Dissertation Submitted for the partial fulfillment of the **M.Sc. (Integrated) Five Years Program AIML** degree to the Department of AIML & Data Science.

M.Sc. Project Dissertation

IMAGE CAPTION GENERATOR USING DEEP LEARNING

submitted to

गुजरात युनिवर्सिटी

।योगः कर्मसु कौशलम्।

By

Patel Aneri Girishbhai
Semester- VII

M.Sc. (Integrated) Five Years Program AIML

Department of AIML & Data Science
School of Emerging Science and Technology

Gujarat University

December, 2022
DECLARATION

This is to certify that the research work reported in this dissertation entitled **Image Caption Generator Using Deep Learning** for the partial fulfilment of Master of Science in Artificial Intelligence and Machine Learning degree is the result of investigation done by myself.

Place: Ahmedabad
Date: 19/12/2022

Patel Aneri Girishbhai

Index

Chapter 1
Abstract & Key Words

Abstract

Image captioning is now one of the most frequently required tools in today's world. In addition, deep neural network models are utilized by in-built applications that generate and provide captions for particular images. The process of writing a description for an image is known as image captioning.. Recognizing important objects, their characteristics, and the image's relationships between them are necessary. Sentences that are grammatically and semantically correct are produced by it. A deep learning model for using computer vision and machine translation to describe images and generate captions is presented in this paper. The purpose of this paper is to generate captions by recognizing the relationships between various objects in an image and detecting them. The dataset utilized is Flickr8k and the programming language utilized was Python3, and a ML strategy called Move Learning will be executed with the assistance of the Xception model, to show the proposed explore. The various neural networks involved will also be described in greater detail in this paper. A crucial component of computer vision and natural language processing is the generation of image captions. Facebook and Google Photos use image segmentation, and image caption generators can also be used to create captions for video frames. They will quickly automate the process of interpreting images. Additionally, it has enormous potential for assisting visually impaired individuals.

Keywords: Image, Caption, CNN, Xception, RNN, LSTM, Neural Networks.

Chapter 2
Introduction

A long-standing challenge in artificial intelligence is teaching a computer system to recognise items and describe them using natural language processing (NLP).

To understand the context of a picture and explain it in a natural language like English or another language, image caption generation uses image processing and natural language processing ideas.

While it is simple for humans to perform, it requires a powerful algorithm and a lot of computing power for a computer system to do. There have been numerous attempts to deconstruct this challenge into smaller, more manageable issues like object identification, image classification, and text synthesis. A model is created by mapping photos into a computer system using the input images as two-dimensional arrays.

While it is simple for humans to perform, it requires a powerful algorithm and a lot of computing power for a computer system to do. There have been numerous attempts to deconstruct this challenge into smaller, more manageable issues like object identification, image classification, and text synthesis. A computer system maps images using the input images as two-dimensional arrays.

Furthermore, benchmark datasets also need quick, precise, and competitive evaluation metrics to stimulate quick development. While new datasets frequently spark significant innovation, The ability to automatically describe an image's content using properly constructed English phrases may be quite difficult, but it might have a great impact, for instance by assisting those who are visually handicapped in understanding the content of internet images. The well-studied picture classification or visual perception tasks, which are a primary focus within the computer vision community, are considerably easier than this problem. Deep learning techniques have produced cutting-edge outcomes for caption generating issues.

The most amazing aspect of these methods is that they frequently just demand for a single end-to-end model, as opposed to a pipeline of specially created models, to predict a caption given a picture.

Chapter 3
Basic Terminology

The basic terminologies used in this project are Image, Caption, CNN, Xception, RNN, LSTM, Neural Networks. The following is the brief explanation of this words.

1. Image :

A digital picture is a binary representation of visual data, whereas an image is a visual representation of anything. These visuals can be in the form of still images from a video, graphics, or photographs. In this context, an image is a photograph that has been made, copied, and saved electronically.

2. Caption :

An image is accompanied by a caption that gives the image's name, a brief description, and attribution. Captions don't have a set structure. Description. Mention any elements of the picture that stand out to you as important or relevant.

3. CNN :

A deep learning network architecture that learns directly from data is a convolutional neural network (CNN or ConvNet). CNNs are very helpful for recognising objects, classes, and categories in photos by looking for patterns in the images. They can be quite useful for categorising signal, time-series, image and audio data.

4. Xception :

Depthwise Separable Convolutions are used in the deep convolutional neural network architecture known as Xception. Researchers from Google created it. A convolutional neural network with 71 layers is called Xception. The ImageNet database contains a pre-trained version of the network that has been trained on more than a million images. The pre-trained network can categorise photos into 1000 different object categories, including several animals, a keyboard, a mouse, and a pencil.

5. RNN :

An example of a neural network with loops that allows data to be saved within the network is a recurrent neural network. Recurrent neural networks, in other words, employ their reasoning based on past experiences to predict future events. Recurrent models are useful because they can sequence vectors, which allows the API to handle more challenging jobs.

6. LSTM :

A sophisticated RNN, or sequential network, called a long short term memory network, permits information to endure. It is capable of resolving the RNN's vanishing gradient issue. RNNs, also referred to as recurrent neural networks, are utilised for persistent memory.

Let's say you recall the previous scene when viewing a movie or the prior chapter's events while reading a book. Similar to how RNNs function, they retain the prior knowledge and apply it to the processing of the incoming data. Due to diminishing gradient, RNN have the flaw of being unable to recall long-term dependencies. Long-term dependency issues are specifically avoided when designing LSTMs.

7. Neural Networks :

Artificial intelligence, machine learning, and deep learning algorithms can discover trends and address common issues because neural networks mimic the activity of the human brain.

Chapter 4

Literature Review

A written description must be provided for a given image as part of the difficult artificial intelligence challenge known as caption creation. It takes both computer vision techniques to comprehend the image's content and a language model from the discipline of natural language processing to translate that understanding into the appropriate sequence of words. On examples of this problem, deep learning techniques recently produced state-of-the-art results.

Deep learning techniques have produced cutting-edge outcomes for caption generating issues. The most amazing aspect of these methods is that, rather than requiring complex data preparation or a pipeline of specially created models, a single end-to-end model can be developed to predict a caption given a photo.

1. Convolutional Neural Network, first (CNN) :-

Convolutional Neural Network (CNN) is a Deep Learning method that takes in an input image and gives priority (learnable weights and biases) to different characteristics and objects in the image to help it distinguish between distinct images.

Image classification is one of this architecture's most well-liked uses. The neural network combines nonlinear, pooling, and multiple convolutional layers. The output of the first convolution layer becomes the input for the second layer after the image has gone through one convolution layer. For each layer after that, this process is repeated.

It is important to attach a completely connected layer following a series of convolutional, nonlinear, and pooling layers. The output data from convolutional networks is used in this layer. An N-dimensional vector, where N is the number of classes from which the model chooses the desired class, is produced by attaching a fully linked layer to the end of the network.

2. Long-term Memory. :-

Recurrent neural networks (RNNs) of the Long Short-Term Memory (LSTM) type are able to recognise order dependence in sequence prediction issues. The most common applications of this are in difficult issues like speech recognition, machine translation, and other issues. When training conventional RNNs, this issue was observed because as we go further into a neural network, if the gradients are very small or zero, little to no training can occur, resulting in poor predicting performance. Since there may be lags of uncertain length between significant occurrences in a time series, LSTM networks are well-suited for categorizing, processing, and making predictions based on time series data.

3. CNN-LSTM architecture :-

The CNN-LSTM architecture combines LSTMs to facilitate sequence prediction with CNN layers for feature extraction on input data. This approach is especially made for problems involving the sequence prediction of spatial inputs, such as photos or videos. They are frequently utilised in activities including activity recognition, image and video description, and many others.

The following are the research paper that are studied while doing project are as follows :

1. **Visual Image Caption Generator :**

https://www.researchgate.net/publication/333214768_Visual_Image_Caption_Generator_Using_Deep_Learning

2. **Image caption generator based on neural networks :**

https://www.math.ucla.edu/~minchen/doc/ImgCapGen.pdf

3. Image caption Generator using Attention Mechanism :

https://ieeexplore.ieee.org/document/9579967

4. Image caption :

https://www.sciencegate.app/keyword/761035

Chapter 5
Methodology

A. Libraries Used :-

In this project, I have used 9 libraries namely numpy, PIL, os, pickle, keras, tensorflow, tqdm, tkinter, string.

B. Data Gathering :-

I'll be using the Flickr 8K dataset to train the model for the image caption generator in this project. The Flickr8k Text folder holds text files with the image captions, whereas the Flickr8k Dataset folder has 8091 pictures. These two elements comprise the dataset.

C. Data Pre-processing :-

In this Flickr8k Dataset folder, the file name Flickr 8k.token contain all images and it's caption. Each image has 5 captions numbered 0 to 4. In this step -

1) I am going to map all images with their caption in dictionary form.

2) I will perform text cleaning.

3) I will create a vocabulary from all the unique words extracted out from descriptions.

4) I will store all preprocessed description into a file named Descriptions.txt

D. Extracting feature vector :-

Now we are going to use pre-trained model named "Xception" which we can access from keras.applications. The goal of this step is to extract features for all images and store this feature dictionary into pickle file.

E. Loading dataset for model training :-

Our Flickr 8k test folder has a file called "Flickr 8k.trainImages.txt." A list of 6000 image names that are used for training purposes can be found in this file. In this step, I will create three function-

- Load photos- This function will return the list of image names by loading the text file into a string.
- Load clean descriptions - This function will store the captions of each image from the list of photos to a dictionary.
- Load features - The extracted feature vectors from the Xception model and the dictionary for the photos are returned by this function.

F. Tokeninzing the vocabulary :-

Since machines can't understand complicated English words, they need a simple numerical representation to process model data.This is why we assign a unique index value to each vocabulary word.To generate tokens from our vocabulary, the Keras library includes a built-in tokenizer function. They can be saved in a pickle file called "tokenizer.p" for us.

G. Create a data generator :-

We need to provide the model with input and output sequences in order to train it as a supervised learning task.Our training sets contain a total of 6000 images with captions denoted by numbers and a 2048-length feature vector.Because it is impossible to store such a large amount of data in memory, we will employ a generator technique that will produce batches.

H. Define the CNN-RNN model :-

I will use the Keras Model to define the model's structure from the Functional API. It contains:

- Feature Extractor - It will extract the feature from images of size 2048 using a dense layer, and we will reduce the dimensions to 256 nodes.
- Sequence Processor: This embedded layer takes care of textual input after the LSTM i.e long short-term memory networks layer.
- Decoder: To arrive at the final prediction, we will combine the results of the previous two layers and process the dense layer.

I. Training the image caption generator model :-

With 6000 training images, we will generate the input and output sequences necessary to train our model.To fit the batches to the model, we create the function named model.fit_generator().Finally, the model is saved in our models folder.

J. Testing the image caption generator model :-

Now we can used different images to check whether our model is accurately generating the captions or not.

K. Build a GUI.

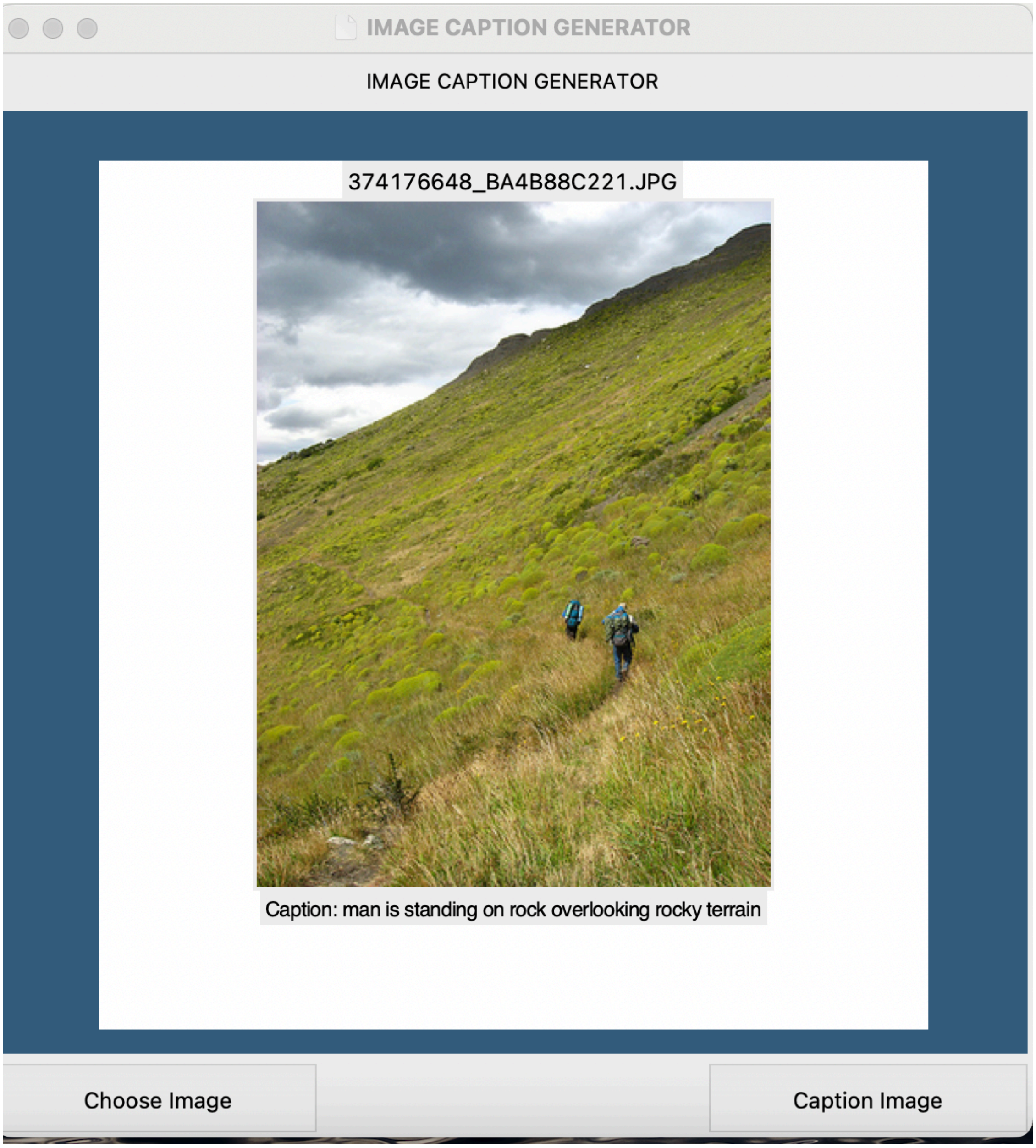
tkinter library has been used to build a graphical user interface. which has option to upload image image and after that to generate caption users have to click the button.

Chapter 6

Result & Discussion

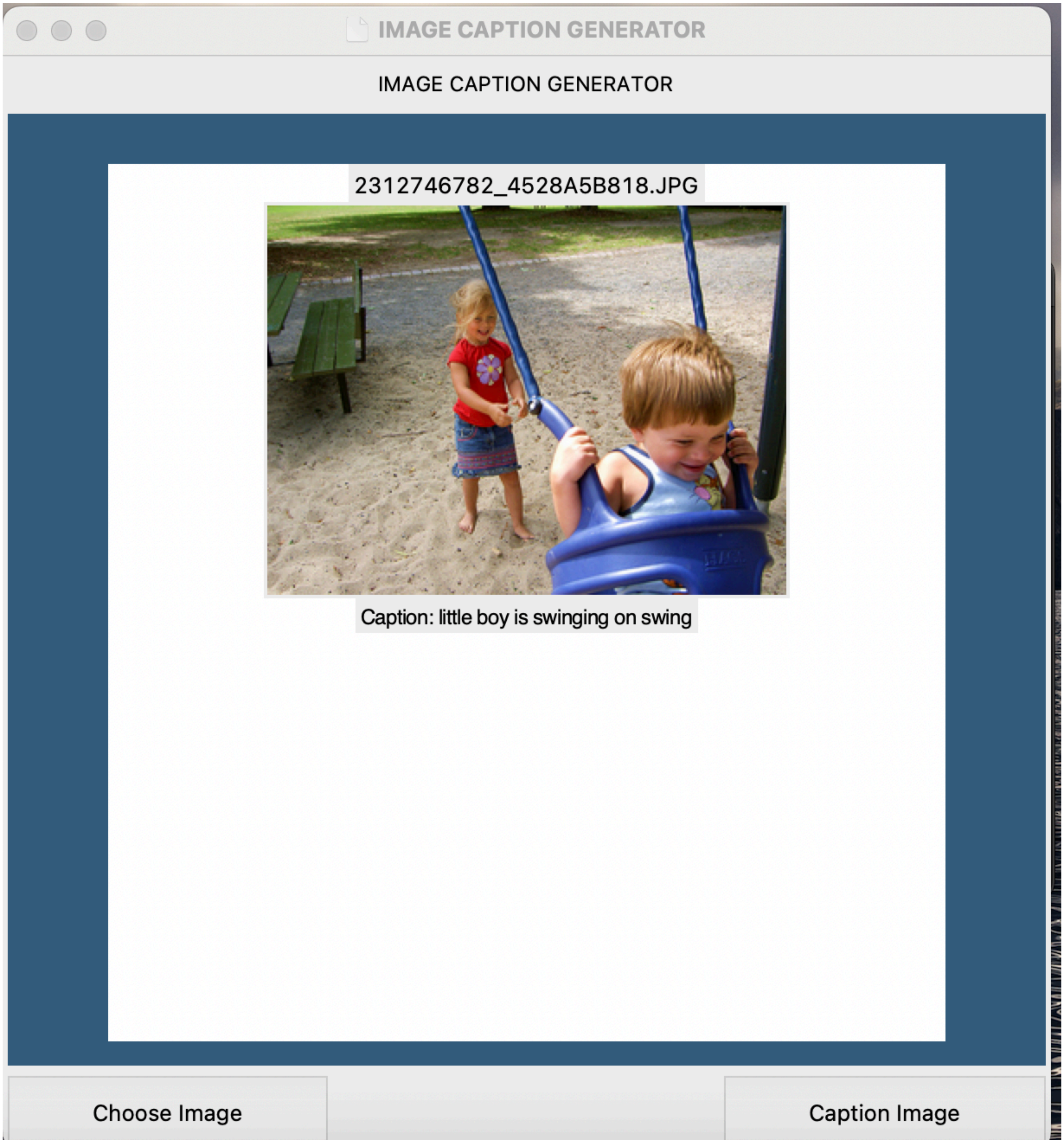
The following are the some results we got when we run model on some unseen data :-

Test Image 1



Caption: man is standing on rock overlooking rocky terrain

Test Image 2



2312746782_4528A5B818.JPG

Caption: little boy is swinging on swing

Test Image 3

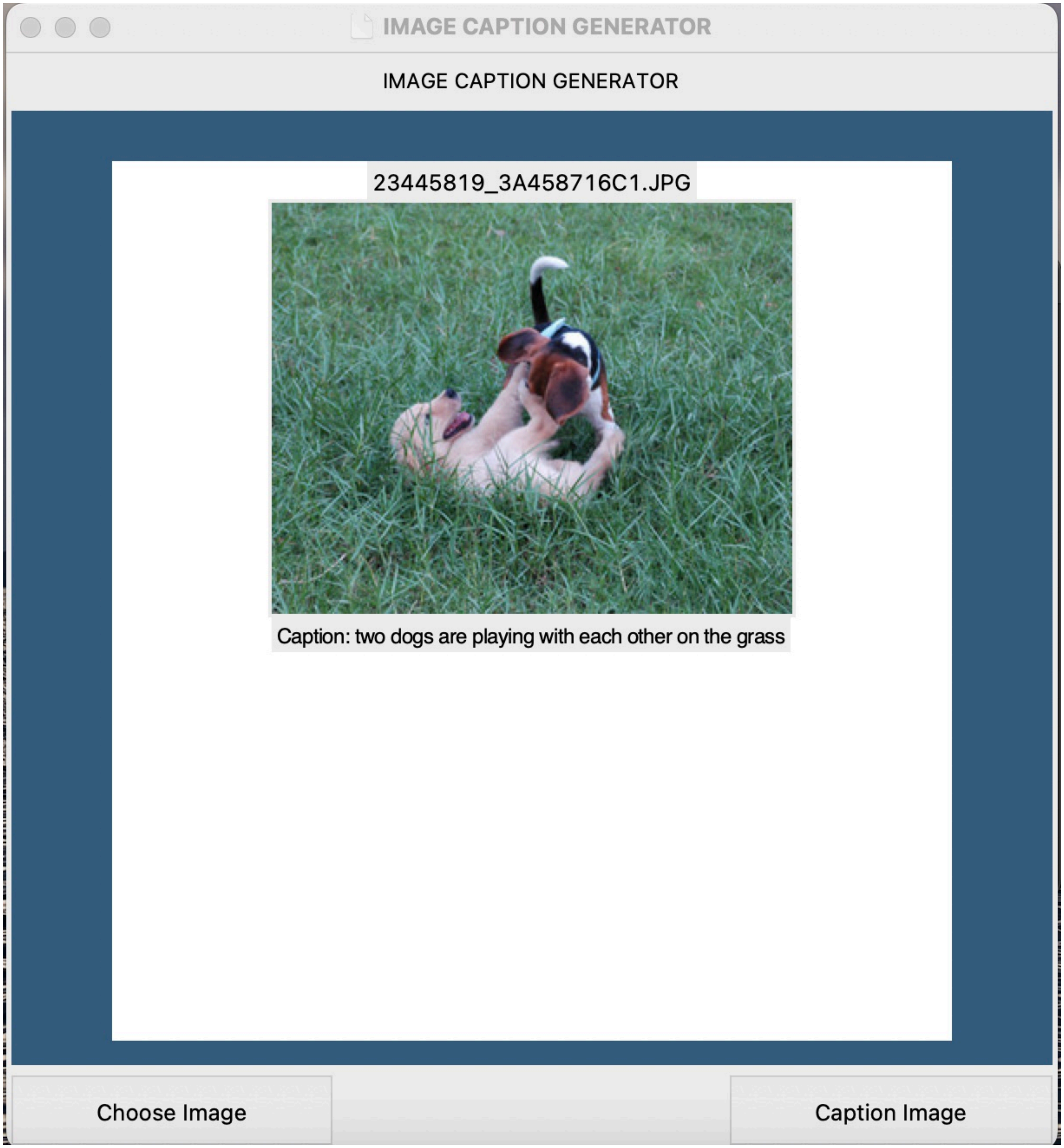


IMAGE CAPTION GENERATOR

IMAGE CAPTION GENERATOR

23445819_3A458716C1.JPG

Caption: two dogs are playing with each other on the grass

Chapter 7
Conclusion

We can see from the data that the deep learning technology employed here produced fruitful outcomes. Together, the CNN and LSTM were able to determine the relationship between objects in images by synchronising their operations.

Using the BLEU (Bilingual Evaluation Understudy) score, we can compare the projected captions to the target captions in our Flickr8k test dataset to assess how accurate they are. Text translation uses BLEU ratings to compare translated text to one or more reference translations. Similar to the one described here, hybrid picture caption generators have been made over the years using a variety of other neural network approaches. for instance, using the GRU model instead of the STM model or the VGG16 model in place of the Xception model.

Chapter 8
Bibliography

References

1. **HaoranWang , Yue Zhang, and Xiaosheng Yu, "An Overview of Image Caption Generation Methods", (CIN-2020)**

2. B.Krishnakumar, K.Kousalya, S.Gokul, R.Karthikeyan, and D.Kaviyarasu, "IMAGE CAPTION GENERATOR USING DEEP LEARNING", (international Journal of Advanced Science and Technology- 2020)

3. MD. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga, "A Comprehensive Survey of Deep Learning for Image Captioning" ,(ACM-2019)

4. Rehab Alahmadi, Chung Hyuk Park, and James Hahn, "Sequence-to-sequence image caption generator", (ICMV-2018)

5. Oriol Vinyals, Alexander Toshev, SamyBengio, and Dumitru Erhan, "Show and Tell: A Neural Image Caption Generator",(CVPR 1, 2-2015)

6. Priyanka Kalena, Nishi Malde, Aromal Nair, Saurabh Parkar, and Grishma Sharma, “Visual Image Caption Generator Using Deep Learning”, (ICAST-2019)

7. J. Redmon, S. Divvala, Girshick and A. Farhadi, "You only look once: Unified real-time object detection", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

8. D. Bahdanau, K. Cho, and Y. Bengio. “Neural machine translation by jointly learning to align and translate.arXiv:1409.0473”, 2014.

9. Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi, “Understanding of a convolutional neural network”, IEEE - 2017.