IMAGE CAPTIONING

A PROJECT REPORT

Submitted by

SNEGA R
CB.SC.I5DAS18035

As part of the curriculum requirement for the course 18CSC491 — Mini Project

in partial fulfilment of the requirements for the award of the degree of

INTEGRATED MASTER OF SCIENCE

IN

DATA SCIENCE

AMRITA SCHOOL OF ENGINEERING

AMRITA VISHWA VIDYAPEETHAM

COIMBATORE 641112

JUNE 2022

AMRITA VISHWA VIDYAPEETHAM

AMRITA SCHOOL OF ENGINEERING, COIMBATORE, 641112

BONAFIDE CERTIFICATE

This is to certify that the project report entitled **"IMAGE CAPTIONING"** submitted by **CB.SC.I5DAS18035 SNEGA R** as part of the curriculum requirement under the course **18CSC491 — Mini Project**, in partial fulfilment of requirements for the award of **Degree of Integrated Master of Science (M.Sc.) in DATA SCIENCE** is a bonafide record of the work carried out at Amrita School of Engineering, Coimbatore.

Class Advisor

Designation

Chairperson
Department of Mathematics
Dr. J. Ravichandran

The project was evaluated by us on:

Internal Examiner

External Examiner

AMRITA VISHWA VIDYAPEETHAM

AMRITA SCHOOL OF ENGINEERING, COIMBATORE, 641112

DECLARATION

I, **Snega R**, hereby declare that this project report entitled **Image Captioning**, is the work done by me, as **part of the curriculum requirement** under the course "**18CSC491 — Mini Project**", in partial fulfilment of requirements for the award of **degree of Integrated Master of Science (M.Sc.) in Data Science** at Amrita Vishwa Vidyapeetham Coimbatore. The work which is being presented in the report submitted to Department of Mathematics at Amrita Vishwa Vidyapeetham Coimbatore is an authentic record of work.

Signature of the Student

ACKNOWLEDGEMENTS

I would like to warmly acknowledge and express my deep sense of gratitude and indebtedness to my class advisor Dr. Prakash. P Department of Mathematics, Amrita Vishwa Vidyapeetham, Coimbatore, for his constant encouragement and prudent suggestions during the course of my study and preparation of the final manuscript of this Project.

I take this opportunity to express my sincere thanks to Dr. Sasangan Ramanathan, Dean of Engineering for extending his support in giving me this opportunity.

I express my sincere thanks to Dr. Ravichandran J, Chairperson, Department of Mathematics, Amrita School of Engineering, for his encouragement and for providing necessary facilities in the Department.

I acknowledge my deep sense of gratitude and dedicate this work to my family members and friends who have always been a source of inspiration to me throughout my study. I am so grateful for the patience, love and care they have shown during the period of my project work.

SNEGA R

CONTENTS

ABSTRACT

Captioning images automatically is one of the basic and important skills of the human visual system. There are various advantages where the scenes surrounded by them are automatically captioned and revert back the caption as a plain message. This paper presents a model based on CNN-LSTM neural networks which automatically detects the objects in the images and generates descriptions for the images. It uses pre-trained model to perform the task of detecting objects and uses CNN and LSTM to generate the captions. This model can perform two operations. The first one is to detect objects in the image using Convolutional Neural Networks and the other is to caption the images using RNN based LSTM (Long Short-Term Memory)

Caption generation is one of the interesting and focussed areas of Artificial Intelligence which has many challenges to pass on. Caption generation involves various complex scenarios starting from picking the dataset, training the model, validating the model, creating pre-trained models to test the images, detecting the images and finally generating the captions.

ABBREVATIONS

CNN	-	Convolutional Neural Network.
RNN	-	Recurrent Neural Network.
LSTM	-	Long Short-Term Memory.
NLTK	-	Natural Language Tool Kit
NLP	-	Natural Language Processing.
BLEU	-	Bi lingual Evaluation Understudy

CHAPTER 1

INTRODUCTION

"A picture is worth a thousand words. But sometimes we actually want the words."

Image Captioning is the process of generating a textual description for given images. It has emerged as a challenging and important research area following advances in statistical language modelling and image recognition. It includes the labelling of an image with English keywords with the help of datasets provided during model training.

The generation of captions from images has various practical benefits, ranging from aiding the visually impaired, to enabling the automatic and cost-saving labelling of the millions of images uploaded to the Internet every day. The field also brings together state-of-the-art models in Natural Language Processing and Computer Vision, two of the major fields in Artificial Intelligence. NVIDIA is using image captioning technologies to create an application to help people who have low or no eyesight.

Image captioning can be regarded as an end-to-end Sequence to Sequence problem, as it converts images, which is regarded as a sequence of pixels to a sequence of words. For this purpose, we need to process both the language or statements and the images.

For the Language part, we use recurrent Neural Networks and for the Image part, we use Convolutional Neural Networks to obtain the feature vectors respectively.

CHAPTER 2

PROPOSED MODELS

2.1 CNN

CNN is a subfield of Deep learning and specialized deep neural. One of the most popular applications of this architecture is image classification. The neural network consists of several convolutional layers mixed with nonlinear and pooling layers. When the image is passed through one convolution layer, the output of the first layer becomes the input for the second layer. This process continues for all subsequent layers.

After a series of convolutional, nonlinear and pooling layers, it is necessary to attach a fully connected layer. This layer takes the output information from convolutional networks. Attaching a fully connected layer to the end of the network results in an N dimensional vector, where N is the number of classes from which the model selects the desired class.

2.2 LSTM

LSTM stands for Long short-term memory; they are a type of RNN (recurrent neural network) which is well suited for sequence prediction problems. Based on the previous text, we can predict what the next word will be. It has proven itself effective from the traditional RNN by overcoming the limitations of RNN which had short term memory. LSTM can carry out relevant information throughout the processing of inputs and with a forget gate, it discards non-relevant information.

LSTMs use gated cells to store information outside the regular flow of the RNN. With these cells, the network can manipulate the information in many ways, including storing information in the cells and reading from them. The cells are individually capable of making decisions regarding the information and can execute these decisions by opening or closing the gates. The ability to retain information for a long period of time gives LSTM the edge over traditional RNNs in these tasks.

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—removes information that is no longer necessary for the completion of the task. This step is essential to optimizing the performance of the network.
- Input gate—responsible for adding information to the cells.
- Output gate—selects and outputs necessary information

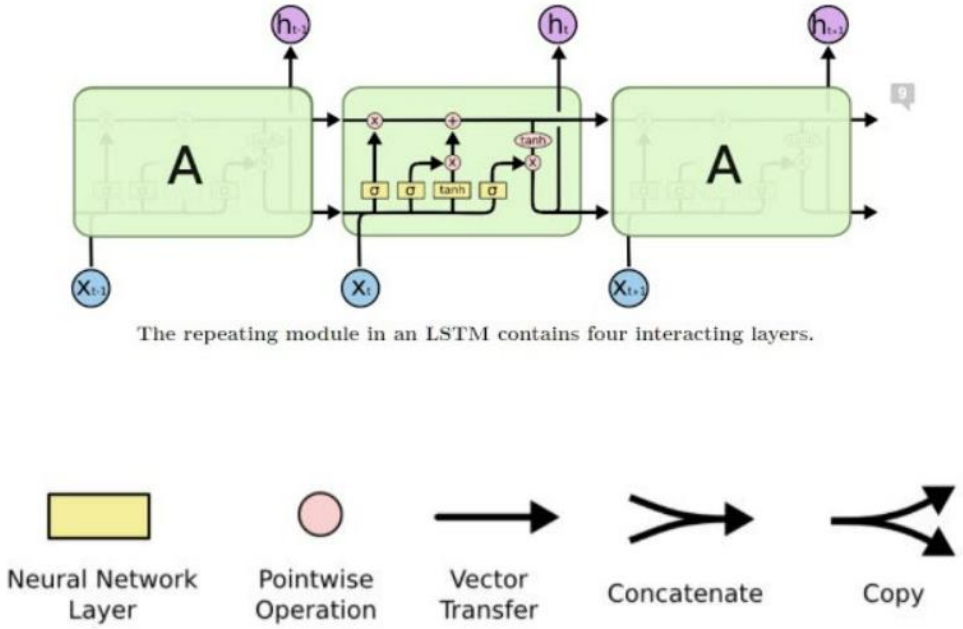

Figure: LSTM networks

2.3 MODEL ARCHITECTURE

To build an image caption generator model we have to merge CNN with LSTM. We can drive that:

Image Caption Generator Model (CNN-RNN model) = CNN + LSTM.

- CNN - To extract features from the image. A pre-trained model called Xception is used for this. ImageNet dataset is used to train the CNN model called Xception. Xception is responsible for image feature extraction. These extracted features will be fed to the LSTM model which in turn generates the image caption.
- LSTM - To generate a description from the extracted information of the image.

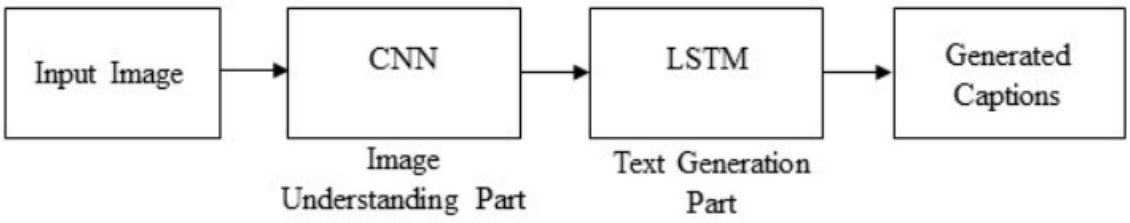

Figure: A block diagram of simple CNN-LSTM architecture-based image captioning

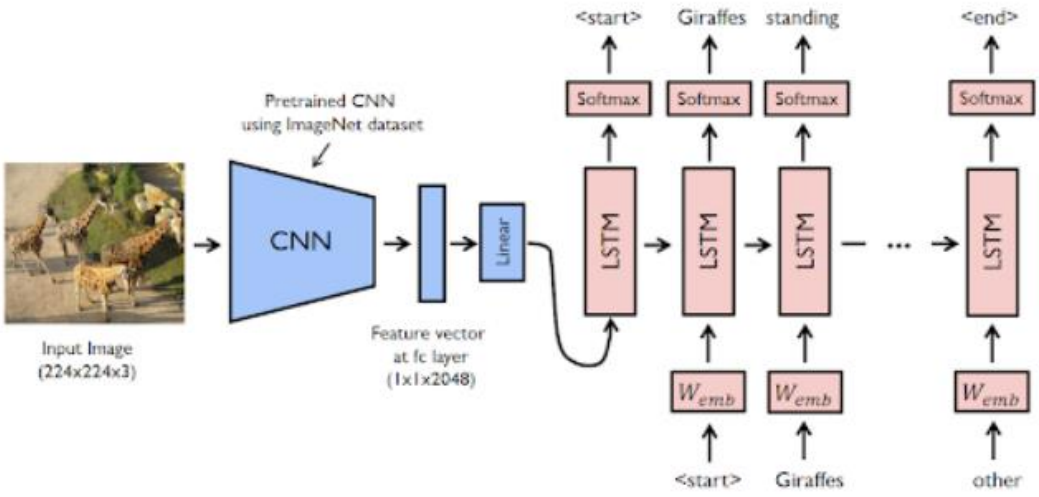

Figure: CNN-LSTM architecture-based image captioning

CHAPTER 3

DATASET PREPARATION

3.1 DATASET FOR IMAGE CAPTION GENERATOR

Flickr8k dataset is a public benchmark dataset for image to sentence description. This dataset consists of 8000 images with five captions for each image. These images are extracted from diverse groups in Flickr website. Each caption provides a clear description of entities and events present in the image. The dataset depicts a variety of events and scenarios and doesn't include images containing well-known people and places which makes the dataset more generic. The dataset has 6000 images in training dataset, 1000 images in development dataset and 1000 images in test dataset. Features of the dataset making it suitable for this project are:

• 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.

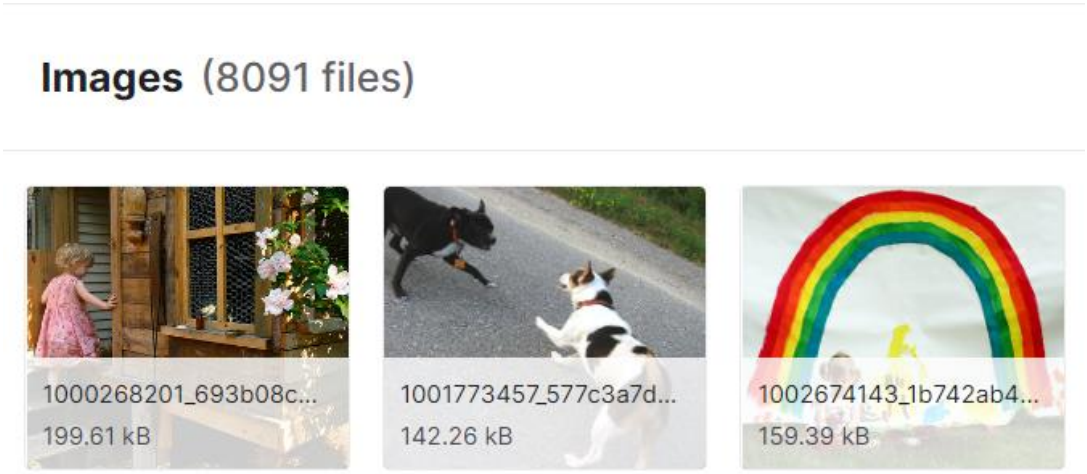

Figure: Snip of an image dataset

captions.txt - Notepad

File Edit Format View Help

```
image,caption
1000268201_693b08cb0e.jpg,A child in a pink dress is climbing up a set of stairs in an entry way .
1000268201_693b08cb0e.jpg,A girl going into a wooden building .
1000268201_693b08cb0e.jpg,A little girl climbing into a wooden playhouse .
1000268201_693b08cb0e.jpg,A little girl climbing the stairs to her playhouse .
1000268201_693b08cb0e.jpg,A little girl in a pink dress going into a wooden cabin .
1001773457_577c3a7d70.jpg,A black dog and a spotted dog are fighting
1001773457_577c3a7d70.jpg,A black dog and a tri-colored dog playing with each other on the road .
1001773457_577c3a7d70.jpg,A black dog and a white dog with brown spots are staring at each other in the street .
1001773457_577c3a7d70.jpg,Two dogs of different breeds looking at each other on the road .
1001773457_577c3a7d70.jpg,Two dogs on pavement moving toward each other .
1002674143_1b742ab4b8.jpg,A little girl covered in paint sits in front of a painted rainbow with her hands in a bowl .
1002674143_1b742ab4b8.jpg,A little girl is sitting in front of a large painted rainbow .
1002674143_1b742ab4b8.jpg,A small girl in the grass plays with fingerpaints in front of a white canvas with a rainbow on it .
1002674143_1b742ab4b8.jpg,There is a girl with pigtails sitting in front of a rainbow painting .
1002674143_1b742ab4b8.jpg,Young girl with pigtails painting outside in the grass .
```

Figure: Snip of a caption dataset

3.2 CAPTION DATA PREPARATION

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

3.2.1 DATA PRE-PROCESSING

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.

Original Caption	Captions after cleaning data
A little girl climbing into a wooden playhouse.	a little girl climbing into a wooden playhouse
A dog in a snowy area.	a dog in a snowy area
The kid is playing hopscotch.	the kid is playing hopscotch

Figure: Data cleaning of captions.

CHAPTER 4

IMPLEMENTATION

4.1 PRE-REQUISITES

- NumPy
- Pandas
- Keras
- TensorFlow
- nltk

4.2 PERFORMING DATA CLEANING

As we see all image captions are available in the Flickr 8k.token file of the Flickr_8k_text folder. If you analyse this file carefully, you can drive the format of image storing, each image and caption separated by a new line and carry 5 captions numbered from 0 to 4 along with.

```
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
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 .
```

Figure: Flickr dataset text format

Now we are going to define 5 functions for cleaning (in code):

1. load_fp(filename) – To load the document file and read the contents of the file into a string.
2. img_capt(filename) – To create a description dictionary that will map images with all 5 captions.
3. txt_cleaning(descriptions) – This method is used to clean the data by taking all descriptions as input. While dealing with textual data we need to perform several types of cleaning including uppercase to lowercase conversion, punctuation removal, and removal of the number containing words.
4. txt_vocab(descriptions) – This is used to create a vocabulary from all the unique words extracted out from descriptions.
5. save_descriptions(descriptions, filename) – This function is used to store all the pre-processed descriptions into a file.

4.3 EXTRACT THE FEATURE VECTORS

Now we are going to use the pre-trained model called Xception which is already trained with large datasets to extract the features from these models. Xception was trained on an ImageNet dataset with 1000 different classes to classify the images. We can use keras.applications to import this model directly. We need to do a few changes to the Xception model to integrate it with our model. The Xception model takes 299*299*3 image size as input so we need to delete the last classification layer and extract out the 2048 feature vectors.

Extract_features() function is used to extract these features for all images.

4.4 LOADING DATASET FOR TRAINING MODEL

A file named “Flickr_8k.trainImages.txt” is present in our Flickr_8k_test folder. This file carries a list of 6000 image names that are used for the sake of training.
Functions required to load the training datasets:

- load_photos(fname) – This function will take a file name as a parameter and return the list of image names by loading the text file into a string.
- load_clean_descriptions(fname, image) – This function stores the captions for every image from the list of photos to a dictionary. For the ease of the LSTM model in identifying the beginning and ending of a caption, we append the and identifier with each caption.
- load_features(photos) – The extracted feature vectors from the Xception model and the dictionary for photos are returned by this function.

4.5 TOKENIZING THE VOCABULARY

Machines are not familiar with complex English words so, to process model’s data they need a simple numerical representation. That’s why we map every word of the vocabulary with a separate unique index value. An in-built tokenizer function is present in the Keras library to create tokens from our vocabulary.

4.6 DEFINE THE CNN-RNN MODEL

- Feature Extractor –With a dense layer, it will extract the feature from the images of size 2048 and we will decrease the dimensions to 256 nodes.
- Sequence Processor – Followed by the LSTM layer, the textual input is handled by this embedded layer.
- Decoder – We will merge the output of the above two layers and process the dense layer to make the final prediction.

4.7 TRAINING THE MODEL

We will generate the input and output sequences to train our model with 6000 training images. We create a function named model.fit_generator() to fit the batches to the model.

4.8 TESTING THE MODEL

A black dog carries a green toy in his mouth as he walks through the grass

CHAPTER 5

EVALUATION METRICS

BLEU:

BLEU is a quality metric score for machine translation systems that attempts to measure the correspondence between a machine translation output and a human translation. The central idea behind BLEU is that the closer a machine translation is to a professional human translation, the better it is.

BLEU scores only reflect how a system performs on the specific set of source sentences and the translations selected for the test. As the selected translation for each segment may not be the only correct one, it is often possible to score good translations poorly. As a result, the scores don't always reflect the actual potential performance of a system, especially on content that differs from the specific test material.

BLEU

- N-gram overlap between machine translation output and reference translation
- Compute precision for n-grams of size 1 to 4
- Add brevity penalty (for too short translations)

$$\text{BLEU} = \min\left(1, \frac{\textit{output-length}}{\textit{reference-length}}\right) \left(\prod_{i=1}^{4} \textit{precision}_i\right)^{\frac{1}{4}}$$

- Typically computed over the entire corpus, not single sentences

The BLEU metric ranges from 0 to 1; 1 is very rare – only for perfect match.

CHAPTER 6

ADVANTAGES, CONCLUSION AND REFERENCES

6.1 ADVANTAGES

1. Assistance for visually impaired

a. The advent of machine learning solutions like image captioning is a boon for visually impaired people who are unable to comprehend visuals.

b. With AI-powered image caption generators, image descriptions can be read out to the visually impaired, enabling them to get a better sense of their surroundings.

2. Recommendations in editing

a. The image captioning model automates and accelerates the closed captioning process for digital content production, editing, delivery, and archival.

b. Well-trained models replace manual efforts for generating quality captions for images as well as videos.

3. Media and Publishing Houses

a. The media and public relations industry circulate tens of thousands of visual data across borders in the form of newsletters, emails, etc.

b. The image captioning model accelerates subtitle creation and enables executives to focus on more important tasks.

4. Self-driving cars

a. By using the captions generated by image caption generator self-driving cars become aware of the surroundings and make decisions to control the car.

5. Reduce vehicle accidents

a. By installing an image caption generator in the vehicles, vehicles can stop by applying the automatic brake when an object in the surrounding is detected.

6.2 CONCLUSION

Image captioning has made significant advances in recent years. Recent work based on deep learning techniques has resulted in a breakthrough in the accuracy of image captioning. The text description of the image can improve the content-based image retrieval efficiency, the expanding application scope of visual understanding in the fields of medicine,

security, military and other fields, which has a broad application prospect. At the same time, the theoretical framework and research methods of image captioning can promote the development of the theory and application of image annotation and visual question answering (VQA), cross media retrieval, video captioning and video dialog, which has important academic and practical application value.

6.3 REFERENCES:

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

https://neurohive.io/en/popular-networks/vgg16/

https://medium.com/swlh/automatic-image-captioning-using-deep-learning-5e899c127387

http://cse.anits.edu.in/projects/projects1920B11.pdf

https://www.matec-conferences.org/articles/matecconf/pdf/2018/91/matecconf_eitce2018_01052.pdf

https://ijcrt.org/papers/IJCRT2103510.pdf