KIET
GROUP OF INSTITUTIONS
Connecting Life with Learning

A Project Report

on

Image Captioning

submitted as partial fulfillment for the award of

BACHELOR OF TECHNOLOGY

DEGREE

SESSION 2022-23

in

Computer Science and Engineering

By

Tushar Pant (1900290100175)

Sagar Srivastava (1900290100127)

Anshika (1900290100033)

Under the supervision of

Prof. Upendra Mishra

KIET GROUP OF INSTITUTIONS, GHAZIABAD

Affiliated to

Dr. A.P.J. Abdul Kalam Technical University, Lucknow

(Formerly UPTU)

May, 2023

DECLARATION

We hereby declare that this submission is our own work and that, to the best of our knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial event has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgment has been made in the text.

Signature:			
Name:	Tushar Pant	Sagar Srivastava	Anshika
Roll No:	1900290100175	1900290100127	1900290100033
Date:	29/05/2023		

CERTIFICATE

This is to certify that Project Report entitled **"Image Captioning"** which is submitted by **Tushar Pant, Sagar Srivastava and Anshika** in partial fulfillment of the requirement for the award of degree B.Tech. in Department of Computer Science and Engineering of Dr. A.P.J. Abdul Kalam Technical University, Lucknow is a record of the candidates own work carried out under my supervision. The matter embodied in this report is original and has not been submitted for the award of any other degree.

Date: 29/05/2023

29/05/23

Prof. Upendra Mishra
(Assistant Professor)

ACKNOWLEDGEMENT

It gives us a great sense of pleasure to present the report of the B. Tech Project undertaken during B. Tech. Final Year. We owe special debt of gratitude to Professor Upendra Mishra, Department of Computer Science & Engineering, KIET, Ghaziabad, for his constant support and guidance throughout the course of our work. His sincerity, thoroughness and perseverance have been a constant source of inspiration for us. It is only his cognizant efforts that our endeavors have seen light of the day.

We also take the opportunity to acknowledge the contribution of Dr. Vineet Sharma, Head of the Department of Computer Science & Engineering, KIET, Ghaziabad, for his full support and assistance during the development of the project. We also do not like to miss the opportunity to acknowledge the contribution of all the faculty members of the department for their kind assistance and cooperation during the development of our project.

We also do not like to miss the opportunity to acknowledge the contribution of all faculty members, especially faculty/industry person/any person, of the department for their kind assistance and cooperation during the development of our project. Last but not the least, we acknowledge our friends for their contribution in the completion of the project.

Date:

Signature: 
Name: Tushar Pant
Roll No.: 1900290100175

Signature: 
Name: Sagar Srivastava
Roll No.: 1900290100127

Signature: 
Name: Anshika
Roll No.: 1900290100033

ABSTRACT

The goal of this project is to utilize CNN and LSTM in order to generate captions for images using deep learning techniques. With the help of large datasets and computational power, this project aims to create models that can accurately identify and describe images using natural language processing and computer vision. This paper provides a comprehensive overview of image captioning and its various techniques, with a focus on Keras library, numpy, and jupyter notebooks for implementation. Additionally, it covers the flickr dataset and the CNN approach used for image classification.

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

TABLE OF CONTENTS

Page No.

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
CBIR	Content Based Image Retrieval
CRF	Conditional Random Field
NLP	Natural Language Processing
GRU	Gated Recurrent Unit

CHAPTER 1
INTRODUCTION

On a daily basis, we come across a vast amount of images through different channels like the internet, news articles, documents, and advertisements. Although these images may not always come with descriptions, humans can generally understand them without much difficulty. However, in order for machines to provide automatic image captions, some form of interpretation is necessary. Image captioning holds significant importance as it allows for faster and more precise image searches and indexing when captions are available for every image on the internet.

From the time researchers began studying object recognition in images, it was apparent that merely providing the names of recognized objects was insufficient compared to a complete human-like description. Despite machines not possessing the ability to think, talk, and act like humans, generating natural language descriptions remains an ongoing challenge. Image captioning has numerous applications in fields like biomedicine, commerce, web searching, and military, among others. Social media platforms like Instagram and Facebook have the ability to automatically generate captions from images.

1.1 MOTIVATION

The task of generating captions for images is a crucial endeavor that spans both the domains of Computer Vision and Natural Language Processing. The ability of machines to provide descriptions for images, akin to that of humans, represents a noteworthy advancement in the field of Artificial Intelligence. The primary challenge of this task lies in accurately capturing the relationships between objects in an image and expressing them in a natural language such as English. In the past, computer systems have utilized predefined templates to generate text descriptions for images. However, this approach fails to produce the necessary variety required to generate lexically rich descriptions. This limitation has been overcome through the increased effectiveness of neural networks. Many cutting-edge models employ neural networks to generate captions by utilizing the

image as input and predicting the subsequent lexical unit in the output sentence.

1.2 IMAGE CAPTIONING

Process: Image Captioning involves generating a textual description of an image, utilizing both Natural Language Processing and Computer Vision. It is a widely researched area of Artificial Intelligence that focuses on image understanding and generating a descriptive language for the image. Understanding an image necessitates detecting and recognizing objects, comprehending the type or location of the scene, understanding the properties of the objects, and their interactions. To generate well-formed sentences, it is crucial to have syntactic and semantic language understanding. Obtaining image features plays a significant role in comprehending an image and can be utilized for automatic image indexing. Image indexing holds importance for Content-Based Image Retrieval (CBIR) and can be applied to diverse areas like biomedicine, commerce, the military, education, digital libraries, and web searching. Social media platforms like Facebook and Twitter can automatically generate descriptions from images, including the location (beach, cafe), clothing, and activities being performed.

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) [107], Scale-Invariant Feature Transform (SIFT) [87], the Histogram of Oriented Gradients (HOG) [27], 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) [17] 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) [79] 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 captions. Deep learning algorithms can handle complexities and challenges of image captioning quite well.

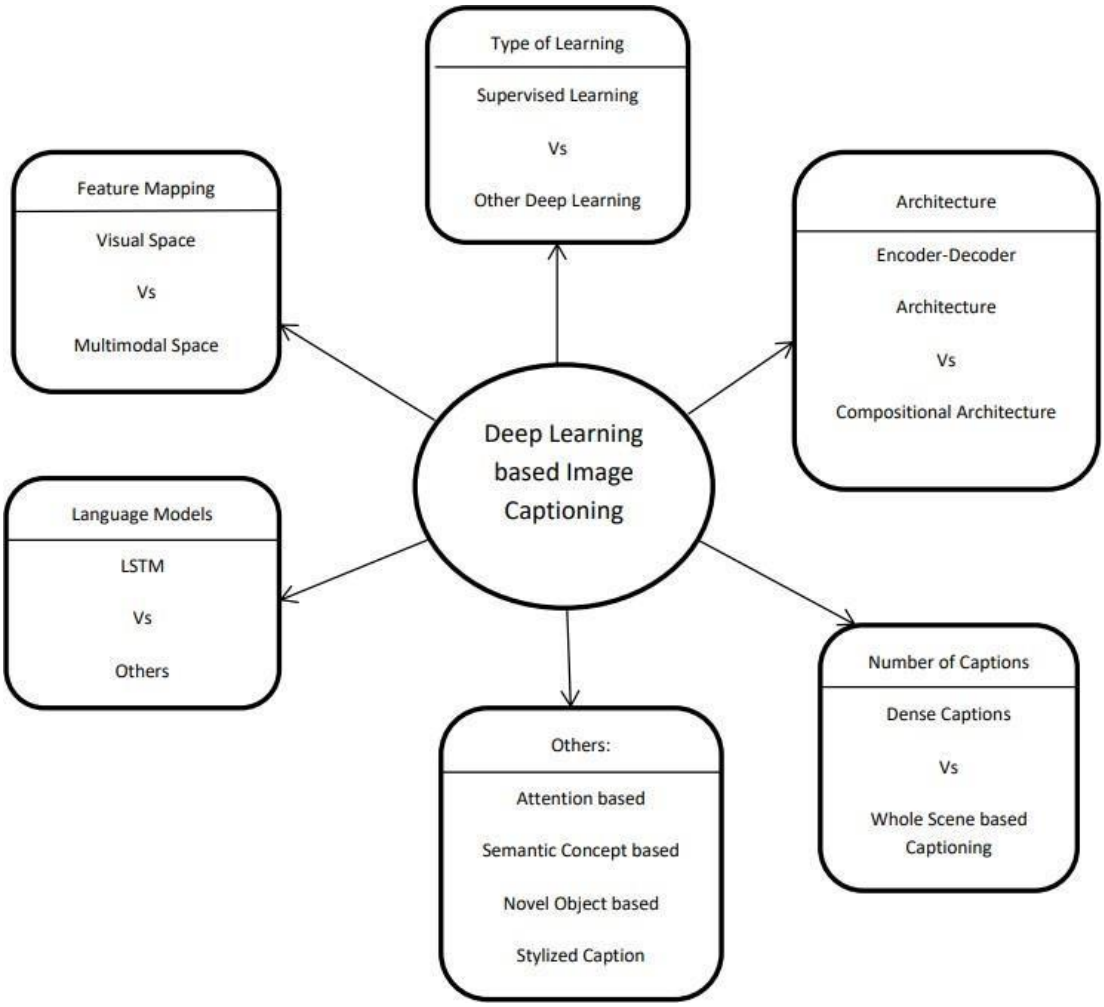

Figure 1. An overall taxonomy of deep learning-based image captioning

CHAPTER 2
LITERATURE REVIEW

Recently, there has been a surge in interest in image captioning, particularly in the realm of natural language. The need for context-based descriptions of images is crucial, and although it may seem difficult, advancements in fields such as neural networks, computer vision, and natural language processing have opened up opportunities to accurately represent the visually grounded meaning of images. We are utilizing cutting-edge techniques like Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), and appropriate image datasets with human-perceived descriptions to achieve this goal. Our alignment model has proven to be successful in retrieval experiments on datasets like Flickr.

2.1 IMAGE CAPTIONING METHODS

Before delving into the current image captioning techniques, it is important to provide an overview of the different categories of methods that exist. These categories include template-based image captioning, retrieval-based image captioning, and novel caption generation, with the latter typically utilizing deep learning techniques in the visual space. Deep learning-based image captioning methods can be further categorized based on the learning techniques used, such as supervised learning, reinforcement learning, and unsupervised learning. Dense captioning is another approach that involves generating captions for different regions of an image. Image captioning methods can also use different architectures, such as simple Encoder-Decoder or Compositional. Attention mechanism, semantic concepts, and different styles can also be employed in image descriptions. Additionally, some methods can generate descriptions for unseen objects. Lastly, while most image captioning methods use LSTM as the language model, there are also techniques that utilize other language models such as CNN and RNN, leading to a "LSTM vs. Others" language model-based category.

2.1.1 TEMPLATE-BASED APPROACHES

Fixed templates with blank slots are used in template-based approaches to generate captions. These approaches detect different objects, attributes, and actions before filling in the blank spaces in the templates. For instance, Farhadi et al. use a triplet of scene elements to fill the template slots, while Li et al. extract phrases related to detected objects, attributes, and their relationships for this purpose. Kulkarni et al. use a Conditional Random Field (CRF) to infer objects, attributes, and prepositions before filling in the gaps. Although template-based methods can generate grammatically correct captions, they have predefined templates and cannot produce variable-length captions. Parsing-based language models are more powerful than fixed template-based methods and have been introduced in image captioning. Therefore, this paper does not focus on template-based methods.

2.1.2 RETRIEVAL-BASED APPROACHES

Retrieval-based approaches in image captioning retrieve captions from either visual or multimodal space. These methods involve searching for visually similar images with their corresponding captions in the training dataset, where the captions are referred to as candidate captions. The captions for a query image are then selected from this pool of candidate captions. While retrieval-based methods generate syntactically correct and general captions, they lack the ability to generate semantically correct and image-specific captions.

2.1.3 NOVEL CAPTION GENERATION

Novel image captions refer to captions generated by combining image features with a language model, rather than selecting from existing captions. This approach addresses the limitations of retrieval-based methods and is considered a more valuable and interesting problem. Novel captions can be generated from visual or multimodal space, and typically involve analyzing the image content and generating captions using a language model. These methods can produce more semantically accurate captions than previous

approaches, and often employ deep machine learning techniques. Therefore, in this literature, we focus mainly on deep learning-based methods for generating novel image captions.

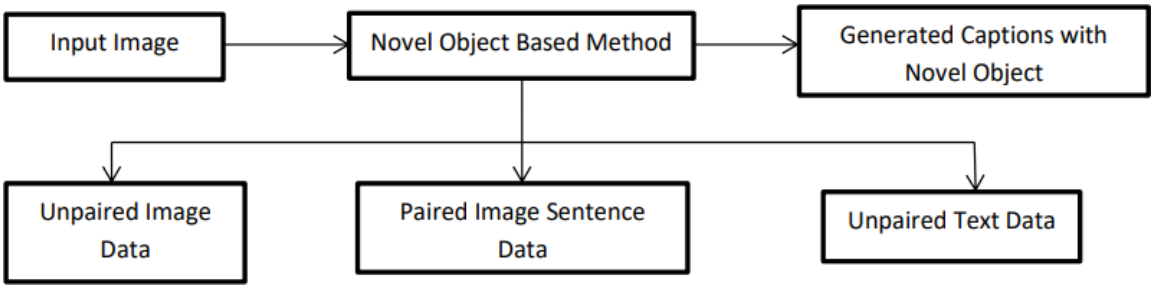

Figure 2. Novel Caption Generation

2.2 DEEP LEARNING BASED IMAGE CAPTIONING METHODS

In Figure 1, we present a comprehensive classification of deep learning-based image captioning methods, which we further discuss based on their similarities and differences. We group them into visual space versus multimodal space, dense captioning versus captions for the entire scene, supervised learning versus other deep learning, encoder-decoder architecture versus compositional architecture, and an "others" group that includes attention-based, semantic concept-based, stylized captions, and novel object-based captioning. Additionally, we create a category called "LSTM vs. Others." In the accompanying table, we provide a brief summary of the deep learning-based image captioning methods, including their names, the types of deep neural networks used for encoding image information, and the language models employed for generating captions. Finally, we assign a category label to each captioning technique based on the taxonomy presented in Figure 1.

2.2.1 VISUAL SPACE VS. MULTIMODAL SPACE

Deep learning-based image captioning techniques have the ability to generate captions

using either visual space or multimodal space. Typically, image captioning datasets contain corresponding captions as text. In visual space-based approaches, the image features and corresponding captions are separately fed into the language decoder. On the other hand, in multimodal space-based methods, a shared multimodal space is learned by combining the images and corresponding caption text. This multimodal representation is then inputted into the language decoder for generating captions.

VISUAL SPACE

Most image captioning methods rely on visual space to generate captions. This approach involves separating the image features and the corresponding captions, which are then fed independently into the language decoder.

MULTIMODAL SPACE

A common architecture for multimodal space-based image captioning methods consists of four main components: a vision part, a language encoder part, a multimodal space part, and a language decoder part. As depicted in Figure 2, the vision part employs a deep convolutional neural network to extract the visual features from the input image, while the language encoder part extracts word features and learns dense embeddings for each word. The extracted semantic temporal context is then passed to the recurrent layers. The multimodal space part maps the image features into a shared space with the word features.

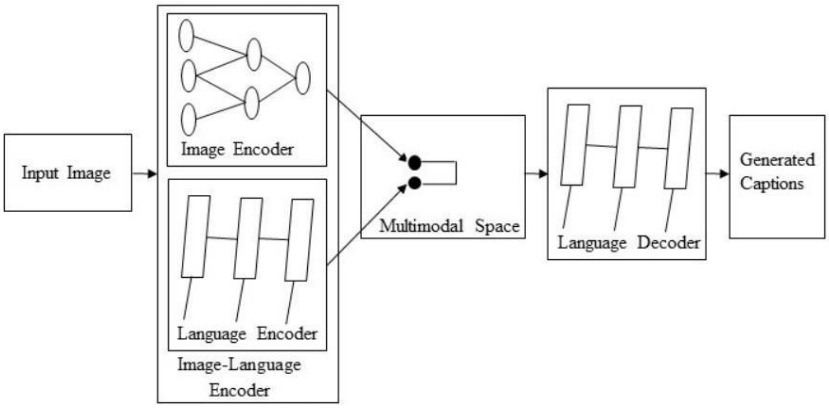

Figure 3. A block diagram of multimodal space-based image captioning

2.3 SUPERVISED LEARNING VS. OTHER DEEP LEARNING

In supervised learning, the training data includes a labeled desired output, while unsupervised learning techniques deal with unlabeled data. Reinforcement learning is another type of machine learning approach where the agent discovers data and/or labels through exploration and a reward signal. Some image captioning methods employ reinforcement learning and GAN-based approaches, which fall under the category of "Other Deep Learning" in terms of labeled data. GANs belong to the category of unsupervised learning.

2.3.1 SUPERVISED LEARNING-BASED IMAGE CAPTIONING

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, (iv) Semantic concept-based, (v) Stylized captions, (vi) Novel object-based, and (vii) Dense image captioning.

2.3.2 OTHER DEEP LEARNING-BASED IMAGE CAPTIONING

Due to the ever-increasing amount of data in our daily lives, the proportion of unlabeled data is growing. This makes it difficult to accurately annotate data. As a result, researchers have recently been placing greater emphasis on reinforcement learning and unsupervised learning-based techniques for image captioning.

2.4 DENSE CAPTIONING VS CAPTIONS FOR THE WHOLE SCENE

In dense captioning, captions are generated for each region of the scene. Other methods generate captions for the whole scene.

2.4.1 DENSE CAPTIONING

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. Johnson et al. [62] proposed an image captioning method called Dense Cap. This method localizes all the salient regions of an image and then it generates descriptions for those regions.

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 4.

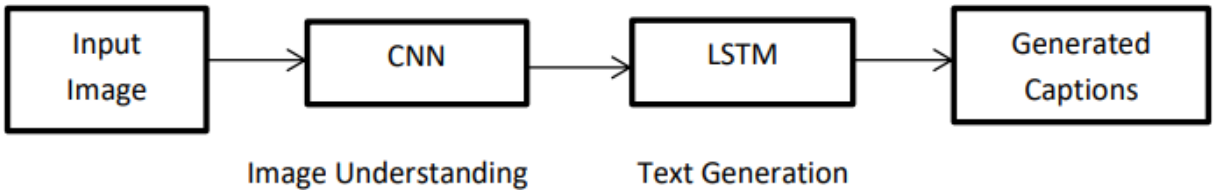

Figure 4. A block diagram of simple Encoder-Decoder architecture-based image captioning

2.4.2 CAPTIONS FOR THE WHOLE SCENE

Single or multiple captions for the entire scene can be generated by a variety of image captioning methods, including Encoder-Decoder architecture, Compositional architecture, attention-based, semantic concept-based, stylized captions, Novel object-based image captioning, and other deep learning network-based techniques.

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

The neural network-based image captioning methods work as just simple end to end manner. These methods are very similar to the encoder-decoder framework-based neural machine translation [131]. In this network, global image features are extracted from the hidden activations of CNN and then fed them into an LSTM to generate a sequence of words. A typical method of this category has the following general steps:

(1) A vanilla 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.

A simple block diagram of this category is given in Figure 5.

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 a language model using the information of Step 1 and Step 2.

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

A common block diagram of compositional network-based image captioning methods is given in Figure 5.

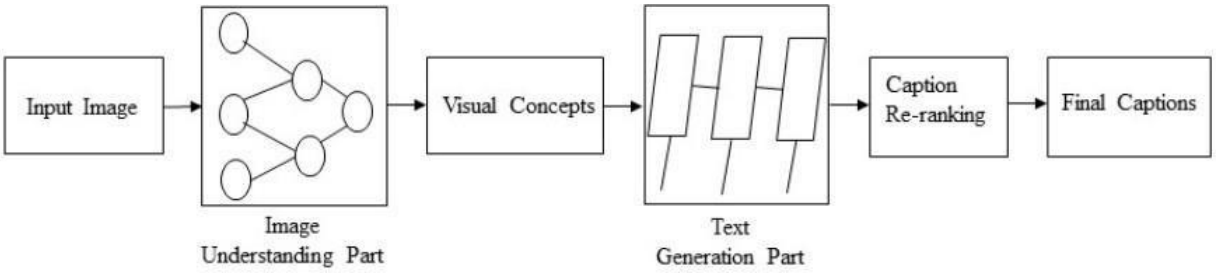

Figure 5. A block diagram of a compositional network-based captioning

2.6 LSTM VS. OTHERS

The field of image captioning lies at the intersection of computer vision and natural language processing (NLP). In NLP, tasks can be framed as sequence-to-sequence learning problems. To address these tasks, a variety of neural language models have been proposed, such as neural probabilistic language models, log-bilinear models, skip-gram models, and recurrent neural networks (RNNs). While RNNs have been widely adopted for sequence learning tasks, they suffer from issues like vanishing and exploding gradient problems and cannot effectively capture long-term temporal dependencies.

LSTM networks are a type of RNN that contain specialized units in addition to standard units. These units include a memory cell that can retain information for long periods of time. In recent years, LSTM-based models have become prevalent in sequence-to-sequence learning tasks. Another type of network, called the Gated Recurrent Unit (GRU), has a similar structure to LSTM but uses fewer gates to control information flow and does not utilize separate memory cells.

While LSTMs are effective in sequence to sequence learning, they have some limitations. For example, they do not consider the hierarchical structure of sentences and require significant storage due to long-term dependencies. In contrast, CNNs can learn internal hierarchical structures and are faster at processing. Thus, convolutional architectures have been used in other sequence to sequence tasks like machine translation and conditional image generation. One method proposed by Gu used a CNN language model for image captioning, but it couldn't effectively model the dynamic temporal behavior of the language model. To address this, they combined a recurrent network with the language

CNN to better capture temporal dependencies. Aneja proposed a convolutional architecture for the task of image captioning. They use a feedforward network without any recurrent function. The architecture of the method has four components: (i) input embedding layer (ii) image embedding layer (iii) convolutional module, and (iv) output embedding layer. It also uses an attention mechanism to leverage spatial image features. They evaluate their architecture on the challenging MSCOCO dataset and shows comparable performance to an LSTM based method on standard metrics.

CHAPTER 3

PROBLEM FORMULATION

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 (RNNs) are a type of deep learning algorithm that is specifically tailored to tackle complex computer tasks, such as speech detection and object classification. These networks are designed to process sequences of events that happen in a specific order, with each event's interpretation based on information from the preceding events

The ability of RNNs to retain memory over a long period of time is desirable for various applications such as stock prediction and speech detection. However, despite their potential, RNNs are not widely used in practical settings due to the vanishing gradient problem.

3.1.1 THE VANISHING GRADIENT PROBLEM

The limited memory capacity of RNNs is a major obstacle that severely affects their performance. In reality, the architecture of RNNs imposes constraints on their long-term memory capacity, allowing them to remember only a few sequences at a time. As a result, the memory of RNNs is effective only for shorter sequences and time intervals.

The Vanishing Gradient problem is a challenge that arises during the training of Artificial Neural Networks due to improper setting of network parameters and hyperparameters. Traditional RNNs are particularly vulnerable to this problem as it limits their memory capabilities. Increasing the number of time-steps can exacerbate this issue, resulting in gradient problems and potential loss of information during backpropagation.

CHAPTER 4
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. We will use the pre-trained model Xception.
- LSTM will use the information from CNN to help generate a description of the image.

4.1 CONVOLUTIONAL NEURAL NETWORK

Convolutional Neural Network (CNN/ConvNet) is a type of deep learning algorithm that can analyze an input image, assign significance to different features or objects in the image using learnable weights and biases, and distinguish between them. Compared to other classification algorithms, CNNs require minimal pre-processing of the input data.
CNNs are a type of deep neural networks that are specialized in processing data in the form of a 2D matrix, such as images. Therefore, CNNs are particularly useful in image-related tasks due to their ability to assign importance (learnable weights and biases) to different aspects/objects in the image. Additionally, compared to other classification algorithms, CNNs require less pre-processing.
By scanning images from left to right and top to bottom, CNNs can extract important features from images and use them to classify the images. Additionally, CNNs can handle various transformations such as translation, rotation, scaling, and perspective changes in images.

4.2 LONG SHORT-TERM MEMORY

LSTM stands for Long Short-Term Memory, and it is a type of recurrent neural network that is particularly useful for sequence prediction tasks. LSTM can predict the next word in a sequence based on the previous text. LSTM has overcome the limitations of traditional RNNs, which had short-term memory. LSTM can store relevant information throughout the input processing and use a forget gate to discard irrelevant information.

LSTMs are built to solve the problem of vanishing gradients and enable them to store information for longer periods compared to traditional RNNs. With the ability to keep a constant error, LSTMs can persist in learning across numerous time-steps and layers and use backpropagation through time.

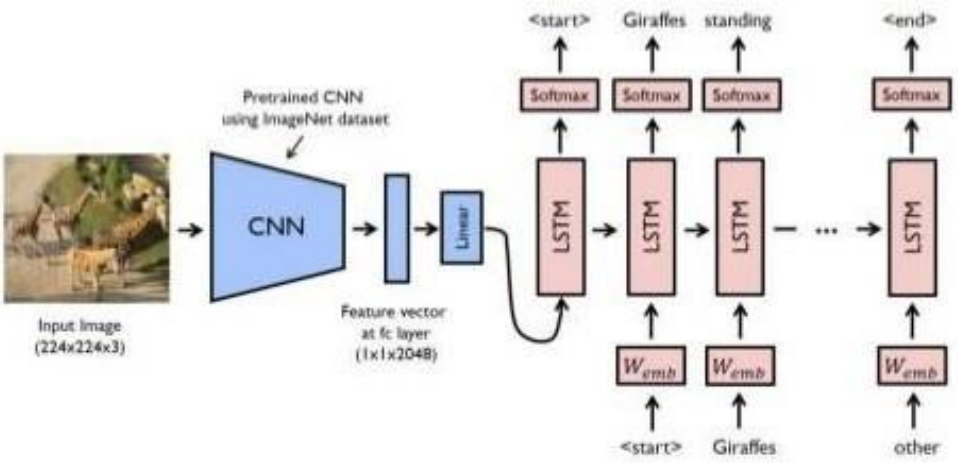

Figure 6. Model, Image Caption Generator

LSTMs incorporate gated cells that can store and manipulate information outside the usual flow of the RNN. These cells allow the network to make decisions regarding information by opening or closing the gates, which includes storing information in the cells and reading from them. Because they can retain information for a longer period, LSTMs are better suited for certain tasks compared to traditional RNNs.

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. The three major parts of the LSTM include:

Forget gate — decides how much of the previous data will be forgotten and how much of the previous data will be used in next steps. The result of this gate is in the range of0-1 while "0" forgets the previous data, "1" uses the previous data.

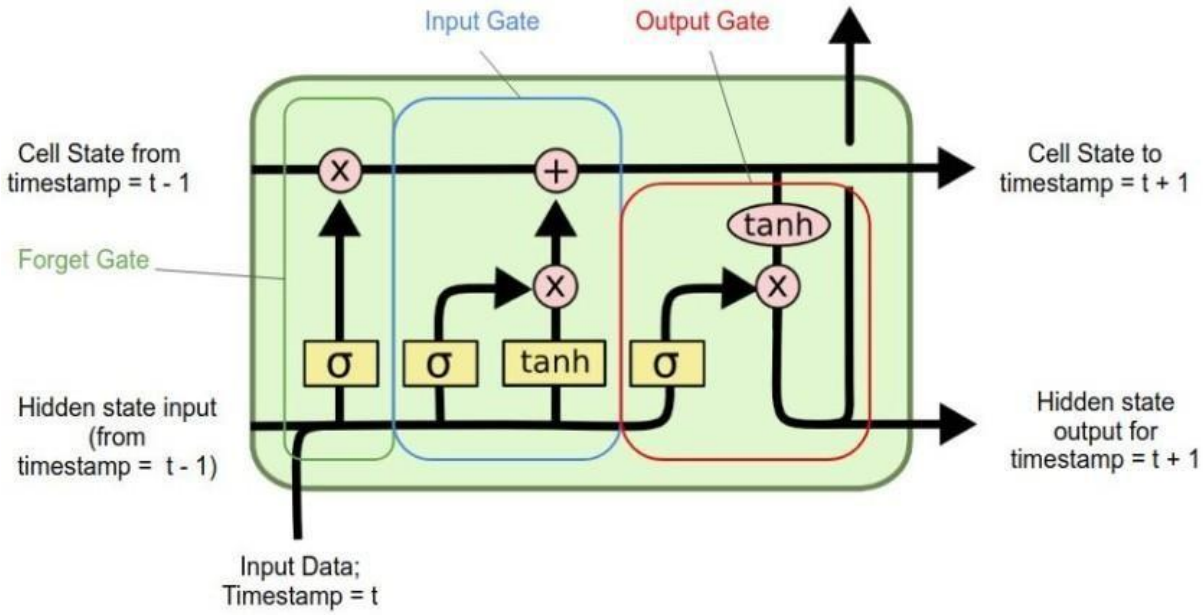

Figure 7. Forget gate, Input gate, Output gate

Input gate — determine how much new information is added to the current cell state. It controls the flow of information from the current input to the current cell state, allowing the LSTM to selectively update its memory.The input gate activation ranges between 0 and 1, indicating the amount of new information to be added to the cell state. A value of 0 indicates that the new information should be ignored, while a value of 1 indicates that it should be fully incorporated.

Output gate — allows the LSTM cell to selectively expose relevant information from the cell state while suppressing irrelevant or less important information. This helps in producing accurate and contextually meaningful outputs based on the input and the accumulated information stored in the cell state.

In the CNN LSTM architecture, a Convolutional Neural Network (CNN) is used to extract features from input data, which is then combined with LSTMs for sequence prediction. Originally known as Long-term Recurrent Convolutional Networks (LRCN), this architecture is now commonly referred to as CNN LSTM, specifically when LSTMs use a CNN as a front end.

This type of architecture is commonly employed in the task of generating textual descriptions of images. The crucial component is a CNN that has been pre-trained on a complex image classification task, which is then adapted to serve as a feature extractor for the task of generating image captions.

CHAPTER 5
SYSTEM DESIGN

This project requires a dataset which has both images and their caption. The dataset should be able to train the image captioning model.

5.1 FLICKR 8K DATASET

The Flickr8k dataset is a widely used benchmark dataset for image captioning. It comprises a collection of 8000 images that are sourced from diverse groups on the Flickr website, with each image having five descriptive captions. The captions provide clear explanations of the entities and events depicted in the corresponding image, and the dataset covers a wide range of scenarios and events. Notably, the dataset does not include images of well-known people or places, which makes it more generalizable. The dataset is partitioned into a training set with 6000 images, a development set with 1000 images, and a test set with 1000 images.

• Multiple captions mapped for a single image makes the model generic and avoids overfitting of the model.

• Diverse category of training images can make the image captioning model to work for multiple categories of images and hence can make the model more robust.

5.2 IMAGE DATA PREPARATION

To train an image using a deep learning model, it is necessary to extract suitable features from it. This process of feature extraction is mandatory. In this regard, Convolutional Neural Network (CNN) with the Visual Geometry Group (VGG-16) model is used to extract the features. This model has been successful in winning the ImageNet Large Scale Visual Recognition Challenge in 2015, where it was used to classify images into 1000 classes. Thus, this model is well-suited for image captioning as it requires the identification of images.

VGG-16 is a neural network architecture with 16 layers that uses 3x3 convolutional layers for better feature extraction from images. It incorporates max pooling layers to reduce the size of the image volume. The final classification layer of the network is removed, and the internal representation of the image just before classification is extracted as a feature. The input image must have a dimension of 224x224, and the model returns a 1D vector of 4096 elements representing the features extracted from the image.

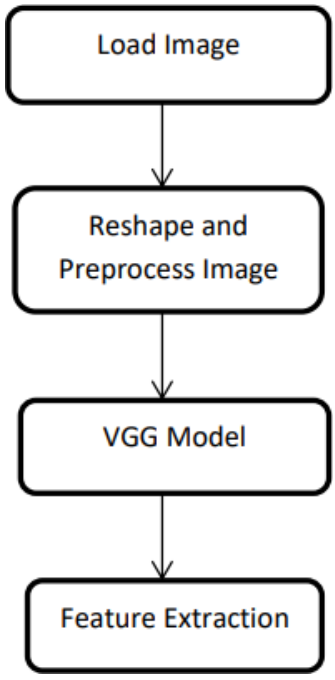

Figure 8. Feature Extraction in Image using VGG

5.3 CAPTION DATA PREPARATION

During the data preparation phase, a dictionary is created where the key represents the image id and the values correspond to the multiple descriptions provided for that particular image in the Flickr 8k dataset.

5.3.1 DATA CLEANING

In order to make the text dataset work in machine learning or deep learning models, raw text should be converted to a usable format. The following text cleaning steps are done before using it for the project:

• Removal of punctuations.

• Removal of numbers.

• Removal of single length words.

• Conversion of uppercase to lowercase characters. Stop words are not removed from the text data as it will hinder the generation of a grammatically complete caption which is needed for this project. Table 1 shows samples of captions after data cleaning.

Original Captions	Cleaned Captions
A dog sitting on a beach enjoying the sunny weather.	Dog sitting on beach, enjoying the sun.
People walking in a crowded city street.	People walking on a busy city street.
A plate of delicious spaghetti with meatballs.	Plate of spaghetti topped with meatballs.
A little girl rides in a child's swing.	Little girl rides in a child swing.

Table. 5.1: Data cleaning of captions

CHAPTER 6

IMPLEMENTATION

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 have installed all the following necessary libraries:

- pip install tensorflow
- keras
- pillow
- numpy
- tqdm
- jupyterlab

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 Xception pre-trained 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.

6.3 BUILDING THE PYTHON BASED PROJECT

Let's start by initializing the jupyter notebook server by typing jupyter lab in the console of your project folder. It will open up the interactive Python notebook where you can run your code. Create a Python3 notebook and name it training_caption_generator.ipynb.

6.3.1 GETTING AND PERFORMING DATA CLEANING

The main text file which contains all image captions is **Flickr8k.token** in our **Flickr_8k_text** folder.

```
{'1000268201_693b08cb0e': ['A child in a pink dress is climbing up a set of stairs in an entry way .',
  'A girl going into a wooden building .',
  'A little girl climbing into a wooden playhouse .',
  'A little girl climbing the stairs to her playhouse .',
  'A little girl in a pink dress going into a wooden cabin .'],
 '1001773457_577c3a7d70': ['A black dog and a spotted dog are fighting',
  'A black dog and a tri-colored dog playing with each other on the road .',
  'A black dog and a white dog with brown spots are staring at each other in the street .',
  'Two dogs of different breeds looking at each other on the road .',
  'Two dogs on pavement moving toward each other .'],
 '1002674143_1b742ab4b8': ['A little girl covered in paint sits in front of a painted rainbow with her hands in a bowl .',
  'A little girl is sitting in front of a large painted rainbow .',
  'A small girl in the grass plays with fingerpaints in front of a white canvas with a rainbow on it .',
  'There is a girl with pigtails sitting in front of a rainbow painting .',
  'Young girl with pigtails painting outside in the grass .'],
 '1003163366_44323f5815': ['A man lays on a bench while his dog sits by him .',
  'A man lays on the bench to which a white dog is also tied .',
  'a man sleeping on a bench outside with a white and black dog sitting next to him .',
  'A shirtless man lies on a park bench with his dog .',
  'man laying on bench holding leash of dog sitting on ground'],
 '1007129816_e794419615': ['A man in an orange hat starring at something .',
  'A man wears an orange hat and glasses .',
  'A man with gauges and glasses is wearing a Blitz hat .',
  'A man with glasses is wearing a beer can crocheted hat .',
  'The man with pierced ears is wearing glasses and an orange hat .'],
 '1007320043_627395c3d8': ['A child playing on a rope net .',
  'A little girl climbing on red roping .',
  'A little girl in pink climbs a rope bridge at the park .',
  'A small child grips onto the red ropes at the playground .',
  'The small child climbs on a red ropes on a playground .'],
```

Figure 9. Flickr DataSet text Format

The format of our file is image and caption separated by a new line ("\n").

Each image has 5 captions and we can see that #(0 to 5)number is assigned for each caption.

We will define 5 functions:

- **load_doc(filename)** – For loading the document file and reading the contents inside the file into a string.
- **all_img_captions(filename)** – This function will create a **descriptions** dictionary that maps images with a list of 5 captions. The descriptions dictionary will look something like the Figure.

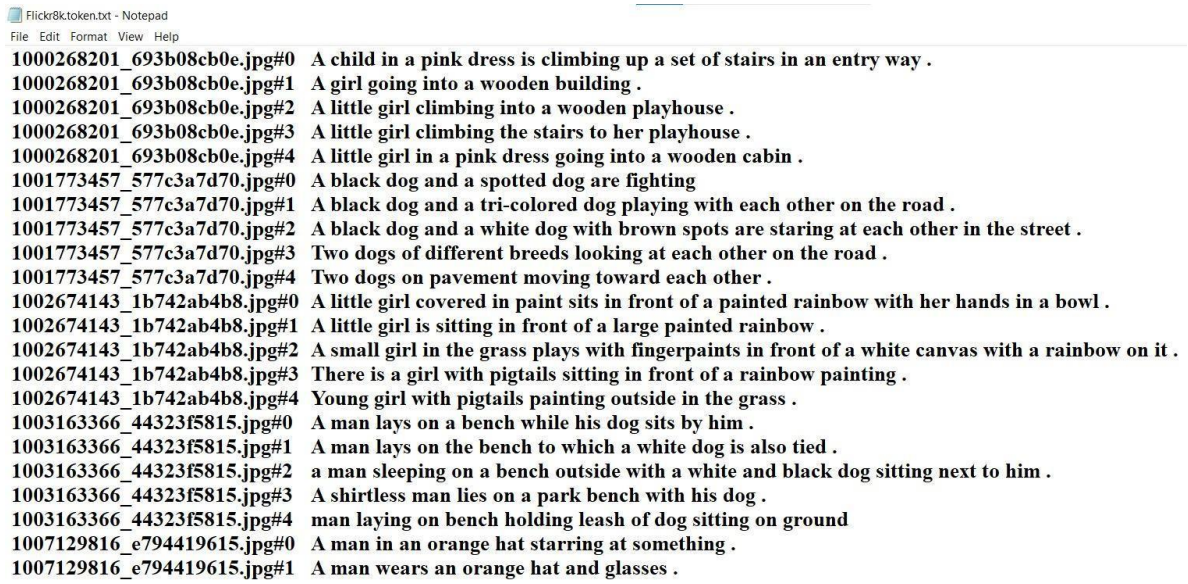
Flickr8k.token.txt - Notepad
File Edit Format View Help
1000268201_693b08cb0e.jpg#0 A child in a pink dress is climbing up a set of stairs in an entry way .
1000268201_693b08cb0e.jpg#1 A girl going into a wooden building .
1000268201_693b08cb0e.jpg#2 A little girl climbing into a wooden playhouse .
1000268201_693b08cb0e.jpg#3 A little girl climbing the stairs to her playhouse .
1000268201_693b08cb0e.jpg#4 A little girl in a pink dress going into a wooden cabin .
1001773457_577c3a7d70.jpg#0 A black dog and a spotted dog are fighting
1001773457_577c3a7d70.jpg#1 A black dog and a tri-colored dog playing with each other on the road .
1001773457_577c3a7d70.jpg#2 A black dog and a white dog with brown spots are staring at each other in the street .
1001773457_577c3a7d70.jpg#3 Two dogs of different breeds looking at each other on the road .
1001773457_577c3a7d70.jpg#4 Two dogs on pavement moving toward each other .
1002674143_1b742ab4b8.jpg#0 A little girl covered in paint sits in front of a painted rainbow with her hands in a bowl .
1002674143_1b742ab4b8.jpg#1 A little girl is sitting in front of a large painted rainbow .
1002674143_1b742ab4b8.jpg#2 A small girl in the grass plays with fingerpaints in front of a white canvas with a rainbow on it .
1002674143_1b742ab4b8.jpg#3 There is a girl with pigtails sitting in front of a rainbow painting .
1002674143_1b742ab4b8.jpg#4 Young girl with pigtails painting outside in the grass .
1003163366_44323f5815.jpg#0 A man lays on a bench while his dog sits by him .
1003163366_44323f5815.jpg#1 A man lays on the bench to which a white dog is also tied .
1003163366_44323f5815.jpg#2 a man sleeping on a bench outside with a white and black dog sitting next to him .
1003163366_44323f5815.jpg#3 A shirtless man lies on a park bench with his dog .
1003163366_44323f5815.jpg#4 man laying on bench holding leash of dog sitting on ground
1007129816_e794419615.jpg#0 A man in an orange hat starring at something .
1007129816_e794419615.jpg#1 A man wears an orange hat and glasses .

Figure 10. Flickr DataSet Python File

- **cleaning_text(descriptions)** – During data preparation, a function is used to clean all descriptions. Textual data requires cleaning before it can be used effectively. Cleaning involves removing any irrelevant or unnecessary elements from the text. In this case, the function removes all punctuation marks, converts all text to lowercase and eliminates any words that contain numbers. For example, the caption "A man riding on a three-wheeled wheelchair" would be transformed into "man riding on 3 wheeled wheelchair" after cleaning.
- **text_vocabulary(descriptions)** – This is a simple function that will separate all the unique words and create the vocabulary from all the descriptions.
- **save_descriptions(descriptions, filename)** – This function will create a list

of all the descriptions that have been preprocessed and store them into a file. We will create a descriptions.txt file to store all the captions. It will look something like this:

6.3.2 EXTRACTING THE FEATURE VECTOR FROM ALL IMAGES

The technique we are using is known as transfer learning, where we utilize pre-trained models that have already been trained on large datasets, and extract the features from these models to be used for our specific task. Specifically, we are using the Xception model, which has been trained on the imagenet dataset with 1000 different classes for classification. We can import this model directly from the keras.applications library, and the weights will automatically download (internet connection required). However, since the Xception model was originally designed for the imagenet dataset, we need to make some changes to integrate it with our own model. It is worth noting that the Xception model requires input images to be of size 2992993. We will remove the last classification layer and extract the 2048 feature vector for our image captioning task.

model = Xception(include_top=False, pooling="avg")

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.

This process can take a lot of time depending on your system. I am using an Nvidia 1050 GPU for training purpose so it took me around 7 minutes for performing this task. However, if you are using CPU then this process might take 1-2 hours. You can comment out the code and directly load the features from our pickle file.

6.3.3 LOADING DATASET FOR TRAINING THE MODEL

In our Flickr_8k_test folder, we have Flickr_8k.trainImages.txt file that contains a list of 6000 image names that we will use for training. For loading the training dataset, we need more functions:

- **load_photos(filename)** – This will load the text file in a string and will return the list of image names.

- **load_clean_descriptions(filename, photos)** – This function will create a dictionary that contains captions for each photo from the list of photos. We also append the and identifier for each caption. We need this so that our LSTM model can identify the starting and ending of the caption.
- **load_features(photos)** – This function will give us the dictionary for image names and their feature vector which we have previously extracted from the Xception model.

6.3.4 TOKENIZING THE VOCABULARY

Computers don‟t understand English words, for computers, we will have to represent them with numbers. So, we will map each word of the vocabulary with a unique index value. Keras library provides us with the tokenizer function that we will use to create tokens from our vocabulary and save them to a "tokenizer.p" pickle file. 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.

6.3.5 Create Data generator

Let us first see how the input and output of our model will look like. To make this task into a supervised learning task, we have to provide input and output to the model for training. We have to train our model on 6000 images and each image will contain 2048 length feature vector and caption is also represented as numbers. This amount of data for 6000 images is not possible to

hold into memory so we will be using a generator method that will yield batches. The generator will yield the input and output sequence.

For example: 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.

Features	Text Sequence	Word to Predict
Image features	"Start"	"I"
Image features, "I"	"Start I"	"want"
Image features, "I want"	"Start I want"	"to"
Image features, "I want to"	"Start I want to"	"go"
Image features, "I want to go"	"Start I want to go"	"to"
Image features, "I want to go to"	"Start I want to go to"	"the"
Image features, "I want to go to the"	"Start I want to go to the"	"beach"

Table. 6.1. Word Prediction Generation Step by Step

6.3.6 Defining the CNN-RNN model

To define the structure of the model, we will be using the Keras Model from Functional API. It will consist of three major parts:

- **Feature Extractor** – The feature extracted from the image has a size of 2048, with a dense layer, we will reduce the dimensions to 256 nodes.
- **Sequence Processor** – An embedding layer will handle the textual input, followed by the LSTM layer.
- **Decoder** – By merging the output from the above two layers, we will process by the dense layer to make the final prediction. The final layer will contain the number of nodes equal to our vocabulary size. Visual representation of the final model is given in the figure

6.3.7 Training the model

To train the model, we will be using the 6000 training images by generating the input and output sequences in batches and fitting them to the model using model.fit_generator() method. We also save the model to our models folder. This will take some time depending on your system capability.

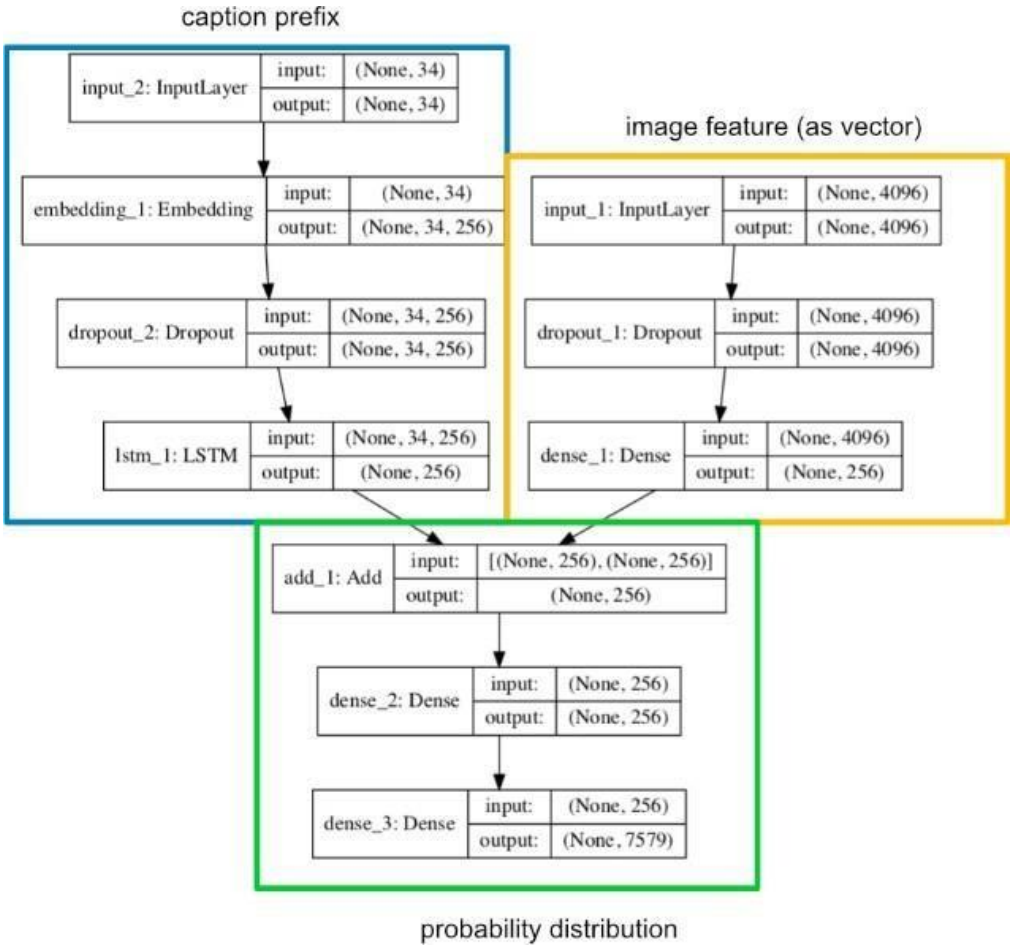

Figure 11. Final Model Structure

6.3.8 Testing the model

The model has been trained, now, we will make a separate file testing_caption_generator.py which will load the model and generate predictions. The predictions contain the max length of index values so we will use the same tokenizer.p

pickle file to get the words from their index values.

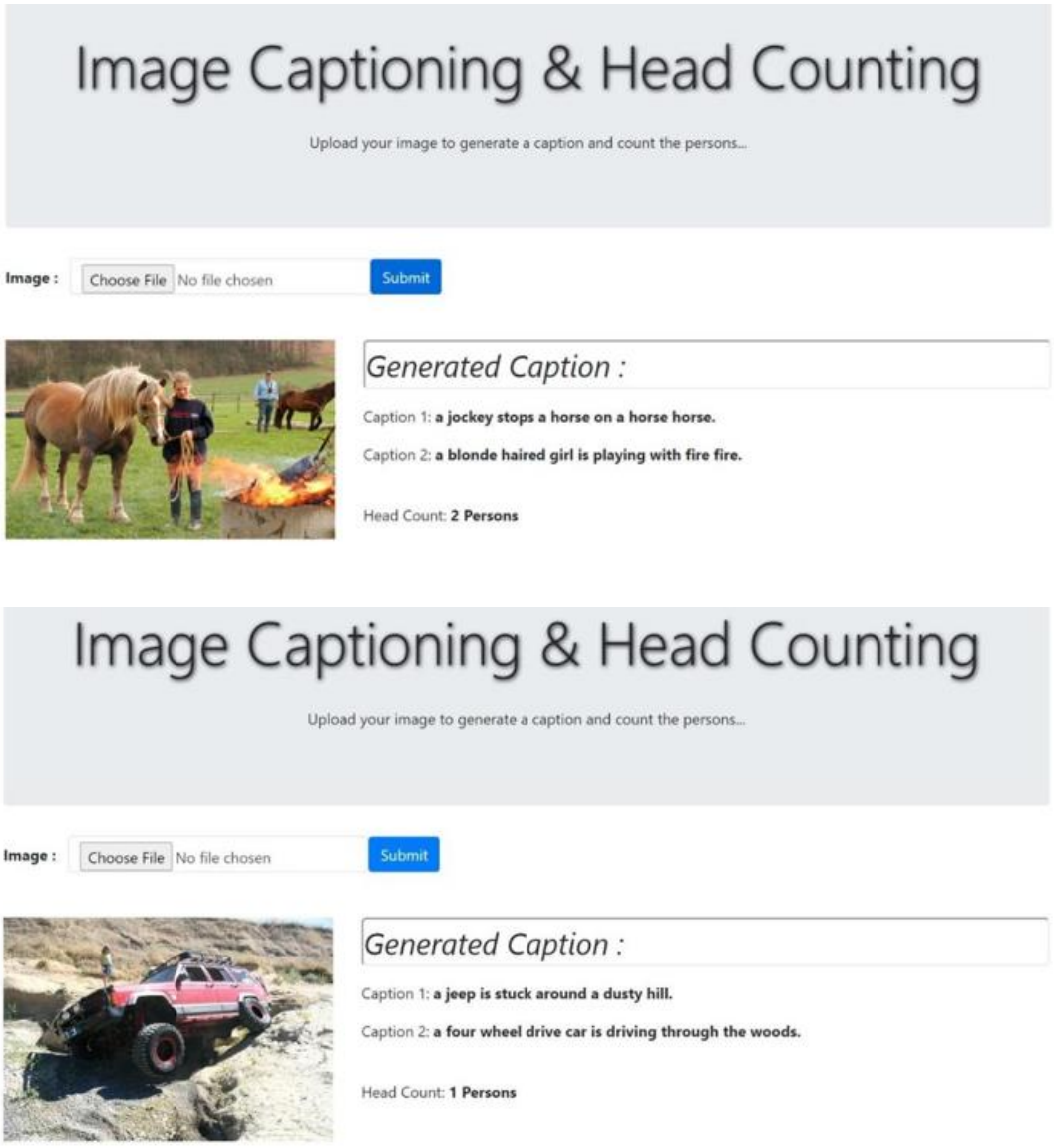

Figure 12. Output Caption of Given Image

CHAPTER 7

CONCLUSION, LIMITATION AND FUTURE SCOPE

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.

7.1 CONCLUSION

This paper provides an overview of deep learning-based image captioning techniques. A taxonomy of these techniques is presented, along with a generic block diagram of major groups and their advantages and disadvantages. Evaluation metrics and datasets, including their strengths and weaknesses, are also discussed. The experimental results are briefly summarized. Potential research directions in this area are outlined. Despite the significant progress made in deep learning-based image captioning, there is still a need for a robust method that can generate high-quality captions for nearly all images. As new deep learning network architectures emerge, automatic image captioning will continue to be an active research area.

For our project, we utilized the Flickr_8k dataset, which consists of approximately 8000 images along with their corresponding captions stored in a text file. Despite significant advancements in deep learning-based image captioning methods, a reliable method for generating high-quality captions for almost all images has yet to be achieved. As new deep learning network architectures emerge, automatic image captioning will continue to be an active area of research. With the increasing number of users on social media who regularly post photos, the potential applications of image captioning are vast and this project can greatly assist in this regard.

7.2 LIMITATIONS

The framework of the neural image caption generator is valuable for teaching the mapping from images to human-like image captions. Through training with a vast

number of image-caption pairs, the model acquires the ability to grasp pertinent semantic information from visual features. When using a static image, the image caption generator tends to focus on image classification features, rather than features that are useful for generating captions. To address this limitation and increase the amount of task-relevant information, we can train the image embedding model (in this case, the VGG-16 network used to encode image features) as a component of the caption generation model. This allows us to fine-tune the image encoder to better suit the task of generating captions.

Also, if we actually look closely at the captions generated, we notice that they are rather mundane and commonplace. Take this possible image-caption pair for instance:

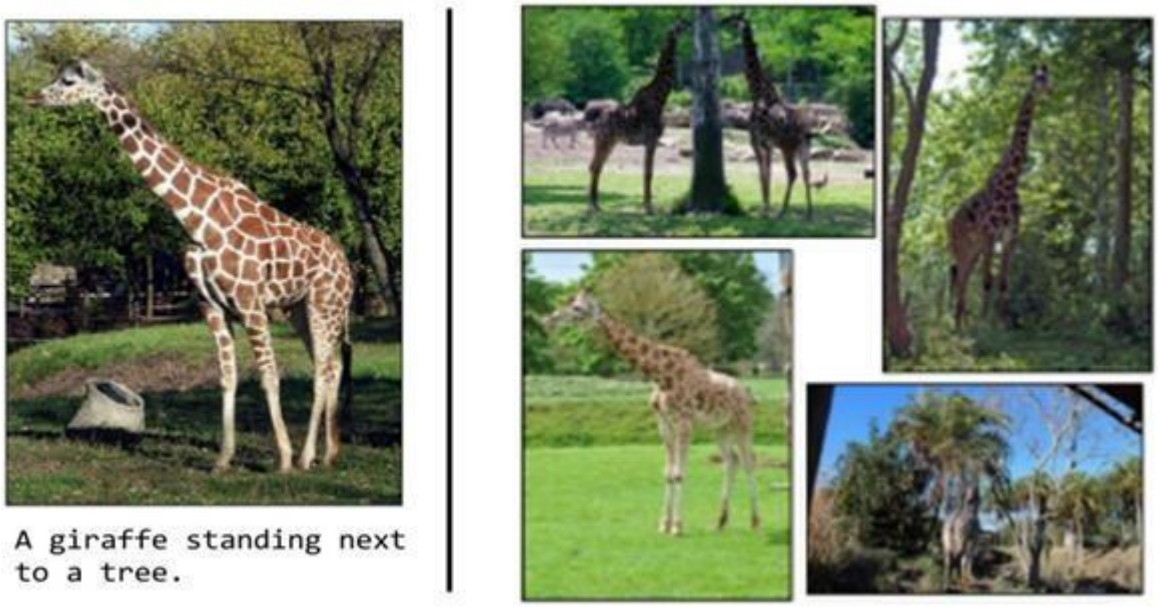

Figure 13. The above picture depicts clear limitation of the model because it relies mostly on the training dataset

This is most certainly a “giraffe standing next to a tree.” However, if we look at other pictures, we will likely notice that it generates a caption of “a giraffe next to a tree” for any picture with a giraffe because giraffes in the training set often appear near trees.

7.3 FUTURE SCOPE

Image captioning has emerged as a significant challenge due to the increasing number of images available on social media and the internet. This report provides an overview of previous research on image retrieval, highlighting the techniques and methodologies used. As feature extraction and similarity calculation are complex tasks in this field, there is great potential for further research in the future. Current image retrieval systems rely on features such as color, tags, histograms, etc. to calculate similarity. Due to the lack of consideration of image context, the existing methodologies for image retrieval cannot provide completely accurate results. Therefore, future research in image retrieval can benefit from incorporating image captioning techniques to better understand the context of the images. Additionally, to improve the precision of identifying classes, this project can be enhanced by incorporating more image captioning datasets for training. Furthermore, combining this methodology with previous image retrieval techniques such as histograms and shapes can be explored to evaluate if the image retrieval results can be further improved.

REFERENCES

1. Md Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, Hamid Laga (2018), Murdoch University. A Comprehensive Survey of Deep Learning for Image Captioning, ACM Computing Surveys 51(6)
2. Shuang Bai, Shan An (2018), Beijing Jiaotong University. A Survey on Automatic Image Caption Generation.
3. Naeha Sharif, Lyndon White, Mohammed Bennamoun, Syed Afaq Ali Shah (2018), University of Western Australia. Learning-based Composite Metrics for Improved Caption Evaluation, Proceedings of ACL 2018, Student ResearchWorkshop.
4. Cristian Bodnar (2018), University of Cambridge. Text to Image Synthesis UsingGenerative Adversarial Networks.
5. Jyoti Aneja, Aditya Deshpande, Alexander G. Schwing (2018), University of Illinois, Urbana-Champaign. Convolutional Image Captioning, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
6. Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney (2018). Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
7. Bo Dai, Sanja Fidler, Raquel Urtasun, Dahua Lin (2017). Towards Diverse and Natural Image Descriptions via a Conditional GAN, IEEE International Conference on Computer Vision (ICCV)

APPENDIX

Automated Image Captioning: A Study In Deep Learning For Generating Accurate And Descriptive Captions

Upendra Mishra,a) Sagar Srivastava,b) Tushar Pant,c) and Anshikad)

KIET Group of Institutions, Delhi-NCR, Ghaziabad, India

a)upendra.mishra@kiet.edu

b)srivastavasagar2001@gmail.com

c)tusharpant2001@gmail.com

d)info2anshika@gmail.com

Abstract. *It is challenging to label Earth from space with air conditions, many types of land cover, and land use. To help the international community comprehend where, how, and why deforestation occurs all across the world, we proposed an algorithm. Upcoming developments in satellite image science are anticipated to create new opportunities for a more precise examination of both significant and minute changes, including deforestation, that are taking place on Earth. Around 5% of the Amazon rainforest has been destroyed in the past 40 years. The forest is being analyzed and estimated by this application. In order to train deep convolutional neural networks (CNNs) on visual information from satellite images, a variety of classification frameworks, such gate recurrent unit label captioning, are used. In this study, we introduce a machine learning model that uses captions to explain images. This study focuses on finding various objects in an image, figuring out how they relate to one another, and creating captions for those objects. The suggested experiment will be demonstrated using the Flickr 8K dataset, Python 3, and a machine learning technique known as Transfer Learning. The potential for using image caption generators to assist persons who are blind is enormous.*

Keywords: ML, AI, Image, Caption, CNN, RNN, LSTM, Neural Networks

INTRODUCTION

A. Overview

On a daily basis, we come across numerous images from diverse origins such as the internet, news articles, document diagrams, and advertisements. These images often lack

descriptions, leaving the viewers to interpret them on their own. While humans can typically understand the meaning behind images without detailed captions, machines require some form of interpretation to generate automatic image captions that are understandable to humans. Image captioning holds significant importance due to its ability to aid in accessibility for individuals with visual impairments, enhance search engine optimization, assist in e-commerce by providing detailed product information, suggest ideas for content creation, and facilitate the creation of educational materials for students. Providing captions for every image on the internet could potentially result in faster and more accurate image searches and indexing. Initially, researchers focused on identifying objects within images, but it soon became apparent that simply providing the names of recognized objects was insufficient. A more comprehensive and human-like description of the image was necessary to make a lasting impression. Until machines can replicate human thought processes, communication, and behavior, generating natural language descriptions will continue to pose a challenge. Nonetheless, image captioning has numerous practical applications in diverse fields, including biomedicine, e-commerce, web search, and military operations. Social media platforms such as Instagram and Facebook can automatically generate captions from images. [13]

1) Process: The process of generating a textual description of a picture using computer vision and natural language processing methods is known as image captioning. It is a well-known area of study within Artificial Intelligence (AI) that involves visual comprehension and the creation of a corresponding linguistic description. To understand an image, object recognition and detection play a crucial role. Understanding the type of scene or location that is being shown in the image as well as the characteristics of the things there and how they relate to one another is also necessary for image comprehension. Both syntax and semantics must be well-understood in order to construct intelligible and grammatically accurate phrases. Identifying pertinent picture elements is essential for understanding an image. Automatic picture indexing using image captioning technology is essential for content-based image retrieval (CBIR) and has practical applications in various industries such as biomedicine, business, the military, education, digital libraries, and web search. Additionally, prominent social media sites like Facebook and Twitter enable the automatic generation of image descriptions, which might contain details about the scene's setting (such as a beach or cafe), clothing, and going-on.

2) Techniques: There are two primary categories of techniques used for this purpose, namely conventional machine learning and deep machine learning approaches. Traditional machine learning techniques rely on manually created features, such as Local Binary Patterns (LBP), Scale-Invariant Feature Transforms (SIFT), Histograms of Oriented Gradients (HOG), and combinations thereof. In these methods, characteristics

are taken from the input data and used to categorize an item using a classifier like Support Vector Machines (SVM). [11] The drawback of employing handcrafted features is that they are customized to particular activities, and it is impractical to extract features from a large variety of diverse data. Real-world data, including photographs and videos, are very complicated and have a wide range of semantic interpretations. Deep machine learning approaches, in contrast, can handle a wide variety of photos and videos and automatically discover features from the training data. While classifiers like soft max are used for classification, Convolutional Neural Networks (CNN) are a classic example of deep machine learning approaches that excel at feature learning. To create subtitles, CNNs are frequently combined with recurrent neural networks (RNN) or long short-term memory networks (LSTM). Deep learning [11] algorithms are skilled at navigating the difficulties and complexity of creating image captions.

B. Research Problem

With the air quality and various descriptions of the land cover or land use, it is challenging to label the satellite image. The results of the algorithms will assist the international community in determining the best course of action as well as what, how, and why deforestation is happening all across the world. Furthermore, it is usually difficult for present methods to discern between human and natural causes of forest degradation. Although it has already been shown that using higher resolution photographs in this way is very beneficial, stable methods for imaging planets have not yet been developed. In order to address this issue, our plan is to develop an encoder-decoder architecture based on the CNN and RNN algorithms. The advent of deep learning has revolutionized the field of computer vision, allowing machines to recognize and classify images with remarkable accuracy. One of the challenging tasks in this area is the automatic generation of image captions, which involves not only identifying the objects and actions depicted in an image, but also understanding the context and generating a coherent and semantically meaningful description. The creation of image captions has several useful uses, including helping those who are blind, raising search engine rankings, and improving social networking platform user experiences. The purpose of this research study is to create and assess a deep learning-based picture caption generator that can provide accurate and grammatically correct descriptions of images automatically. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs), two cutting edge deep learning approaches, will be used by the proposed system to extract pertinent visual elements from the input image and produce a related caption. The research will explore different architectures and training strategies, as well as evaluate the system's performance on standard benchmarks and real-world datasets. [1]

C. Motivation

Creating descriptions for images is a significant task that pertains to the fields of Computer Vision and Natural Language Processing. It is a remarkable achievement in the realm of Artificial Intelligence to emulate the human capacity of generating descriptions for images using a machine. The key difficulty in generating text descriptions for images is to effectively capture the relationships between objects within the image and express them in a natural language such as English. In the past, computer systems have relied on predefined templates to generate text descriptions for images. However, this approach lacks the necessary diversity needed to produce linguistically sophisticated text descriptions. The deficiency in text description generation for images has been mitigated through the advancement of neural networks. Modern models employ neural networks to generate captions by taking an image as input and predicting the subsequent lexical units in the output sentence. Labeling satellite images with diverse captions that accurately depict land cover or land use, including air conditions, can be a challenging task.

However, the output produced by advanced algorithms can provide valuable insights to the global community about the causes and effects of deforestation worldwide, as well as suggest the best course of action. Unfortunately, existing techniques often fail to differentiate between human-induced and natural causes of forest decline. Utilizing high-resolution images has proven to be a highly effective method for this task. However, reliable techniques for processing Planet imaging have yet to be developed. [5] To address this challenge, our aim is to develop an encoder-decoder architecture that leverages Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) to generate captions for these satellite photos.

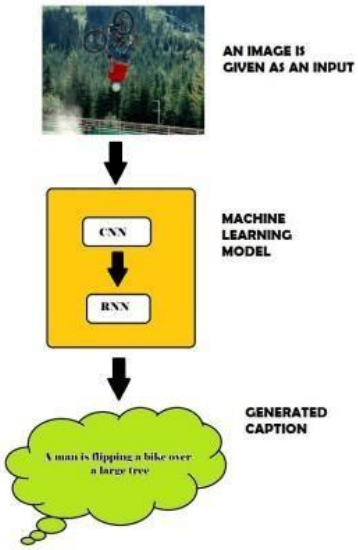

FIGURE 1. Pictorial Representation of the Model

RELATED WORK

Over the years, the problem of generating captions for images on the Internet has been addressed by various researchers using different algorithms and techniques. One such technique was proposed by Krizhevsky and team, who developed a neural network based on non-saturating neurons and an efficient GPU convolution function. To tackle overfitting, they utilized a regularization method called dropout. Their network comprised max pooling layers and a final 1000-way soft max. Another significant contribution to this field was made by Deng and colleagues, who introduced a large collection of images called ImageNet, categorized into a semantic hierarchy with numerous nodes. Karpathy and FeiFei utilized this dataset, along with its sentence descriptions, to explore the connections between visual data and language. They developed a Multimodal Recurrent Neural Network architecture that leveraged the collinear arrangement of features to generate creative image descriptions. Yang and colleagues have put forward a technique that can enhance the comprehension of images by automatically generating natural language descriptions for them. [9] This approach employs a multi-model neural network that incorporates object detection and localization modules, and its similarity to the human visual system is noteworthy. Another study by Aneja and team has proposed a convolutional network model for both machine translation and conditional picture generation. This model is fundamentally sequential across time and addresses the complexity associated with LSTM units. Extensive experimentation with network design was conducted by Pan and colleagues on large datasets containing diverse content, resulting in a novel model that outperformed previous captioning models in terms of accuracy. Meanwhile, Vinyals and team developed a generative model that utilizes computer vision and machine translation to generate natural image descriptions with a high level of accuracy. This model ensures that the generated sentences closely describe the target image. Xu and co-workers proposed an attention-based model that can automatically identify and characterize visual regions. This model was trained using traditional back propagation techniques, maximizing a lower variable.

PROPOSED WORK AND METHODOLOGY

There are numerous datasets of annotated images available for image captioning tasks, including MS COCO, Flickr 8K, and Pascal VOC. In this study, we utilized the Flickr 8K Image Captioning dataset, which comprises 8,092 images sourced from Kaggle.com. This particular dataset features a range of everyday activities, each accompanied by a corresponding caption. The images in the dataset are labeled based on the objects they contain, and a description is generated accordingly. [6] In this study, we split the dataset of 8,000 images into three groups: a development set and a test set, each containing 1,000

images, and a training set containing 6,000 images. The potential of the image-caption pairs to join previously undiscovered images and captions with each other must be assessed in order to evaluate the image-caption pairs. [10] The BLEU (Bilingual Evaluation Understudy) Score can be used to evaluate models that produce natural language sentences. It compares a natural statement to one that was created by a human. It is frequently used to assess the effectiveness of machine translation. [8] For the purpose of calculating a BLEU score, sentences are compared using a modified n-gram precision approach, where accuracy is determined using the following equation:

$$BLEU = BP * exp(\sum_{n=1}^{N} W_n loglog P_n) \quad (1)$$

$$Count_c = min(Count, MaxRefCount) \quad (2)$$

where c = clip

Our image captioning model is built on the backbone of multimodal recurrent network and convolutional neural networks (CNNs). Initially, the image features are extracted with the help of a CNN, and later these features and captions are passed through a recurrent neural network (RNN). The model consists of 3 phases:

1) Image Feature Extraction: Because of the VGG 16 model's superior object identification performance, it is used to extract photo features from the Flickr 8K dataset. A convolutional neural network (CNN) of 16 layers, including 2 convolutional layers, 1 dropout layer, and a fully connected layer at the end, is called the VGG 16. [7] The VGG 16 model, which can learn quickly, has dropout layers to prevent the model from overfitting on the training data. The dense layer produces a 4096-dimensional vector representation of the image as its output, which is then transferred to the LSTM layer for additional processing.

2) Sequence Processor: To deal with the input text, there is a sequence processor which serves as a word embedding layer. The embedded layer consists of a mask which is used to disregard padding data and algorithms to extract out the necessary text features. A LSTM is subsequently joined to the network to complete the image captioning process.

3) Decoder: The model's ultimate stage mixes the input from the Image with other data. Utilizing an extractor phase and the sequence processor phase extra action was then transmitted to a layer of 256 neurons, followed by a dense layer that generates a soft-max prediction as the final output over the full lexicon, of the following word in the caption, which was created from the analyzed text data in the processor phase sequence. [3]

A. Training Phase

During the training phase, we provide the image captioning model with two input images and their accompanying captions. Every potential object in a photo must be recognized for the VGG model to function properly. After seeing the image and all of the words before it, the LSTM portion of the model is trained to predict each word in the phrase. To denote the start and finish of the sequence, we add two additional symbols to each caption. The sentence generator halts and marks the string's end when a stop word is encountered. [12] The loss function for the model is calculated as, where 'I' stands for the input image and 'S' for the output caption. The length of the resulting sentence is N.

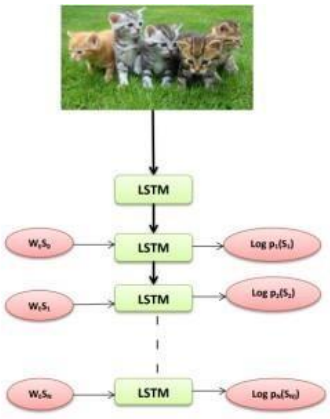

FIGURE 2. Structure of the Model

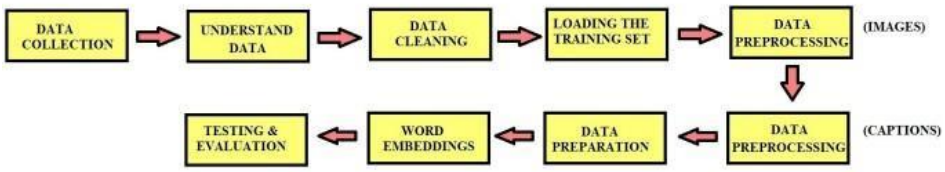

FIGURE 3. Flowchart of the Model

B. Implementation

• The Python SciPy environment was used for the model's implementation. Due to the presence of the VGG net, which was utilized for object recognition, keras 2.0 was utilized to create the deep learning model. The keras framework which is used for creating and training deep neural networks has the TensorFlow library installed as a backend in it. A deep learning library called TensorFlow was created by Google. It offers a platform that is heterogeneous, allowing algorithms to be run on both large-scale

distributed systems with thousands of GPUs and low-power devices like mobile.

• Each compatible device can run a graph once it has been defined. The photo features are computed and saved using the pretrained model. These attributes are then input into our model as the interpretation of a single photo in the dataset, saving us the trouble of having to run each photo through the network each time we want to test a different language model configuration.

• For the real-time implementation of the image captioning model, the picture features are additionally preloaded. [4]

RESULTS & CONCLUSION

After implementing the picture captioning model, we were able to produce captions that were somewhat comparable to those produced by humans. To begin with, the VGG net model assigns probabilities to each object that could potentially be present in the image. The image is transformed into word vector by the model. The LSTM cells receive this word vector as input and use it to construct a sentence. For example, see Fig.5, This gives this image a BLEU score of 57. The BLEU score is a metric used to compare a generated sentence with a reference sentence. A perfect match between the two sentences results in a score of 1.00, whereas an incorrect mismatch leads to a score of 0.00. The authors of a particular study are currently conducting experiments to improve the model's feature extraction in the future. To achieve this, they are exploring the use of alternating pre-trained photo models. Additionally, the scientists intend to use word vectors on a much bigger corpus of data,

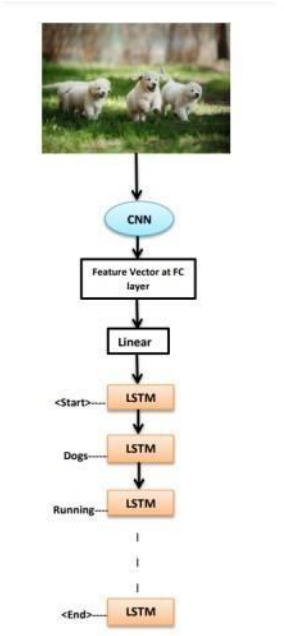

FIGURE 4. Implementation of the Model

such as news stories and other internet sources of data, in order to increase performance. The model's configuration was fine-tuned, however different configurations can be trained to see whether the performance of the picture captioning model can be enhanced.

FIGURE 5. Conversion of an uploaded image into caption

FIGURE 6. Conversion of an uploaded image into caption

FUTURE SCOPE

Image captioning has emerged as a critical problem in recent times due to the rapid proliferation of images across social media and the internet. This report delves into past research on image retrieval, highlighting the various techniques and methodologies employed in such research. Given the difficulties associated with feature extraction and similarity calculation in this domain, there exists significant scope for future research in this field. Presently, image retrieval systems rely on similarity calculations that utilize features like color, tags, histograms, and image retrieval through image captioning. However, these methodologies may not yield entirely accurate results since they do not take into account the context of the image. Therefore, extensive research in image retrieval that leverages the context of images, such as image captioning, could help to address this issue in the future. This project has the potential to be improved by

incorporating more image captioning datasets to enhance the identification of classes with lower precision. Additionally, it can be beneficial to combine this methodology with previous image retrieval techniques, such as histograms and shapes, to evaluate whether it leads to improved image retrieval results. [2]

References

1. Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 6077–6086, 2018.

2. Himanshi Chaudhary, Upendra Mishra, Arpita Gupta, and Amisha Singh. Comparative analysis of rainfall prediction using machine learning and deep learning techniques. In 2022 3rd International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), pages 1–6. IEEE, 2022.

3. Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim. Attend to you: Personalized image captioning with context sequence memory networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 895–903, 2017.

4. Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim. Attend to you: Personalized image captioning with context sequence memory networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 895–903, 2017.

5. Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Show, control and tell: A framework for generating controllable and grounded captions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8307–8316, 2019.

6. Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. From captions

to visual concepts and back. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1473–1482, 2015.

7. Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng. Semantic compositional networks for visual captioning. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 5630–5639, 2017.

8. Yan Huang, Qi Wu, Chunfeng Song, and Liang Wang. Learning semantic concepts and order for image and sentence matching. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 6163–6171, 2018.

9. Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3128–3137, 2015.

10. Andrej Karpathy, Armand Joulin, and Li F Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. Advances in neural information processing systems, 27, 2014.

11. Upendra Mishra, Deepak Gupta, and Barenya Bikash Hazarika. An intuitionistic fuzzy random vector functional link classifier. Neural Processing Letters, pages 1–22, 2022.

12. Vignesh Ramanathan, Armand Joulin, Percy Liang, and Li Fei-Fei. Linking people in videos with “their” names using coreference resolution. In Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13, pages 95–110. Springer, 2014.

13. Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3156–3164, 2015.

Proof of Acceptance of Research Paper

65 of 600

[COPY] Regarding paper decision Inbox ×

ICCS . <iccs@lpu.co.in>
to srivastavasagar2001, tusharpant2001, me

May 1, 2023, 3:00 PM

Dear Sir/Mam,
Greetings from ICCS KILBY 100 conference!

We have reached the final decision regarding your manuscript with ID 1060 to our conference KILBY100 ICCS 2023.

Our decision is to: Accept Submission and it will be submitted to the publication phase.

Submit the camera ready paper on the same mail id. as per the details in Template, Copyright Form and Author Guidelines are available here: https://aip.scitation.org/apc/authors/preppapers

Once you have registered at https://conferences.lpu.in/iccs/RegistrationForm.aspx, please proceed to make the payment as soon as possible and share the proof regarding the same to ensure a successful submission at iccs@lpu.co.in

If you have any queries or require assistance at any stage of this process, feel free to contact us. We will be more than happy to help.
Warm Regards,

Team ICCS
KILBY100
7th International Conference on Computing Sciences 2023

Reply Reply all Forward