

UNITEDWORLD INSTITUTE OF TECHNOLOGY (UIT)

Summative Assessment (SA)

Submitted BY

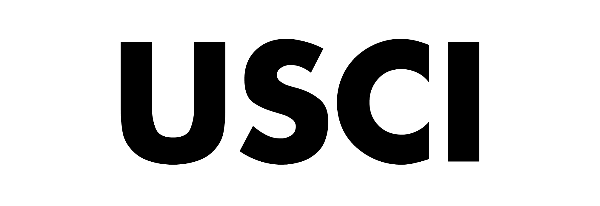
Manav Patel

(Enroll. No.: 20220701061)

**Course Code and Title: –**

B.Sc. (Hons.) Computer Science / Data Science / AIML

IV Semester – Dec – May 2024



Dec/May 2024

Table of Contents

[Image Caption Generator using Machine Learning 3](#_Toc162833670)

[What is Image to Caption Generator? 4](#_Toc162833671)

[What is CNN? 5](#_Toc162833672)

[What is LSTM? 5](#_Toc162833673)

[Dataset for Image Caption Generator 6](#_Toc162833674)

[Building the Image Caption Generator 7](#_Toc162833675)

[Import all the Required Packages 7](#_Toc162833676)

[Perform Data Cleaning 8](#_Toc162833677)

[Extract the Feature Vector 11](#_Toc162833678)

[Loading dataset for model training 12](#_Toc162833679)

[Create a Data generator 13](#_Toc162833680)

[Define the CNN-RNN model 14](#_Toc162833681)

[Training the Image Caption Generator model 15](#_Toc162833682)

[Testing the Image Caption Generator model 15](#_Toc162833683)

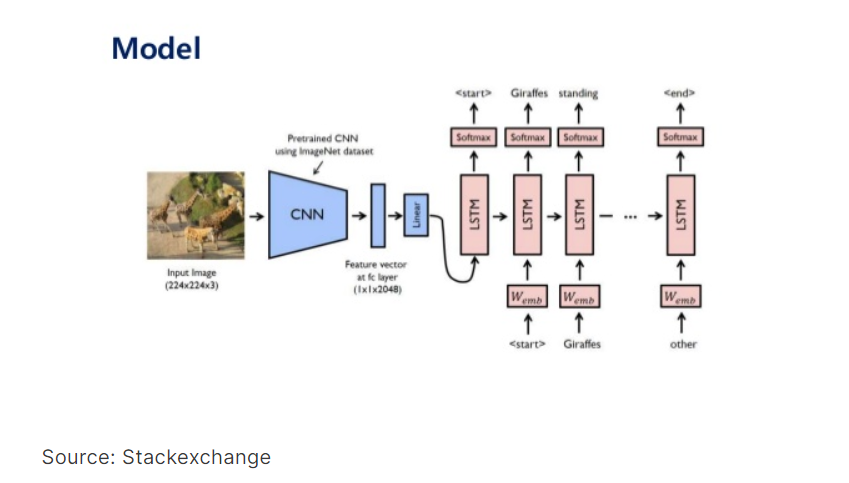
[End Note 19](#_Toc162833684)

# Image Caption Generator using Machine Learning

The basis of our brains is annotating or classifying any images that are demonstrated to us. What about computers, though? How can a device comprehend an image and place a caption that is highly accurate and has appropriate context terms? To say that it was hard to believe was absolutely true only a few years ago. However, there is a continuous advancement in the general-purpose image captions generator because of the progress in Computer Vision and Deep Learning algorithms, also relevant data sets and AI models are now available. The circle of captioning around the world is constantly widening, with many data annotation companies increasing their profit margins every year. By the aid of datasets, we will build a specimen with such properties as to give a useful description of the image in this tutorial. LSTM and CNN technologies will be discussed briefly and also covers a general description about the concepts of Deep learning approaches.

## What is Image to Caption Generator?

With the aid of deep learning and computer vision, an image caption generator will be able to interpret the image and provide relevant captions. It utilizes an image set provided during the model-training to assign these keywords in English. The CNN model Xception is trained using the ImageNet dataset. Xception does the features extraction of images. The LSTM model in charge of the image description will receive these feature vectors as the input.



## What is CNN?

CNN, which is a sub-discipline of deep learning, is an approach used to detect and classify images. It deals with the raw data in the form of matrices that looks like two dimensional pictures. CNN is able to deal with images that are translated, rotated or scaled. It locates important features of the image by looking from top to bottom and from left to right. Therefore, it unites all the components of image categorization in the end.

## What is LSTM?

Sequence prediction problems can be tackled via the use of LSTM (Long Short-Term Memory), which is a type of RNN (Recurrent Neural Network). The main purpose of that model is predicting the next word, as our system takes into account the Google search results determined by the previous text. LSTM has been designed to filter the significant data and eradicate the redundant data as the inputs are processed.  
Consequently, CNN and LSTM has to be incorporated for building image caption generator model.

**CNN + RNN make the CNN-RNN model for image caption generators.**  
  
**CNN:** The first stage involves the description of the image. Therefore, an Xception model that is already trained will be involved.

**LSTM:** To build a description based on information that was disentangled from the image.

## Dataset for Image Caption Generator

Flickr8K is the model training dataset setting of the caption generators. These datasets can be retrieved through links given below. As the dataset is big (1GB in size), the downloading being slow takes some time to complete. Scroll through every file in the Flickr\_8k\_text folder in the image below depicted. Flickr 8k.token, which contains all image titles and captions, the dominant one in the provided files. Folder Flickr\_8k\_text has text files with image captions in it, whereas image dataset of 8091 photos is stored inside the Flickr\_8k\_Dataset folder.

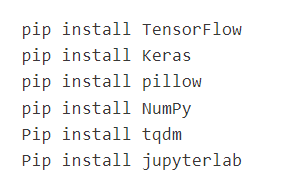
* [Flicker8k\_Dataset](https://github.com/jbrownlee/Datasets/releases/download/Flickr8k/Flickr8k_Dataset.zip)
* [Flickr\_8k\_text](https://github.com/jbrownlee/Datasets/releases/download/Flickr8k/Flickr8k_text.zip)

##### Pre-requisites

The implementation of the caption generator will be done through the utilization of Jupyter notebooks. The Jupyter notebooks described in the article can be found in the above page. The development of these skills will be vital to the understanding, on a deeper level, of Python, Deep Learning and NLP. In that case if this is new for you. I would like to firstly direct you to the link which I have already provided.

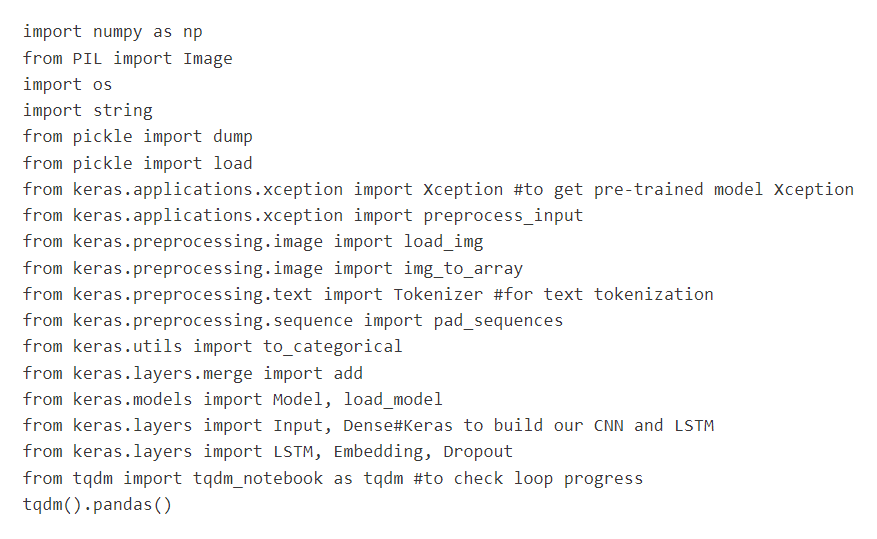
* [Python](https://www.analyticsvidhya.com/blog/2016/01/complete-tutorial-learn-data-science-python-scratch-2/)
* [DeepLearning](https://www.analyticsvidhya.com/blog/2021/05/a-comprehensive-tutorial-on-deep-learning-part-1/)
* [NLP](https://www.analyticsvidhya.com/blog/2017/01/ultimate-guide-to-understand-implement-natural-language-processing-codes-in-python/)

Install below libraries

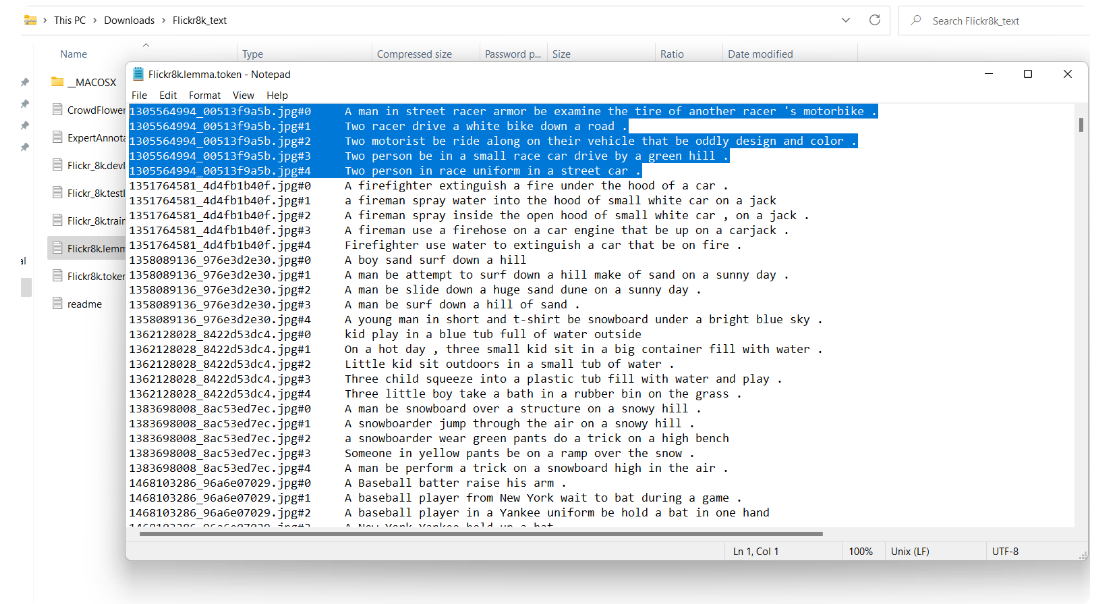


# Building the Image Caption Generator

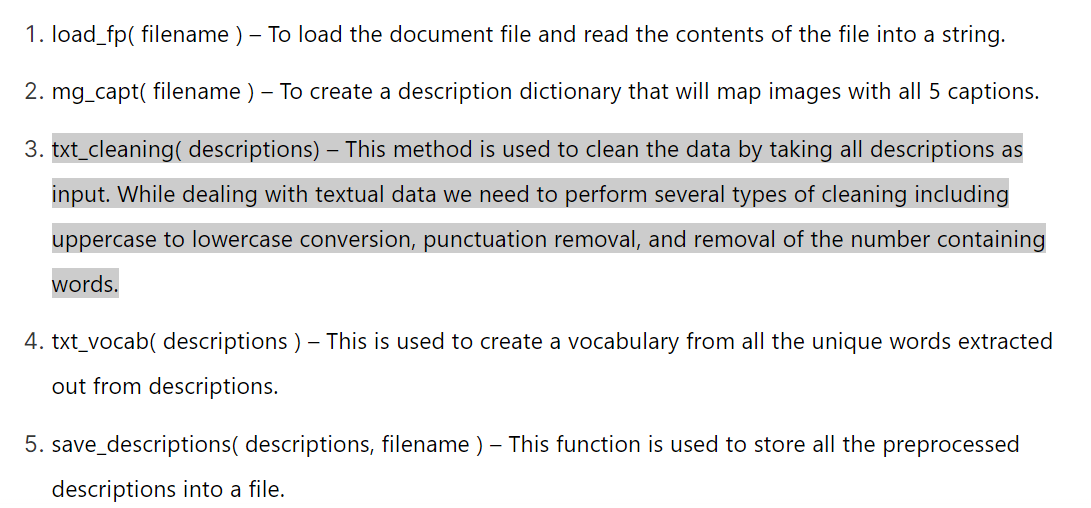
## Import all the Required Packages

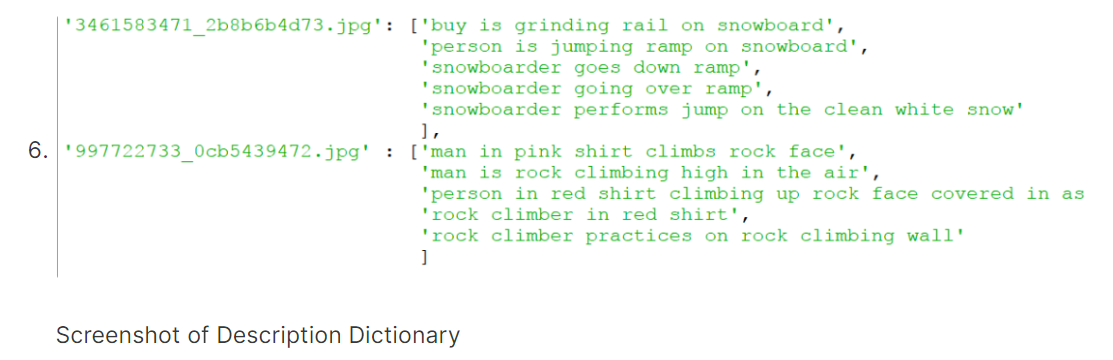


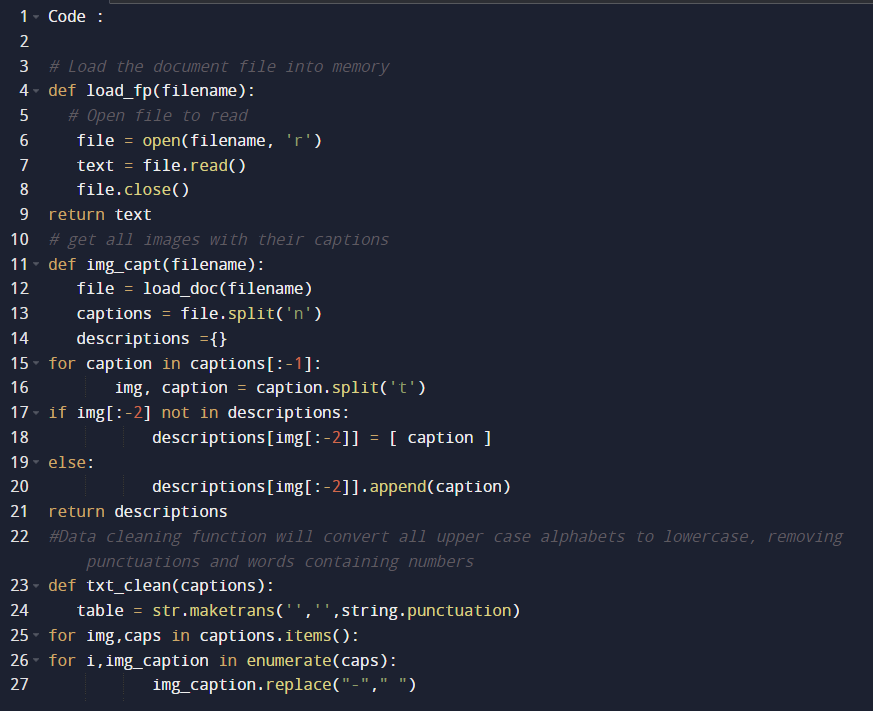
## Perform Data Cleaning

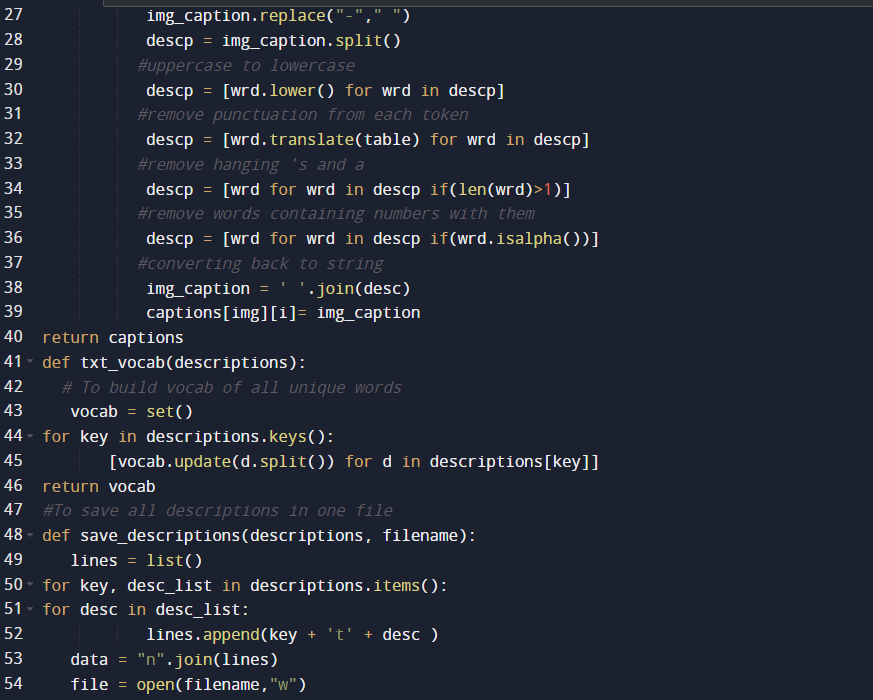


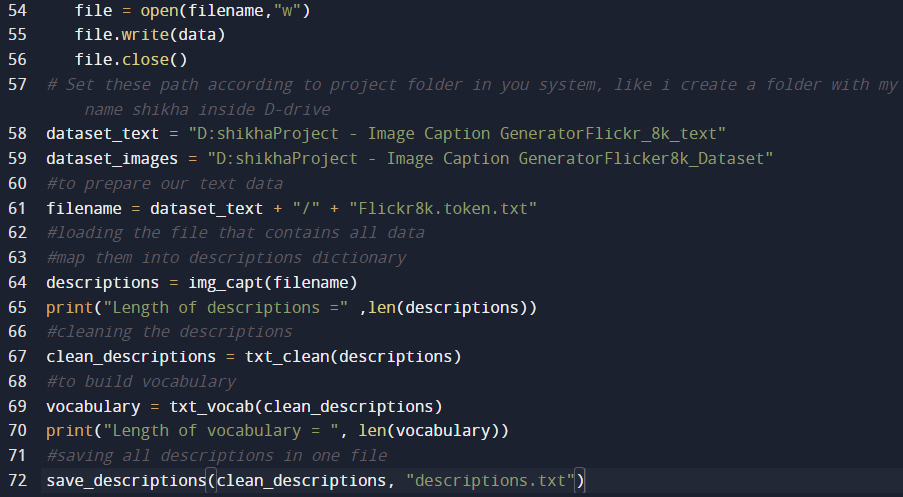
We will now define the following five cleaning functions :







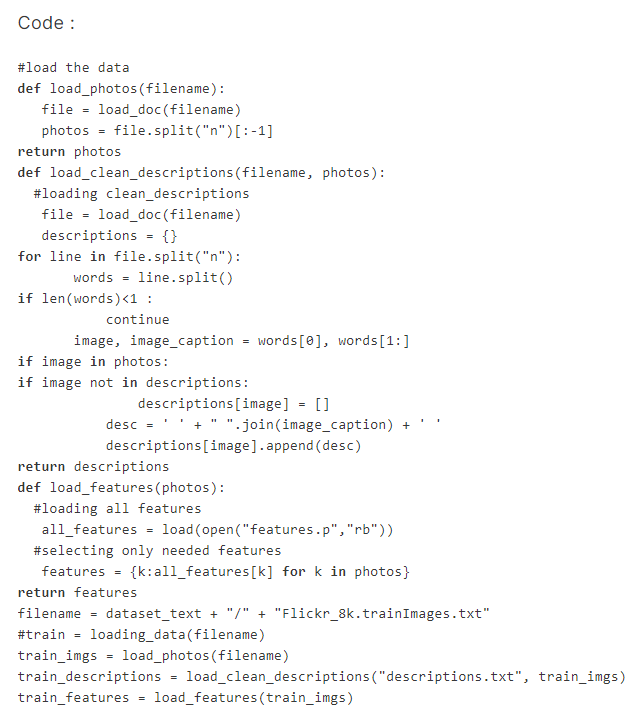




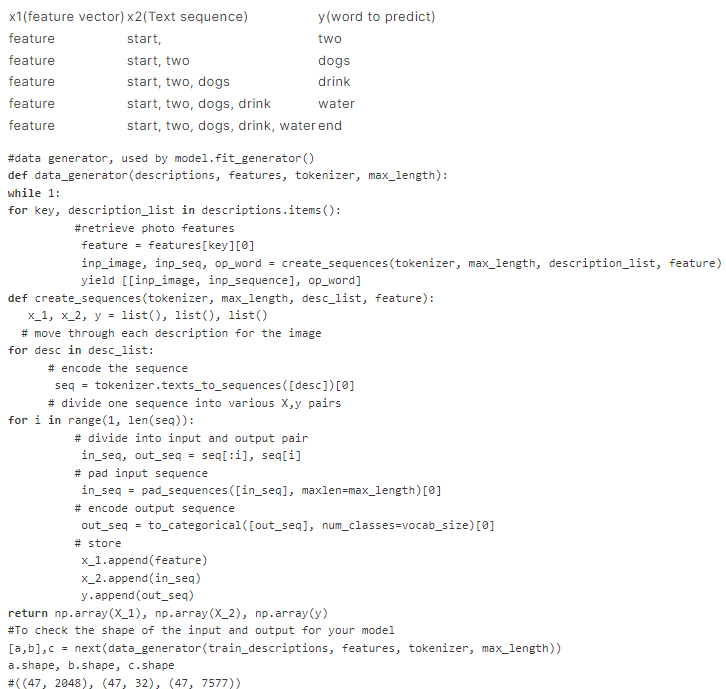
## Extract the Feature Vector

is called Xception and it has been through massive datasets for training. To classify the photos, Xception’s training was carried out on imagenet dataset, which included 1,000 diverse classes. In a rather straightforward manner, the model in use can be imported by using keras.applications. It becomes necessary to make few alterations to the model Xception to integrate it with our framework. As xception model, demands an image size of 299 To extract the general features from the models, we can now use the model for pre-training which \*299\*3 in input, we need to take away the last classifier layer and then extract 2048-feature vector from it.  
  
  

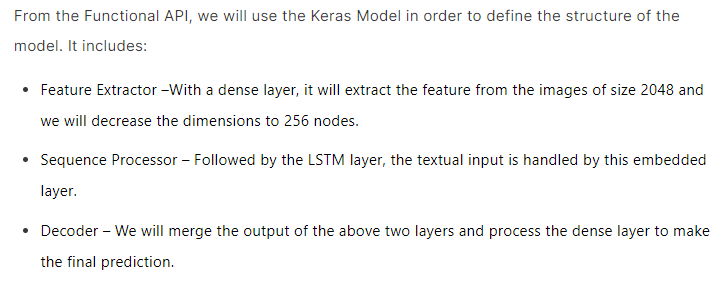

## Loading dataset for model training

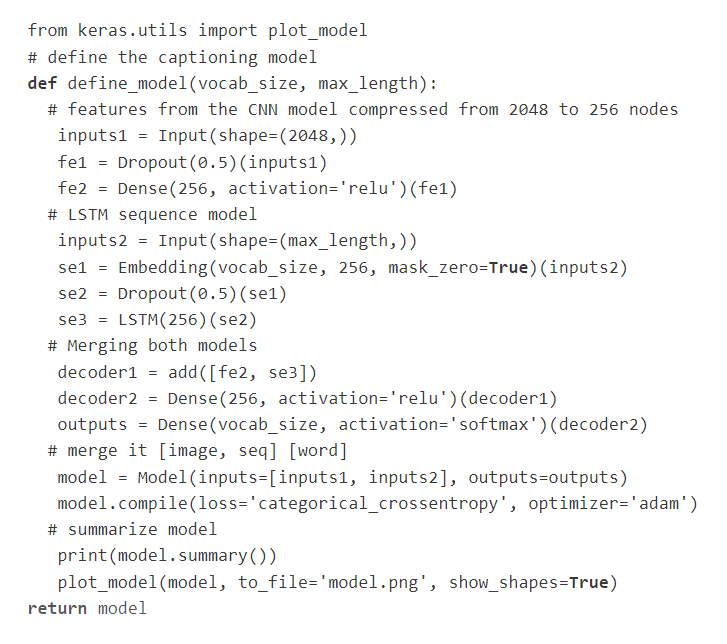
Our Flickr\_8k\_test folder contains the file "Flickr\_8k.trainImages.txt". In this file, there are 6000 names of pictures which mostly serve for training purposes.  
**def load\_photos(fname):** — This function considers a text file as an input, read from a file name, and return it as a list of image names.  
**load\_clean\_descriptions(fname, images)** – This is a function that creates a dictionary with captions for each picture in a list. To define when the caption starts and ends, we add and to every caption that will be presented to the LSTM network.  
**load\_features(photos)** – Function that returns the record of photos together with the features vectors calculated by the Xception model.  


## Create a Data generator

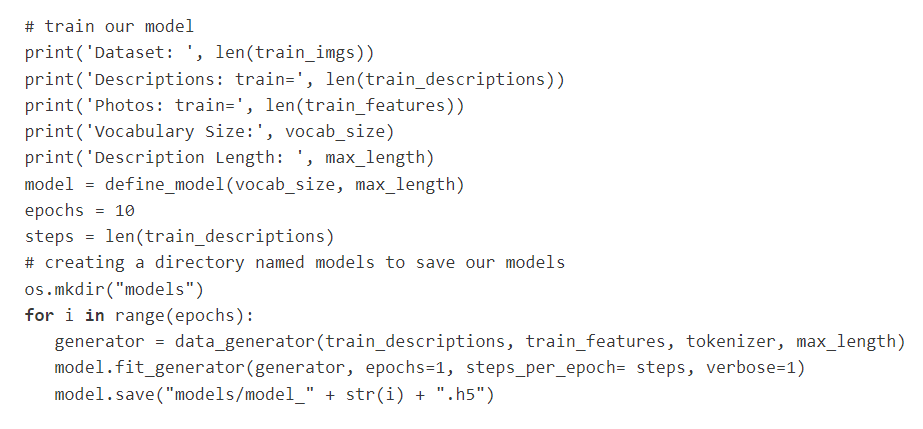


## Define the CNN-RNN model

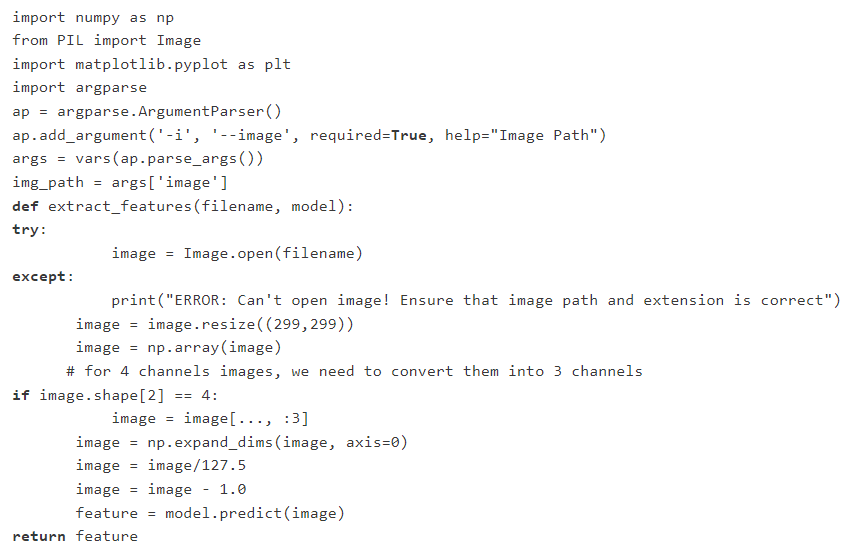


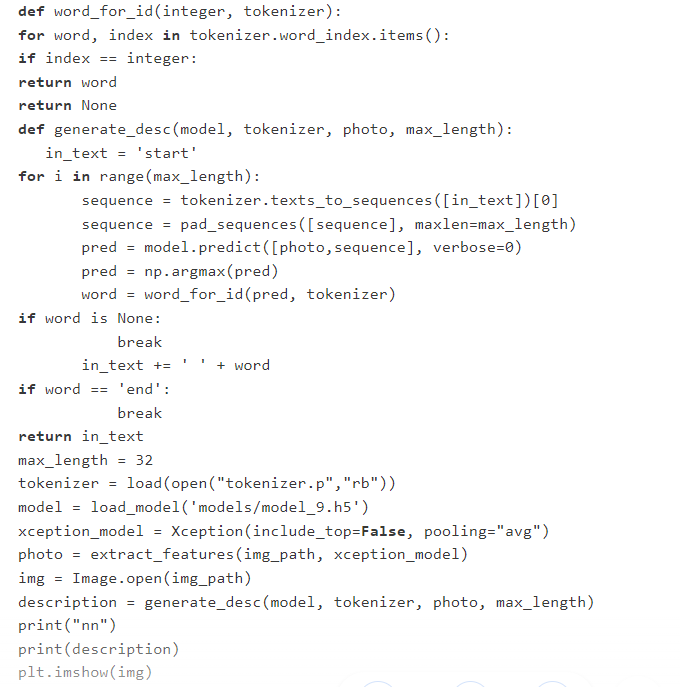


## Training the Image Caption Generator model

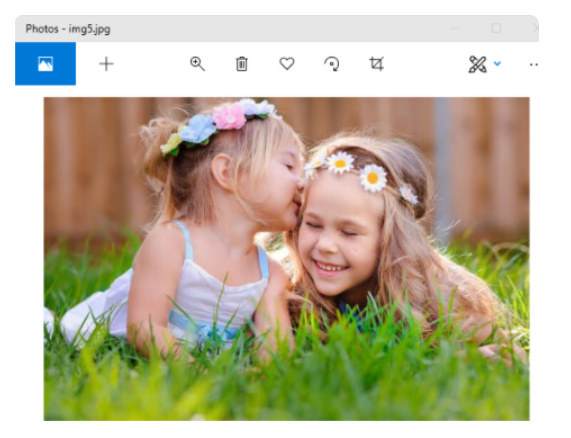


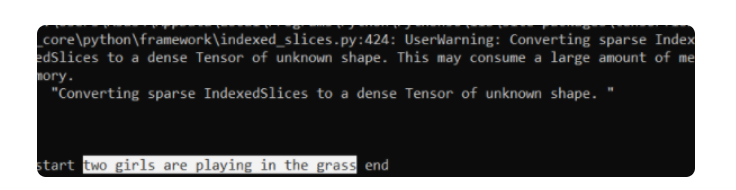
## Testing the Image Caption Generator model

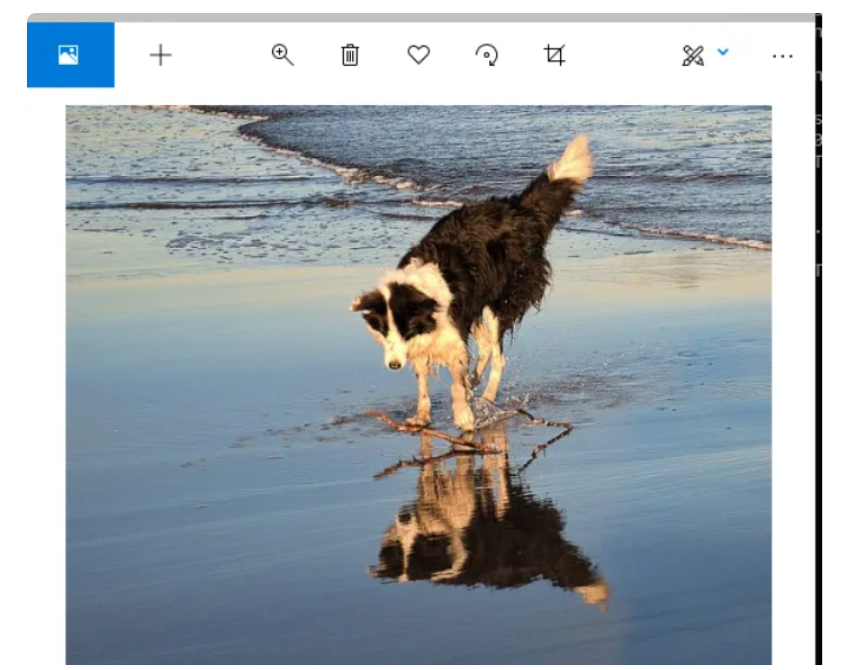


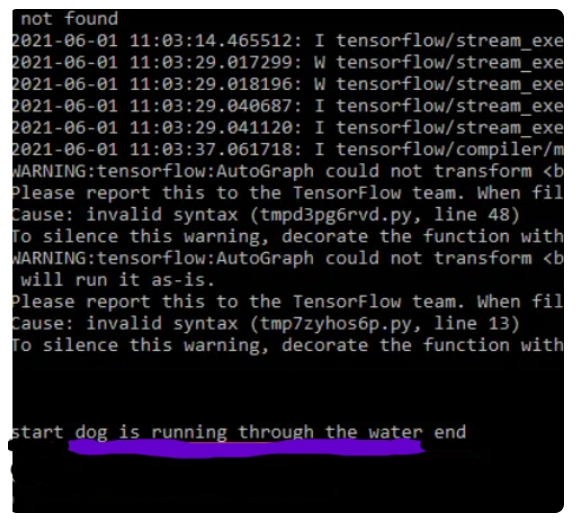


OUTPUT :









## End Note

Within this paper, the combined usage of a convolutional neural network (CNN) and long short-term memory network (LSTM) is demonstrated through which an image caption generator is developed using deep learning method. The initiation of work started with an understanding of the image processing capability of CNNs and the ability of LSTMs to predict sequences. To start with, we bestowed a synergistic CNN-RNN architecture with these models to constitute the fundamental unit of our picture captioning model.  
  
Preparing and preprocessing of data for the training of the model is the key part of this project execution. We controlled the information flow by providing a sensible data which has been cleaned, extracted, and tokenized. This set a firm basis for model's more training.  
  
The design and overhauling the architecture was crucial in our work. Within the Keras library, we then elaborately came up with a very complex neural network architecture before giving the "readymade" dataset to train it. With multiple training sessions containing importance on adjusting parameters and accustoming the model, an image caption generator that performed well was eventually created.  
  
The previous phase revealed the ability of our trained model to generate captions --- linguistically coherent, contextually-relevant, and varied. The process of testing and validation on the data it has not seen before confirmed this ability. This empirical evidence testifies our way to be used in numerous style computers vision and natural language processing tasks and affirms its reliablity.

## Reference

<https://www.analyticsvidhya.com/blog/2021/12/step-by-step-guide-to-build-image-caption-generator-using-deep-learning/>