**IMAGE CAPTION GENERATION USING DEEP LEARNING**

Project submitted to the

SRM University – AP, Andhra Pradesh

for the partial fulfillment of the requirements to award the degree of

**Bachelor of Technology**

**In**

**Computer Science and Engineering**

**School of Engineering and Sciences**

Submitted by

G. Dhakshayani - AP2110010930

* K. Venkatesh - AP21110010933
* Shubham Pandey - AP21110010940
* Ch. Sai Kumar - AP21110010943
* I. Anu Likitha - AP21110010963

**A picture containing text

Description automatically generated**

Under the Guidance of

**(Radha Guha )**

**SRM University–AP**

**Neerukonda, Mangalagiri, Guntur**

**Andhra Pradesh – 522 240**

**Dec, 2023**

# Certificate

Date: 2-Dec-23

I hereby aﬃrm that the project titled "Image Caption Generation" has been conducted by the following individuals under my guidance:

Venkata Naga Saikumar Chunduru - AP21110010943

I certify that the work submitted is legitimate, unique, and suitable for submission to SRM University – AP in order to complete the requirements for the School of Engineering and Sciences' Bachelor of Technology and Master of Technology degrees.

Supervisor (Signature)

Dr. Radha Guha

Designation, Aﬃliation.

# I. Acknowledgments

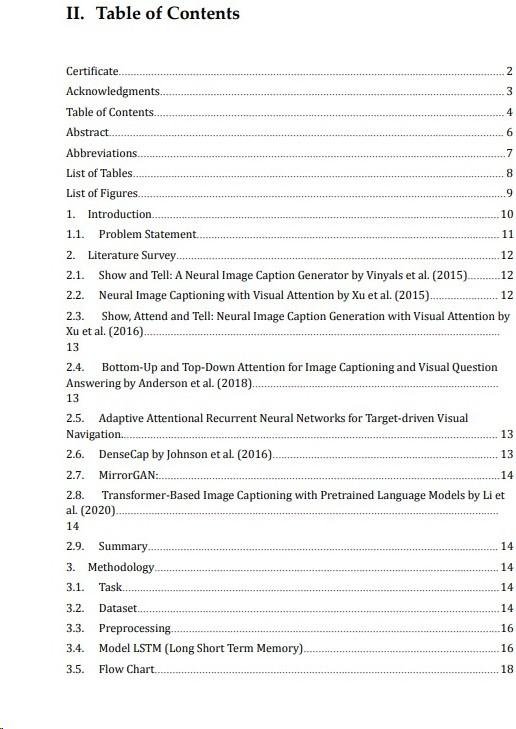
We sincerely thank Dr. Radha Guha, our project guide, for her leadership, mentoring, and steadfast support during the project. Her knowledge and perceptions have been invaluable in determining the course and results of our effort.

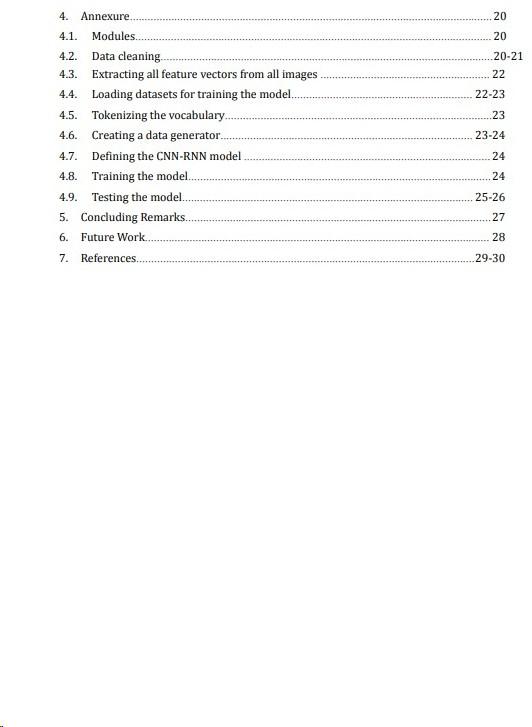
Additionally, we would like to thank the entire project team for their cooperation, hard work, and dedication. Our project's objectives have been met in large part because of the dedication and hard work of every team member, which has also improved and enhanced the working environment.

Our heartfelt thanks go to the esteemed faculty members of SRM AP for their invaluable guidance, support, and commitment throughout our academic journey. Their dedication to excellence, passion for teaching, and willingness to go the extra mile have had a signiﬁcant impact on our education.

We would also want to thank our family and friends for their patience, inspiration, and steadfast support during the project. Their conﬁdence in us and support through diﬃcult times have served as a constant source of inspiration.

Lastly, we appreciate everyone who supported us directly or indirectly in successfully completing this project.





# IV. Abstract

The advanced deep-learning model known as the image caption generator has revolutionized the generation of captions for digital images. In this process, a

convolutional neural network (CNN) is employed to extract visual features from images, followed by the utilization of a long short-term memory (LSTM) network to create a sequence of words forming the image caption. This innovation has significantly

streamlined the previously labor-intensive task of manually generating captions for images.

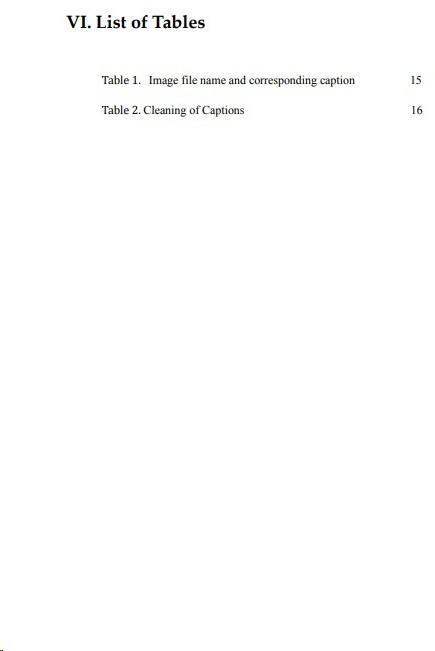
The applications of the image caption generator extend across various domains such as image and video search, social media, and content creation. Its automated generation of descriptive and informative captions contributes to improved search engine

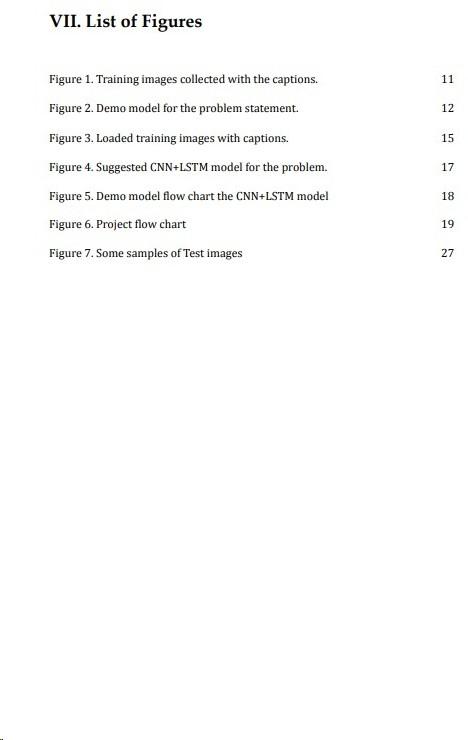
optimization, heightened user engagement, and increased visibility of content. Additionally, the model proves beneficial in automatically generating video subtitles, serving as a valuable asset for video content creators. The primary strength of the image caption generator lies in its proficiency to generate captions that precisely depict the

contents of the images.

The model undergoes training on an extensive dataset comprising images and their corresponding captions. This training enables the model to discern the relationships between visual features and text descriptions, resulting in the generation of captions that are not only descriptive but also contextually relevant and meaningful.







# Introduction

The image caption generator represents a computer vision technology designed to automatically provide textual descriptions for images. This technology has garnered significant attention due to the growing demand for visual content and the necessity for effective descriptive methods. The applications of image caption generators are widespread, spanning various industries such as social media, e-commerce, healthcare, and education. Notably, platforms like Instagram and Facebook employ image captioning to enhance accessibility for individuals with visual impairments, while e-commerce utilizes it to furnish precise product descriptions, thereby enhancing customer experiences and boosting sales.

The image caption generator integrates computer vision, natural language processing, and machine learning techniques. Computer vision identifies objects, people, and actions within an image, natural language processing generates accurate textual descriptions, and machine learning enhances caption accuracy through training on extensive datasets with annotated images. A significant challenge in image captioning is the generation of captions that are both precise and semantically meaningful. This entails providing insightful descriptions that capture diverse aspects of an image. As advancements in computer vision and machine learning continue, the technology is expected to evolve, becoming more efficient, accurate, and accessible in the coming years.



Fig1: Training images collected with the captions.

## Problem Statement

The task of image caption generation includes generating a natural

language resulting from an image. The aim is to create an automated system that can know the contents of an image that produces a

coherent and meaningful sentence that accurately describes the visual content. The problem is quite challenging due to the hardness level of the natural language, the high dimensionality of images, and the need to incorporate contextual information to generate accurate captions.

The image caption generator needs to be trained on more datasets of images along with their captions to develop an understanding of the relationships between visual features and language. The generated

captions should be informative, relevant, and semantically correct, reflecting the content of the image in a human-like way.

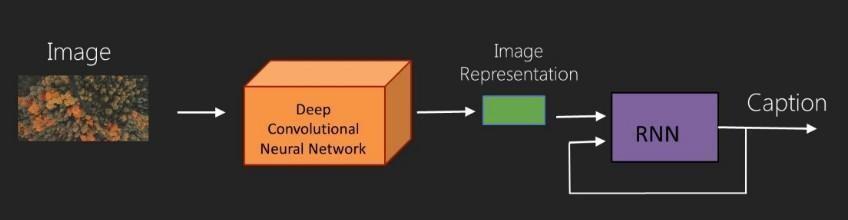


Fig2: Demo model for the problem statement.

# Literature Survey

Research in natural language processing and computer vision has focused on creating captions for images. There are several practical applications for this activity, including picture search, image indexing, and assistive technology for the blind. It entails creating a natural language description of an image's content. Among the most signiﬁcant works on this subject are:

**Vinyals et al. (2015) present Show and Tell:** A Neural Image Caption Generator. This study presented a model that generates captions using a Recurrent Neural Network (RNN) and codes images using a Convolution Neural Network (CNN). At the time, the model produced cutting-edge ﬁndings after being trained on a sizable dataset of image-text pairings.



**Xu et al. (2016) presented Show, Attend, and Tell:** Neural Image Caption Generation with Visual Attention. Through this effort, a novel model architecture combining top-down and

bottom-up attention is built. The model creates subtitles using an LSTM-based decoder and a CNN network to identify viewable regions of the picture.

**Bottom-Up and Top-Down Focus for Visual Question Answering and Image Captioning by Anderson et al. (2018):** In addition, a model architecture combining top-down and bottom-up attention was established in this work, employing an LSTM-based decoder to produce subtitles and a CNN network to identify visible regions of the picture.

**Target-driven Visual Navigation with Adaptive Attentional Recurrent Neural**

**Networks:** This research created a paradigm for virtual world navigation subtitle creation. In order to enable the model to analyze different regions of the image selectively, the researchers developed an adaptive attention mechanism that is dependent on the target's location. The model uses an LSTM-based decoder to generate subtitles.

**DenseCap, Johnson and colleagues (2016):** A method for writing lengthy titles that encompass all items and all of their attributes was presented in this study. The authors employed an LSTM-based decoder to generate subtitles and CNN to identify the items in the picture.

**MirrorGAN:**

Learning text-to-image creation by transformation, Guo et al.

(2019) - this work proposed a new approach to captions, which involves generating images from text descriptions and then

generating captions for the generated images. The authors showed that this approach leads to more accurate and diverse subtitles.

## Transformer-Based Image Captioning with Pretrained Language Models by Li et al. (2020)

In this work, a transformer-based model using pre-trained language models was proposed for captions. The authors showed that this approach leads to better performance in several benchmarks.

## Summary

In summary, these works have made significant contributions to the development of image caption generators, with some of them utilizing attention mechanisms, transformer architectures, and LSTM-based decoders. Our work builds on these foundations by proposing a novel approach that leverages multiple modalities and combines pre-trained language models with fine-tuning on image captioning datasets.

# Methodology

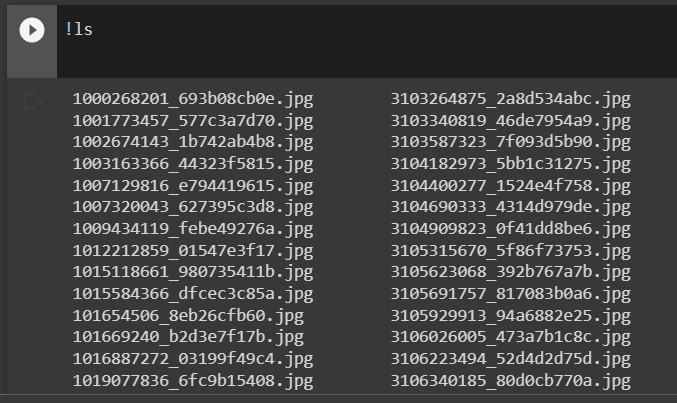
## Task

The objective is to create a system that can receive an image input as a dimensional array and produce a grammatically correct sentence that describes the image as its output.

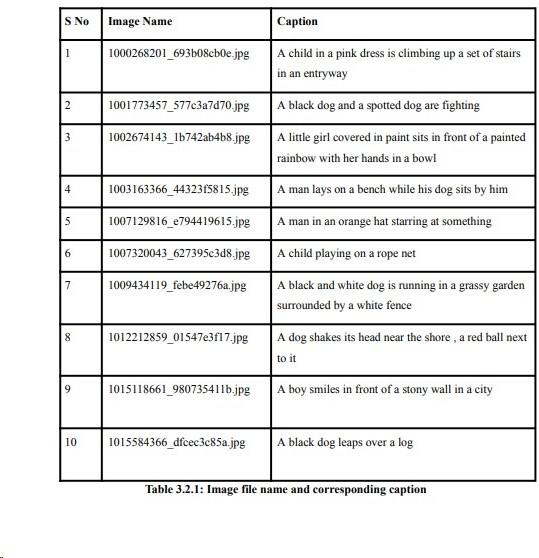
## Dataset

The corpus utilized for this task is the Flickr 8K dataset, which

comprises a total of 8000 images. Dataset contains the image file name and corresponding caption of that image.



**Fig3: Loaded training images with captions.**



## Loss function

In deep learning, the error between predicted and actual values is computed by the loss function, which is crucial for assessing how well a neural network performs in tasks like

regression or classiﬁcation. In order to improve accuracy, this function is reduced during the model's training. Forward propagation, in which input data travels through the network, and backpropagation, in which the model's weights and biases are modiﬁed to minimize the

difference, are the two steps in the calculation of the loss function. Various loss functions are used depending on the task and type of network, including Mean Squared Error (MSE) and

Cross-entropy loss. In order to get a single loss value that represents the performance of the model, the loss function is calculated for each data point during training and averaged.

A loss function used in deep learning for multiclass classiﬁcation issues is called categorical cross-entropy. It is computed as the sum of individual losses for each class label per observation and is also referred to as SoftMax Loss. For a given set of events, the loss

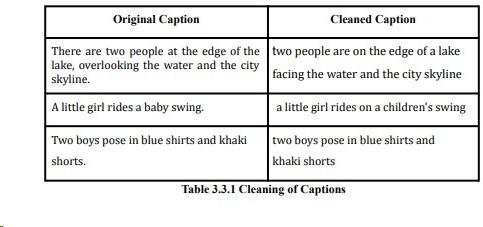
function calculates the difference between two probability distributions. In order to minimize

the cross-entropy, the model weights are iteratively changed during training. Depending on the particular task and network type, different cross-entropy loss functions are used, such as Binary Cross-Entropy Loss and Multinomial Logistic Loss.

**Value of (Y (actual)-Y (predicted)) equals loss.**

## Preprocessing

To prepare the data for the image captioning system, a two-part preprocessing approach was adopted. First, the images and their corresponding captions were cleaned and preprocessed independently. The image data was processed by utilizing the Exception application from the Keras API, which is pre-trained with the ImageNet dataset. This approach facilitated faster training of the images with the help of transfer learning. On the other hand, the captions were cleaned using the tokenizer class available in Keras, which enabled the vectorization of the text corpus and the storage of the resulting vectors in a separate dictionary. Each word of the captions was then processed further to obtain a representation suitable for use in the image captioning model.



## Model LSTM (Long Short Term Memory)

Recurrent neural networks (RNNs) of the Long Short-Term Memory (LSTM) type can recognize long-term dependencies in sequential data. It is appropriate for challenging problem domains like speech recognition and machine translation because it solves the

vanishing gradient issue that traditional RNNs have and may lead to long-term dependence.

Because LSTM can store information for a long time, it is a better option for tasks like speech recognition, language translation, and time series forecasting than traditional RNNs. This is accomplished by means of "gates," such as the input, output, and forget gates, which choose whether to read, output, or forget the current cell value. The hidden states in LSTM are

crucial, as they are transmitted to the next layer and serve as the neural network's memory, enabling it to function effectively in sequential processing tasks.

A popular method for captioning images is to combine convolutional neural networks (CNN) and long short term memory (LSTM) to take an image as input and output a caption. This

enables the model to produce a logical and evocative sentence by efficiently processing the sequential information from the caption and the visual information from the image.

All things considered, LSTM is a strong tool for sequential data processing. It can be utilized

successfully for many different applications, including picture captioning, when coupled with other neural network architectures like CNN.

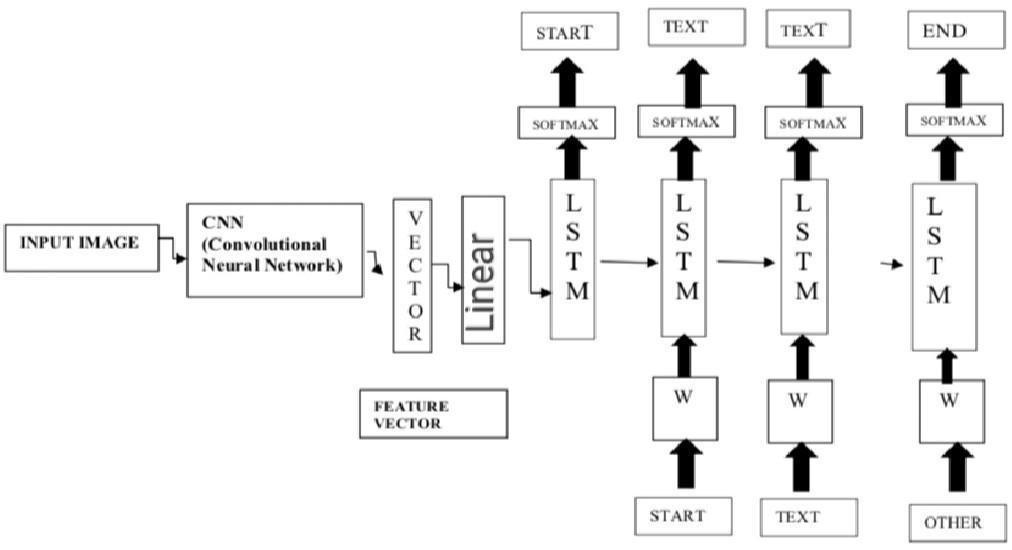


Fig4: Suggested CNN+LSTM model for the problem.

In an image captioning system, the source sentence—typically the image—is mapped by the "encoder" RNN and converted into a fixed-length vector representation. The "decoder" RNN then uses this representation as its hidden initial state to predict the final meaningful sentences. But instead of employing an RNN as the "encoder," a deep Convolutional Neural Network (CNN) embeds the input image into a fixed-length vector to produce a rich representation of the image. The CNN is used as the input for the "decoder" RNN after being trained for the image classification task. It has been demonstrated that by utilizing the advantages of both CNNs and RNNs, this method enhances the performance of image captioning systems.

## 

## 

## Fig5: Demo model flow chart the CNN+LSTM model

## 3.6 Project Architecture

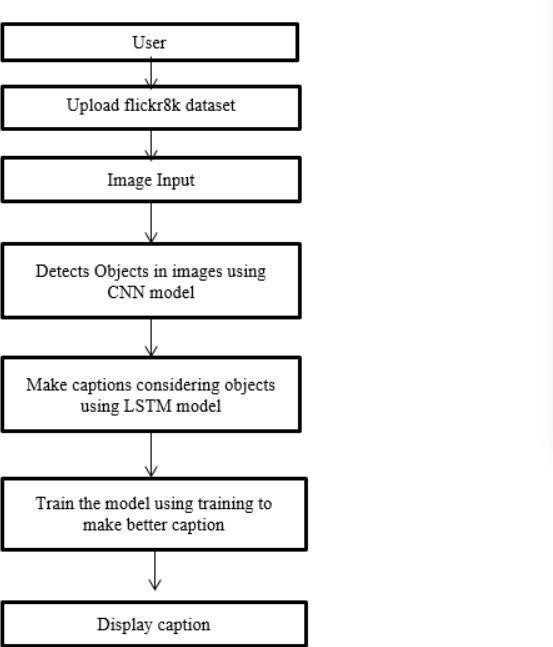


Fig 6. Project flow chart

To build a stacked model, we will utilize the three main components of the Keras Model from the Functional API. The input data will first be reduced in dimension

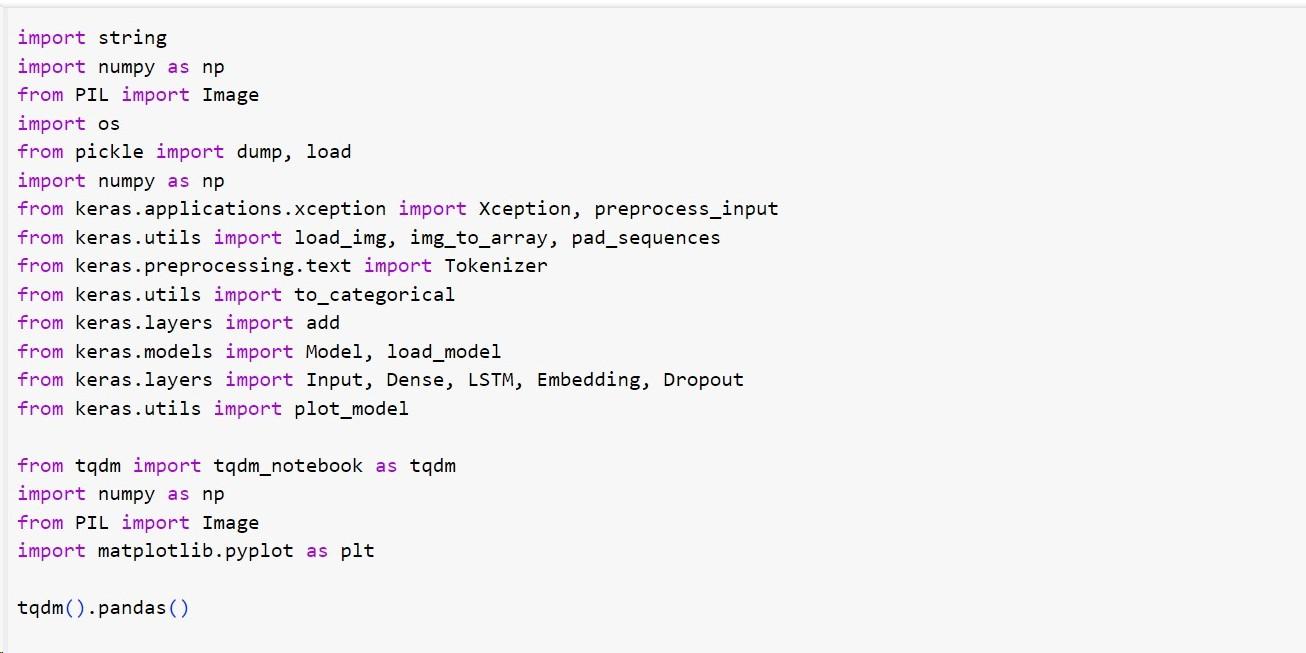
from 2048 to 256 using a feature extractor. To improve generalization, a dropout layer will then be added to the CNN and LSTM. The Xception model, with the exception of the output layer, has been applied to the photos in order to prepare them for feature

extraction. The extracted features will be the input. Second, an LSTM layer will be used after an embedding layer to process the textual input. The final predictions will be generated by combining the outputs of the sequence processor and feature

extractor using a Dense layer.The number of nodes in the final layer is not specified.

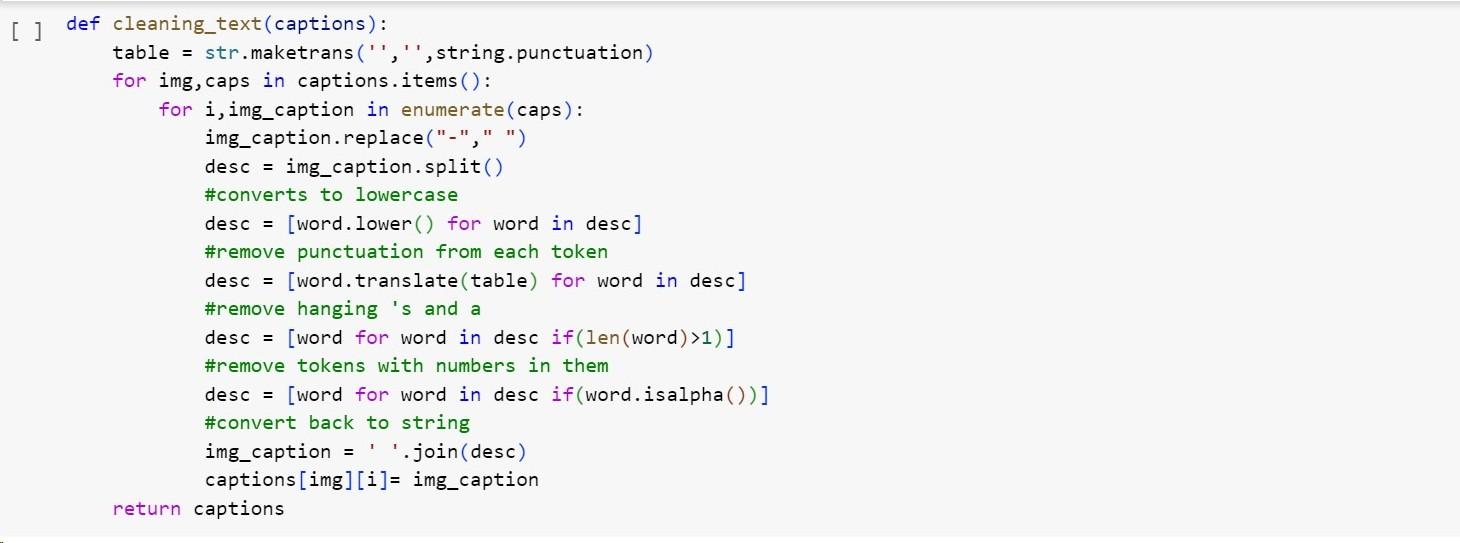
# Annexure

## 4.1. Modules



**4.2 . Data cleaning**







## Extracting all feature vectors from all images

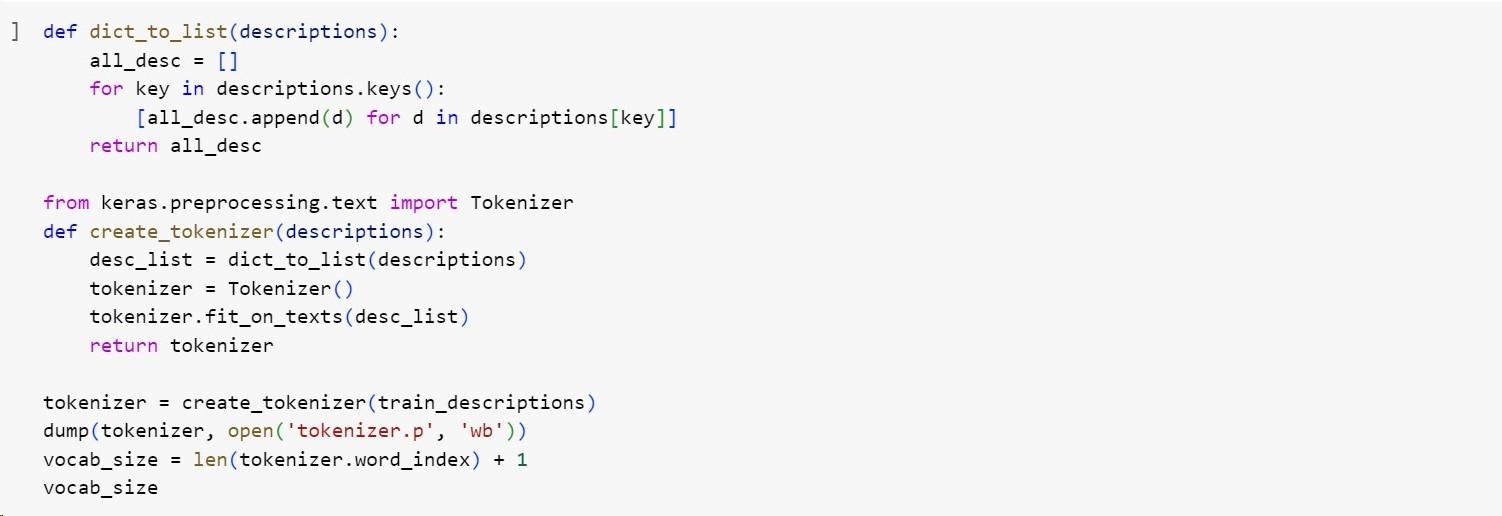


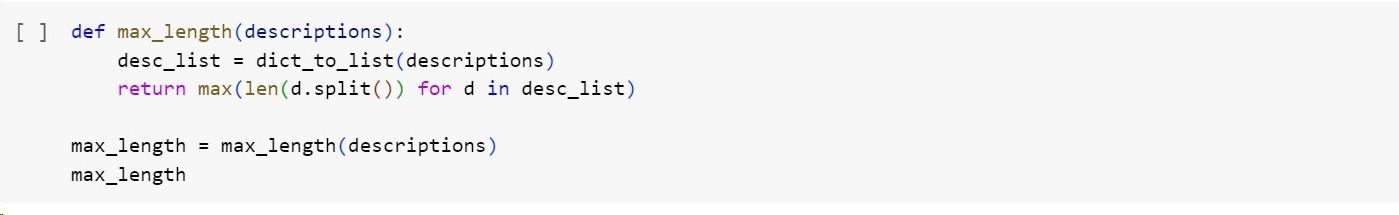
* 1. **Loading the Datasets for training a model**



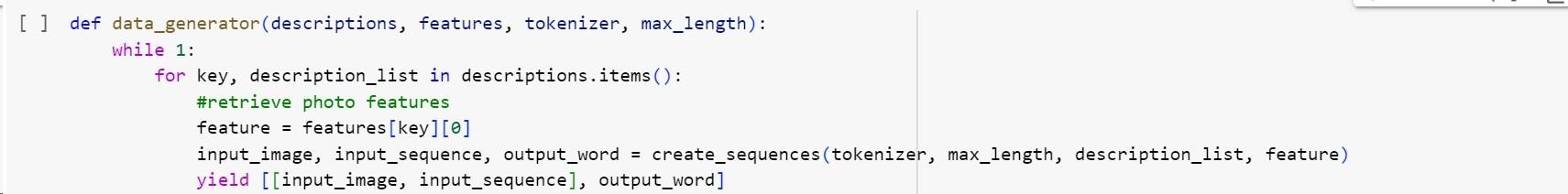


## Tokenizing the Vocabulary





* 1. **Creating a Data Generator**

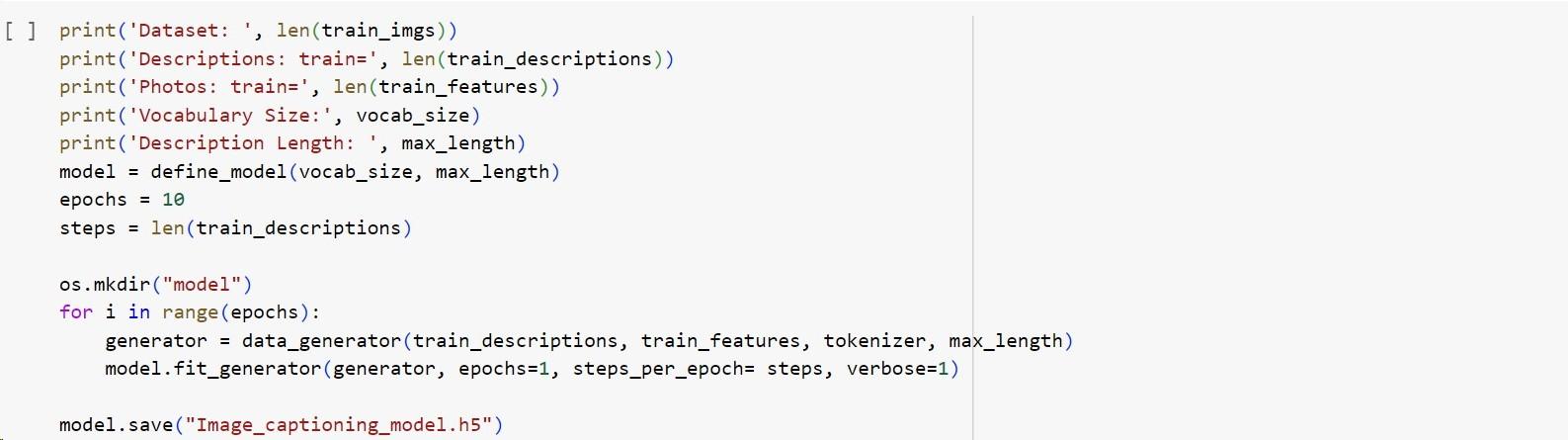




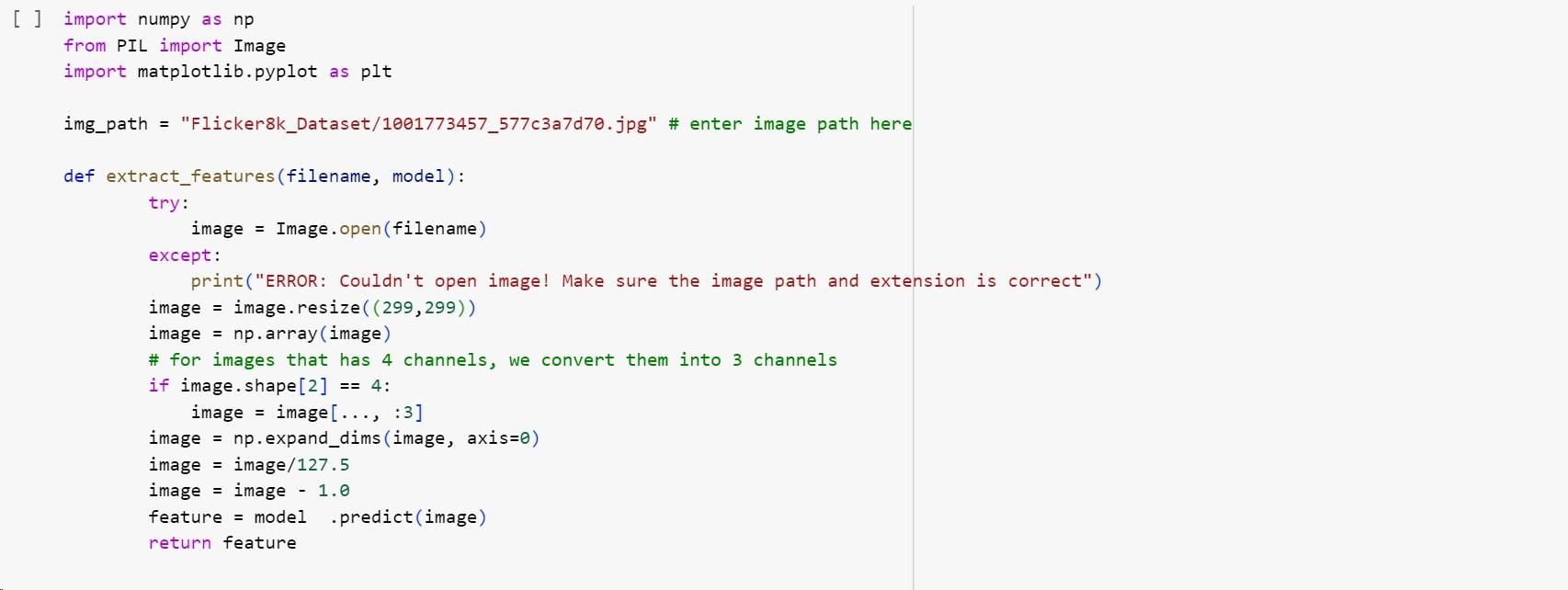
## Defining the CNN-RNN model

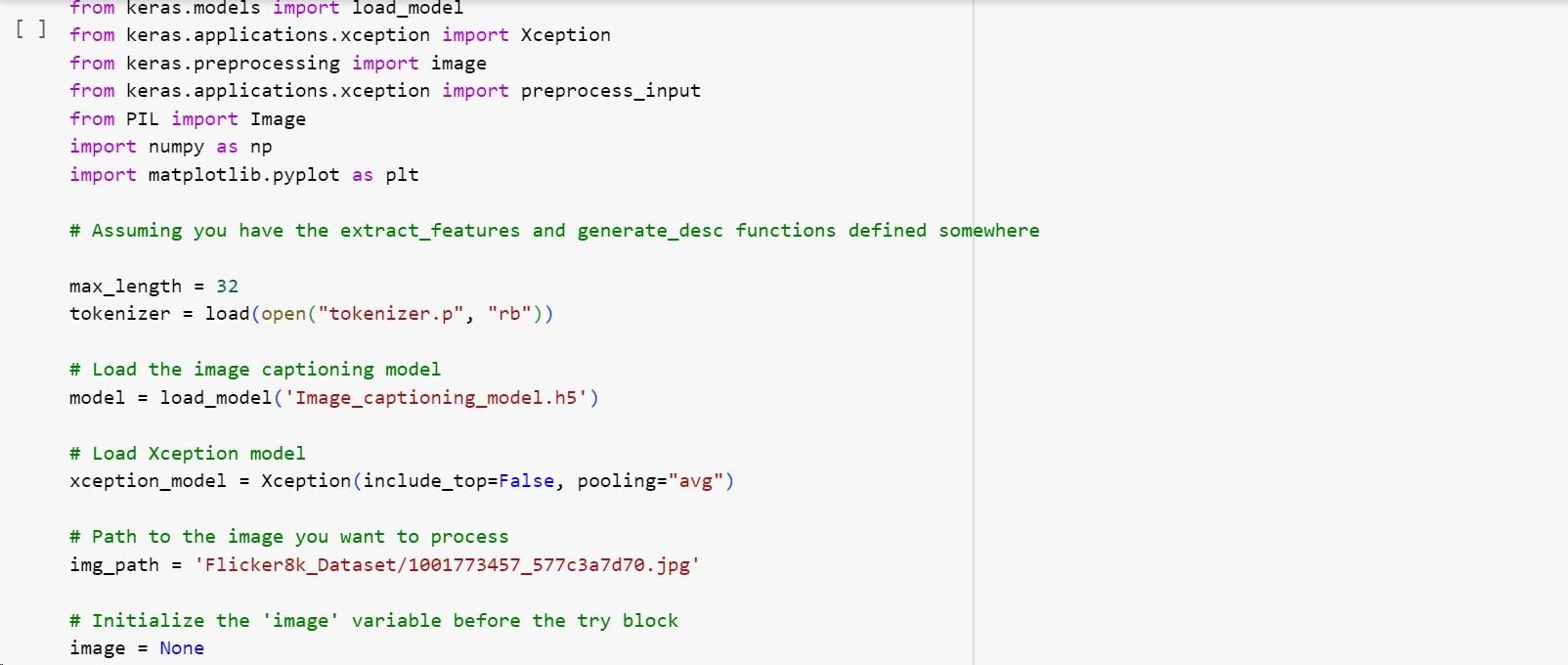


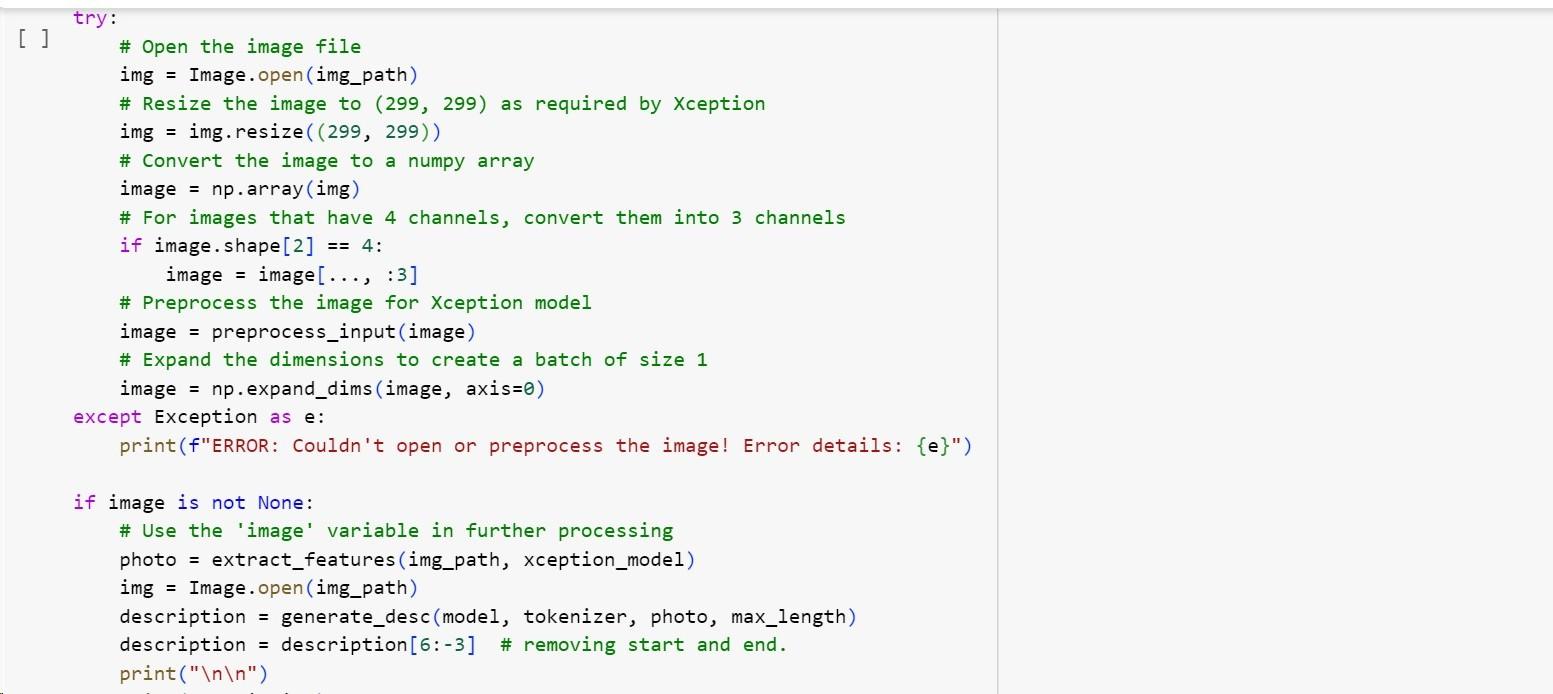
* 1. **Training the Model**

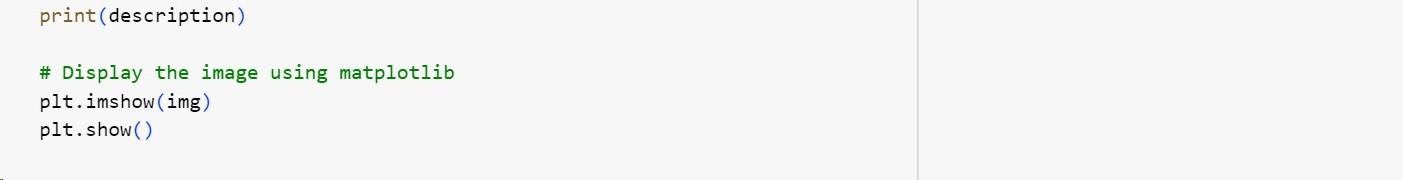


## Testing the model









**Original caption :** 1001773457\_577c3a7d70.jpg, dogs on pavement moving toward each other .

**Generated caption :** 1001773457\_577c3a7d70.jpg, dog is running through the grass



# 6. Future Work

Nevertheless, there exists substantial potential for refining our model to yield superior outcomes. This can be achieved through the augmentation of training data, meticulous fine-tuning of hyperparameters, and an exploration of reinforcement learning techniques and attention mechanisms, all of which are poised to elevate the overall performance of our model.

Looking forward, the envisaged applications for our model encompass image search, picture indexing, and assistive technology tailored for individuals with visual impairments. Furthermore, the adaptability of our model to various languages positions it for global applicability.

In order to rigorously validate the efficacy of our model, it is imperative to subject it to performance assessments using alternative benchmark datasets, such as COCO and Flickr30k. This approach, leveraging a more expansive database, is anticipated to yield nuanced insights into the strengths and weaknesses of our methodology, facilitating the identification of areas warranting refinement.

In summation, our contributions have significantly propelled the field of image captioning, proffering a sound strategy for generating pertinent and diverse descriptions for photographs. The prospective applications of our model are auspicious, and we anticipate its role as a catalyst for continued research and advancements in this domain.



