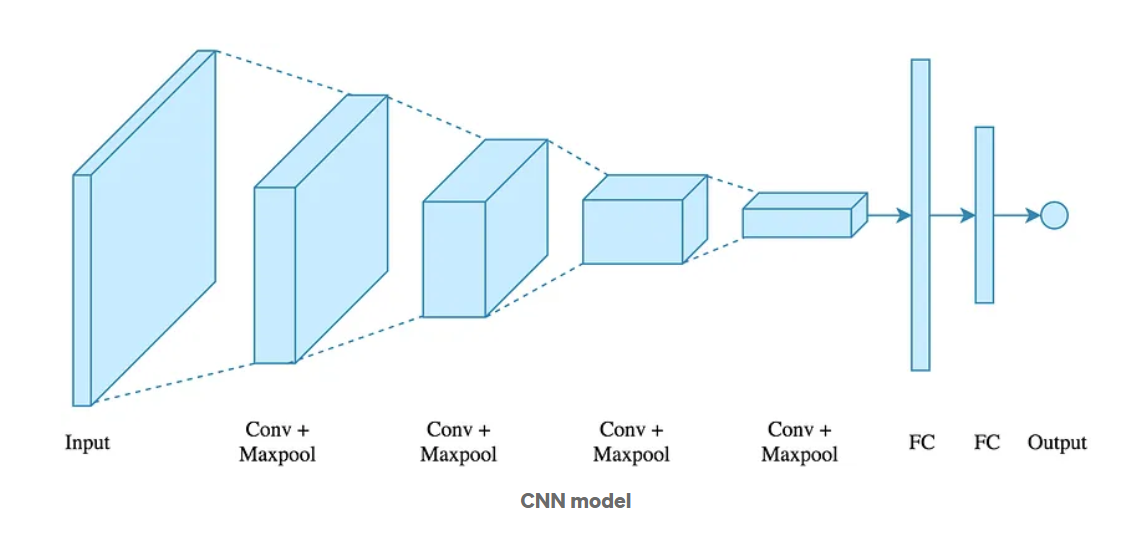
So, to make our image caption generator model, we will be merging these architectures. It is also called a CNN-RNN model.

* CNN is used for extracting features from the image,to get a liner vector of features. We will use the CNN model name VGG-16
* LSTM will use the information from CNN to help generate a description of the image.

**3.4 CONVOLUTIONAL NEURAL NETWORK**

A Convolutional Neural Network (ConvNet/CNN) is a Deep Learning algorithm which can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image and be able to differentiate one from the other.The pre-processing required in a ConvNet is much lower as compared to other classification algorithms

Convolutional Neural networks are specialized deep neural networks which can process the data that has input shape like a 2D matrix. Images are easily represented as a 2D matrix and CNN is very useful in working with images.

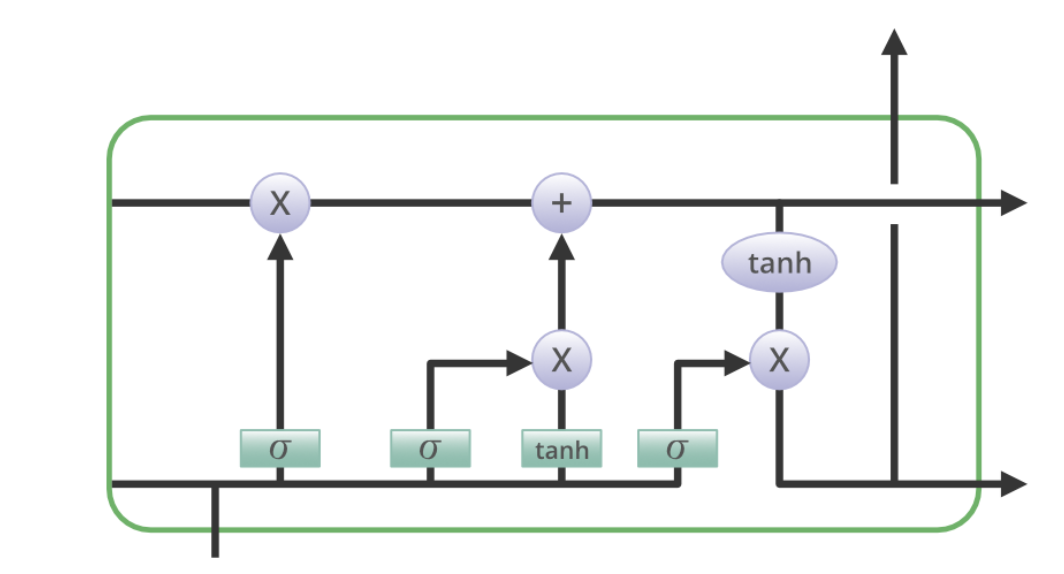


**Figure 4. CNN Model**

It scans images from left to right and top to bottom to pull out important features from the image and combines the features to classify images. It can handle the images that have been translated, rotated, scaled and changes in perspective.

**3.5 LONG SHORT TERM MEMORY(LSTM)**

Long short term memory, or LSTM for short, is a kind of recurrent neural network (RNN) that works well for sequence prediction issues. We know what word will come next because of the previous text. It has outperformed the conventional RNN in terms of effectiveness by resolving the short-term memory issue. During the input processing process, LSTM can process pertinent data while discarding irrelevant data using a forget gate.In contrast to conventional RNNs, LSTMs are made to solve the vanishing gradient issue and enable them to store information for extended periods of time. Because LSTMs are able to sustain a constant error, they can backpropagation through time and layers and continue learning over multiple time-steps.

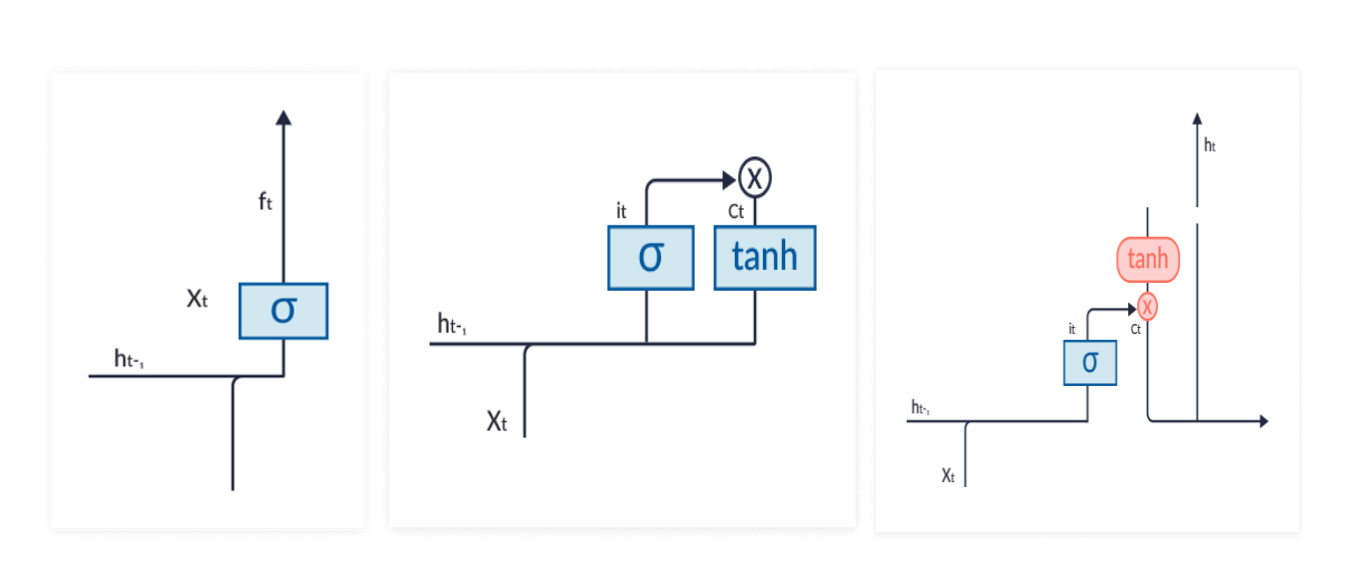


**Figure 5 LSTM model**

The first section determines whether the data from the preceding timestamp should be stored in memory or if it is unimportant and can be ignored. The cell attempts to learn new information from the input to this cell in the second section. Finally, the cell transfers the updated data from the current timestamp to the subsequent timestamp in the third section. A single-time step is this single LSTM cycle.

The terms "gates" refer to these three components of an LSTM unit. They regulate the information that enters and exits the memory cell, also known as the LSM cell. The output gate is the final gate; the forget gate is the first; the input gate is the second; and so on. In a conventional feedforward neural network, an LSTM unit made up of these three gates and a memory cell, also known as an LSTM cell, can be compared to a layer of neurons, with each neuron having a current state and a hidden layer.

The chain-like architecture of LSTM allows it to contain information for longer time periods, solving challenging tasks that traditional RNNs struggle to or simply cannot solve.



**Figure 6 Forget,Input,Output gate**

**Forget gate**-removes information that is no longer necessary for the completion of the task. This step is essential to optimizing the performance of the network.

**Input gate**-responsible for adding information to the cells.

**Output gate**-selects and outputs necessary information.

Textual descriptions of images are generated using this architecture. Using a CNN that has been pre-trained on a difficult image classification task and then modified to serve as a feature extractor for the caption generating problem is crucial.

**CHAPTER-4**

**SYSTEM DESIGN AND IMPLEMENTATION**

This project needs a dataset that includes both photographs and captions. The image captioning model should be able to be trained using the labeled dataset.

**4.1 FLICKR 8K DATASET**

A benchmark for image-to-sentence descriptions that is accessible to the general public is the Flickr 8k dataset. This collection consists of 8091 images, each with five captions. These images were captured from various groups on the Flickr platform. Every caption provides a thorough explanation of the things and happenings depicted in the picture. The collection is more generic since it doesn't include images of famous people or places, but instead depicts a broad variety of occasions and environments. The dataset is perfect for this project because of the following features:

1-When many captions are mapped to a single image, the model becomes more general and

avoids overfitting.

2-Using a variety of training images allows the image captioning model to cope with a

variety of image types, making the model more robust.

**4.2 IMAGE DATA PREPARATION**

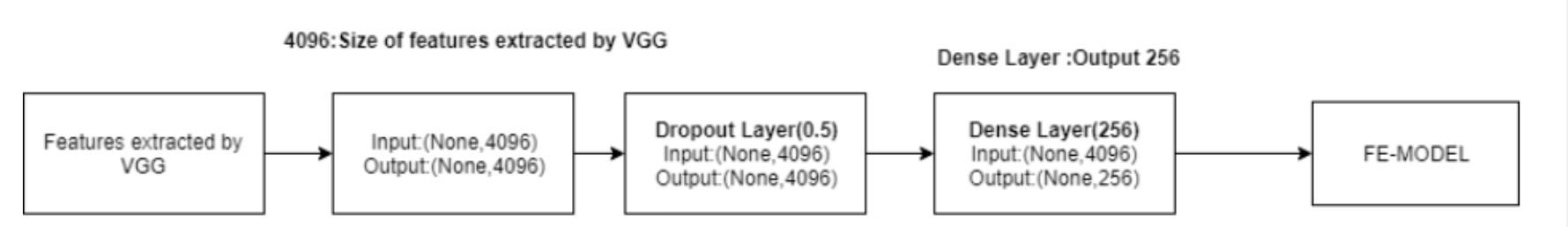
In order to train a deep learning model, the image needs to be transformed into appropriate features. In order to train any image in a deep learning model, feature extraction is a prerequisite.

Convolutional Neural Network (CNN) with Visual Geometry Group (VGG-16) model is used to extract the features. In addition, this model was successful in classifying the images into one of the 1000 classes provided in the ImageNet Large Scale Visual Recognition Challenge,because image identification is necessary for image captioning, this model is perfect for use in this project.

With 16 weight layers in the network, VGG-16 offers superior feature extraction from images due to its deeper number of layers. The VGG-16 network's architecture is straightforward because it employs 3\*3 convolutional layers, and the max pooling layer is used in between to minimize the image's volume size.

The last layer of the image which predicts the classification is removed and the internal representation of the image just before classification is returned as a feature. The dimension of the input image should be 224\*224 and this model extracts features of the image and returns a 1- dimensional 4096 element vector.

To lessen overfitting, a dropout layer with a value of 0.5 is added to the model. The range of 0.5 to 0.8 represents an ideal value, signifying the likelihood of the layer's outputs being eliminated. After the dropout layer, a dense layer is added, which essentially applies a biased activation function to the input kernel. Rectified Linear Units, or "ReLU," is the activation function that is employed, and 256 is the output space size that is specified.



**FIGURE 7**

**4.3 CAPTION DATA PREPARATION**

Flickr8k dataset contains multiple descriptions for a single image. In the data preparation phase, each image id is taken as a key and its corresponding captions are stored as values in a dictionary.

The raw text has to be transformed into a format that can be used by machine learning or deep learning models in order for the text dataset to function. Before using the text for the project, the following cleaning steps are completed:

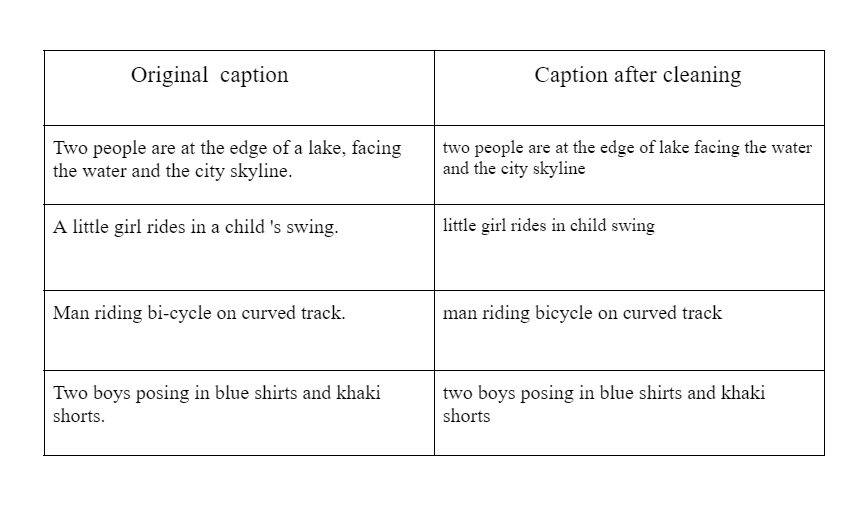
• Eliminating punctuation.

• Eliminating digits.

• Elimination of words that are one length.

• Lowercase characters are converted from uppercase.

Stop words are left in the text data because eliminating them would make it more difficult to create the grammatically correct caption that this project requires. Table 1 displays caption samples following data cleansing.



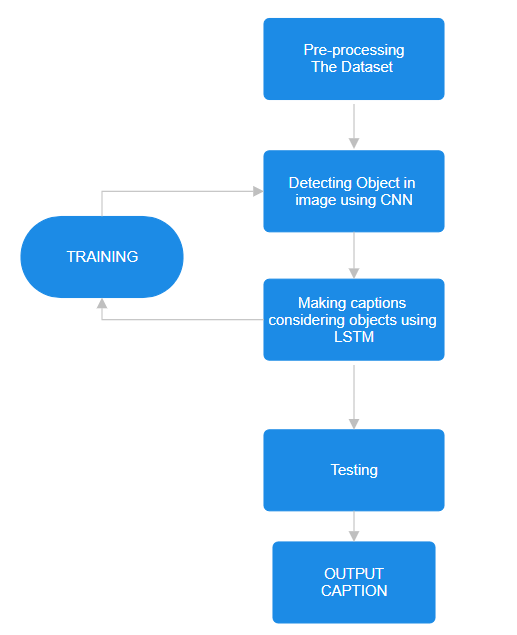
**TABLE 1**

**4.4 HIGH LEVEL DESIGN**

****

**FIGURE 8**

**4.5 WORKFLOW**

****

**FIGURE 9**

Workflow consist of following steps in following order-

1.Data Collection

2.Preprocessing

3.Feature Extraction

4.Text Processing

5.Model architecture & training

6.Evaluation

**4.6 IMPLEMENTATION**

**4.6.1 Pre-requisites**

This project requires good knowledge of Deep learning, Python, working on Jupyter notebooks,

Keras library, Numpy, and Natural language processing

Make sure you install all the necessary libraries and supported by the system:

**-**pip install

-tensorflow

-keras

-pillow

-numpy

-jupyterlab

**4.6.2 PROJECT FILE STRUCTURE**

Downloaded from dataset:

Flicker8k\_Dataset – Dataset folder which contains 8091 images.

Flickr\_8k\_text – Dataset folder which contains text files and captions of images.

The below files will be created by us while making the project.

Models – It will contain our trained models.

Descriptions.txt – This text file contains all image names and their captions after preprocessing.

Features.p – Pickle object that contains an image and their feature vector extracted from the

VGG-16 (CNN)model.

Tokenizer.p – Contains tokens mapped with an index value.

Model.png – Visual representation of dimensions of our project.

Testing\_caption\_generator.py – Python file for generating a caption of any image.

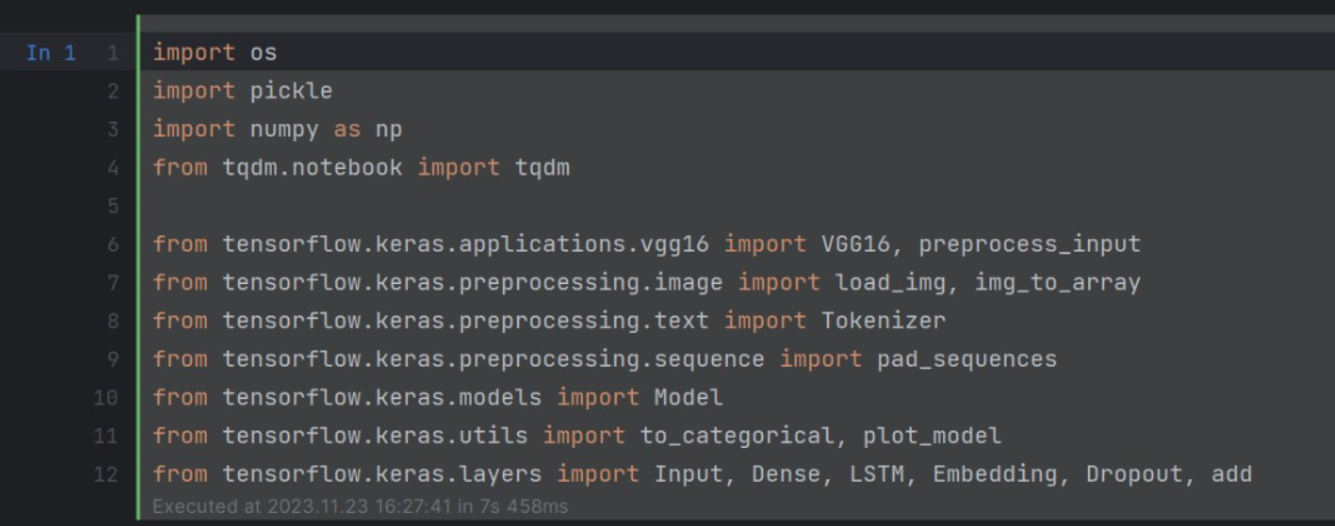
Training\_caption\_generator.ipynb – Jupyter notebook in which we train and build our image caption

Generator.

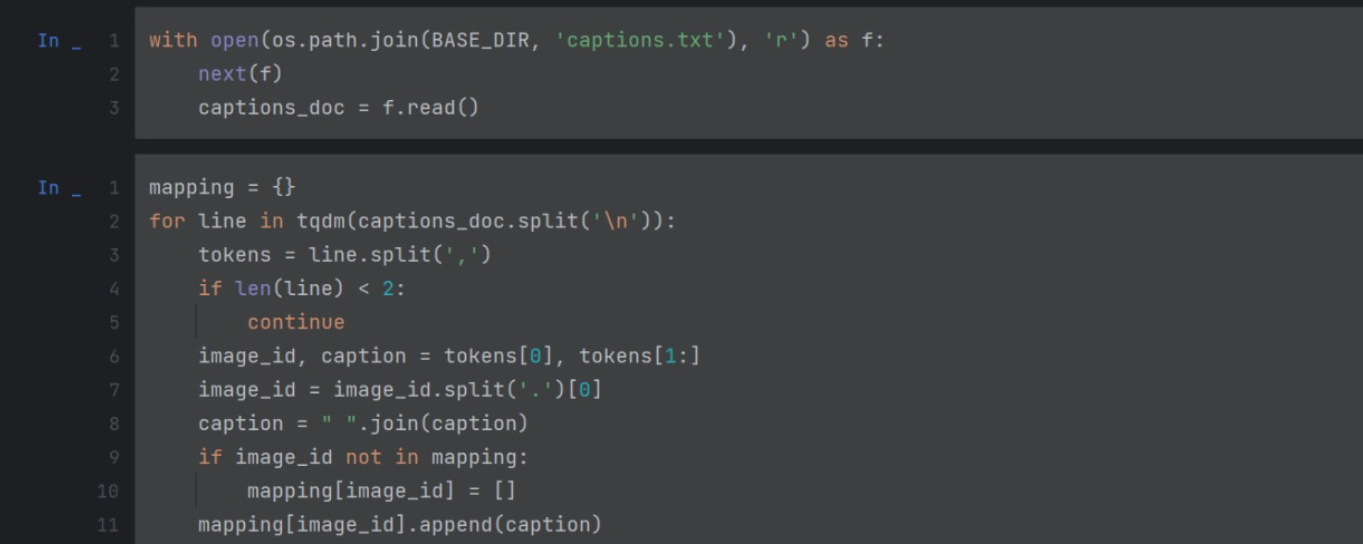
**4.6.3 BUILDING THE PYTHON BASED PROJECT**

To begin, open the console of your project folder and type jupyter lab to initialize the Jupyter Notebook Server. You can run your code in the interactive Python notebook that opens.Call your newly created Python 3 notebook training\_caption\_generator.ipynb.

The main text file which contains all image captions is Flickr8k.token in our main folder



**FIGURE 10 Importing the libraries**

****

**FIGURE 11 Loading caption.text**

**load\_doc** – For loading the document file and reading the contents inside the file into a string.

**all\_img\_captions** – This function will create a descriptions dictionary that maps images with a list of 5 captions.

**cleaning\_text**– This function takes all descriptions and performs data cleaning. This is an important step when we work with textual data, according to our goal, we decide what type of cleaning we want to perform on the text. In our case, we will be removing punctuations, converting all text to lowercase and removing words that contain numbers.So, a caption like “A man riding on a three-wheeled wheelchair” will be transformed into “man riding on three wheeled wheelchair”

**4.7 EXTRACTING THE FEATURE VECTOR FROM ALL IMAGES**

In VGG-16, feature extraction is performed through a series of convolutional and pooling layers. The architecture consists of 16 layers, including 13 convolutional layers and 3 fully connected layers. The convolutional layers use small 3x3 filters with a stride of 1 and zero-padding to maintain spatial resolution.

First the model Accepts the input image, typically of size 224x224 pixels.The convolutional layers use small 3x3 filters with a stride of 1,The activation function used is Rectified Linear Unit (ReLU) to introduce non-linearity.Max-pooling layers with 2x2 filters and a stride of 2 are used to downsample the spatial dimensions.After several convolutional and pooling layers, the spatial dimensions are reduced final fully connected layer produces the output logits for classification

The function extract\_features() will extract features for all images and we will map image names with their respective feature array. Then we will dump the features dictionary into a “features.p” pickle file.

**4.8 TOKENIZING THE VOCABULARY**

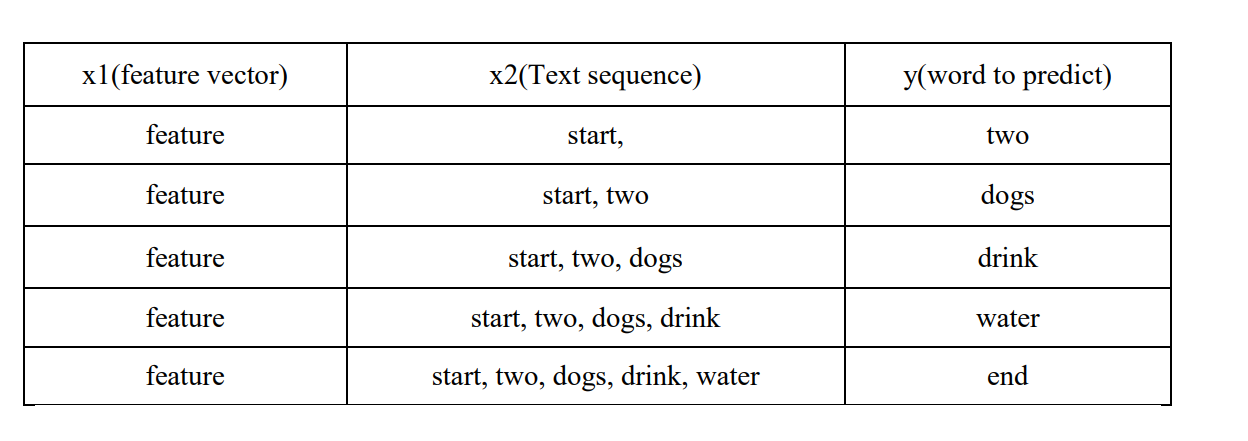
Since English words are not understood by computers, we will have to represent them with numbers. As a result, we will assign a distinct index value to every word in the vocabulary. The tokenizer function in the Keras library is what we'll use to generate tokens from our vocabulary and store them in a pickle file called "tokenizer.p."

Our vocabulary contains 7577 words. We calculate the maximum length of the descriptions. This is important for deciding the model structure parameters. Max\_length of description is 32.

**4.9 Create Data generator**

To initiate our model exploration, let's examine the expected format of both input and output. To transform this task into a supervised learning endeavor, we need to furnish the model with training data comprising 6000 images. Each image is associated with a 2048-dimensional feature vector, and the corresponding caption is represented as numerical values. Due to the substantial data volume for these 6000 images, it's impractical to load all of it into memory simultaneously. Therefore, we'll employ a generator method that will iteratively yield batches of data during the training process. This approach enables us to manage the data efficiently without overwhelming system memory.

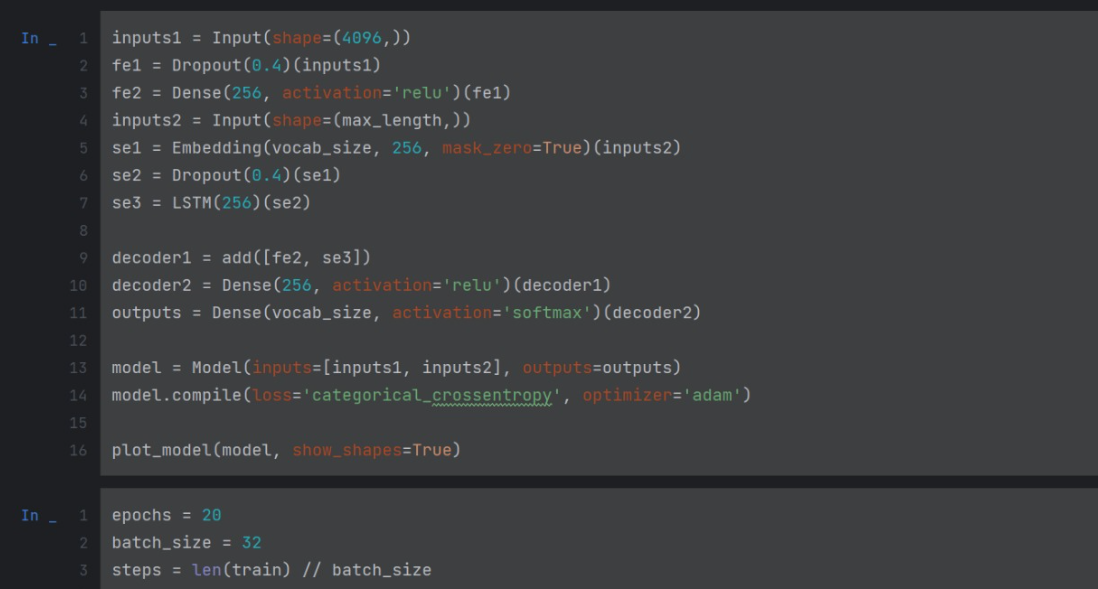
The input to our model is [x1, x2] and the output will be y, where x1 is the 2048 feature vector of that image, x2 is the input text sequence and y is the output text sequence that the model has to predict.

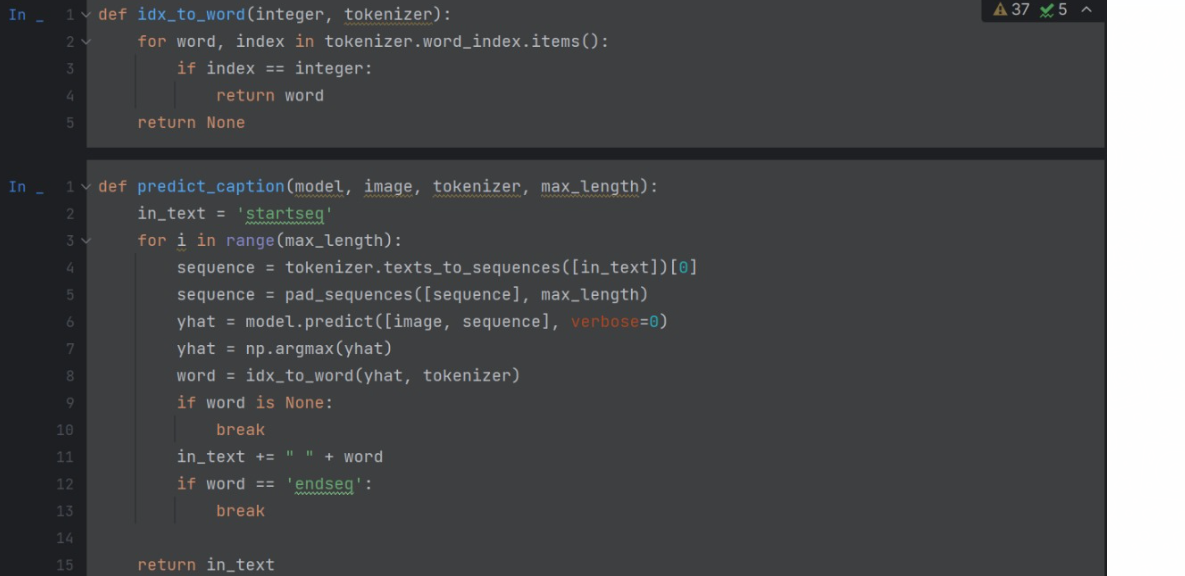


**TABLE 2**

**4.10 Training**

By creating the input and output sequences in batches and fitting them to the model using the model.fit\_generator() function, we will be able to train the model with the 6000 training images. The model is additionally saved to our models folder. Depending on the capabilities of your system, this may take some time.

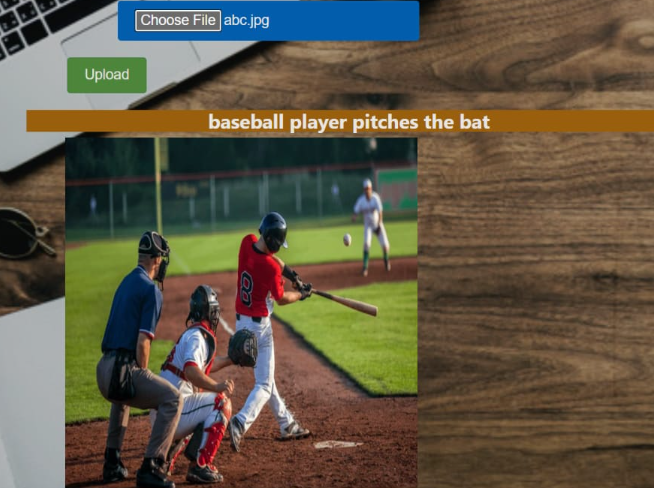




**FIGURE 12 and 13**

**4.11 Testing**

Our image captioning model was accurately able to caption most of the images it was tested on. As the training dataset used was limited, there were a few cases where the captions generated did not well capture the objects in the image, but that improved with the increase in the number of epochs used while training the LSTM model as well as with the increase in the size of the dataset



**Figure 14 & 15**

**CHAPTER-5**

**CONCLUSION**

In this chapter we have thrown some light on the conclusion of our project. We have also underlined the limitation of our methodology. There is a huge possibility in this field, as we have discussed in the future scope section of this chapter.

**5.1 Conclusion**

We have reviewed deep learning approaches for image captioning in this paper. A taxonomy of image captioning methods has been provided, along with a general block diagram of the main categories and a summary of their benefits and drawbacks. We talked about the advantages and disadvantages of various assessment metrics and datasets. The experimental results are also briefly summarized. We gave a brief summary of possible lines of inquiry for this field of study. While there has been significant advancement in deep learning-based image captioning techniques in recent years, a reliable method that can produce captions of high quality for almost all images is still lacking. The field of automatic image captioning will continue to be active for some time due to the introduction of innovative deep learning network architectures.

We utilized the Flickr\_8k dataset, comprising around 8000 images, along with corresponding captions stored in a text file. Despite significant advancements in deep learning-based image captioning methods, achieving a consistently robust system capable of generating high-quality captions for a wide range of images remains a challenge. Ongoing developments in novel deep learning network architectures suggest that automatic image captioning will continue to be a dynamic and evolving research area. As the user base on social media platforms continues to grow, with an increasing number of individuals sharing photos, the significance of image-captioning extends into the future. This project is implemented in a wayto make a meaningful contribution by addressing the evolving needs of users seeking effective and accurate image captioning capabilities.

**5.2 Limitation**

The neural image caption generator gives a useful framework for learning to map from images to human-level image captions. By training on large numbers of image-caption pairs, the model learns to capture relevant semantic information from visual features.

However, with a static image, embedding our caption generator will focus on features of our images useful for image classification and not necessarily features useful for caption generation. To improve the amount of task-relevant information contained in each feature,this can lead to the irrelevant or incorrect caption output,secondly the model is not able to perform well over the completely untrained image that acts as outliers for the model.

**5.3 Future Work**

Future work Image captioning has become an important problem in recent days due to the exponential growth of images in social media and the internet. This report discusses the various research in image retrieval used in the past and it also highlights the various techniques and methodology used in the research.

Decided Future works-

1) Further optimizing the algorithm to increase the accuracy of the generated captions.

2) Implementing voice output to assist blind people

3) Comparing our algorithm with similar algorithms like (inceptionV3 + GRU) and

enhancing our algorithm to make it more efficient

4) Currently, we are using a greedy approach for generating the next word in the sequence

by selecting one with the maximum probability. Beam search instead selects a group of

words with the maximum likelihood and parallel searches through all the sequences

There cannot be completely accurate results as these methodologies do not depend on the context of the image. Hence, a complete research in image retrieval making use of context of the images such as image captioning will facilitate to solve this problem in the future.

**REFERENCES**

[1] CS771 Project Image Captioning by Ankit Gupta , Kartik Hira, Bajaj Dilip.

[2] ”Every Picture Tells a Story: Generating Sentences from Images.” Computer Vision ECCV (2016) by Farhadi, Ali, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hocken-maier, and David Forsyth

[3] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image

[4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2017.

[5] Where to put the Image in an Image Caption Generator by Marc Tanti, Albert Gatt, Kenneth P. Camilleri.

