IMAGE CAPTION GENERATOR

Prof. Sandhya [1], Aishwarya S Salunke [2], Apoorva Pai [3], Tejaswini P Angadi [4], Usha Nandini V

***Abstract*—Image caption, automatically generating natural language descriptions according to the content observed in an image, it is an important part of scene understanding, which combines the knowledge of computer vision and natural language processing. The objective of our project is to learn the concepts of a CNN and LSTM model and build a working model of Image caption generator by implementing CNN with LSTM. We will be implementing the caption generator using CNN (Convolutional Neural Networks) and LSTM (Long short-term memory) which is further converted to speech. The image features will be extracted from Xception which is a CNN model trained on the image dataset and then we feed the features into the LSTM model which will be responsible for generating the image captions and the generated captions are converted into speech to benefit the visually impaired. This paper summarizes the related methods and focuses on the attention mechanism, which plays an important role in computer vision and is recently widely used in image caption generation tasks.**

***Keywords*— Convolutional Neural Network, Long Short-term Memory.**

INTRODUCTION

Describing the content of images automatically using natural language like English is a very challenging task. With the evolution in computing power and with the availability of datasets that are huge, constructing models that are capable of generating captions [7] for an image is possible.

Humans are easily able to define the environments they are in. Given a picture, a person can easily explain an image with details in a fast glance. Even though huge development has been done in computer vision, tasks like recognizing objects, classifying actions, images and attributes and scene recognition are achievable but it is a new task for a computer to describe an image that is sent to it in the form of a sentence.

Based on the semantics of the photographs, this purpose of image captioning should be captured and expressed in the desired form of natural languages. The fact that the generated captions can be spoken has a significant influence in the actual world, such as improving the comprehension of online image information by persons who are blind. So, we will combine CNN-RNN architectures to create our image caption generator model. CNN is used for feature extraction from pictures. The pre-trained model Exception was used. The LSTM [8] then employs the data from CNN to produce a description of the image.

The generic descriptions of the visual content that are produced using these methods omit any contextual information. Such generic descriptions fall short in emergent circumstances because they simply duplicate the information found in the images and omit to provide the comprehensive details [6] that are required for comprehending emergent circumstances.

Our project's goal is to create a web-based interface that allows users to access an image's [5] description, which is then translated into voice, as well as to create a classification scheme that will allow us to distinguish between distinct photos based on their descriptions. Additionally, it can simplify the difficult process of SEO (Search Engine Optimization), which involves maintaining vast volumes of data.

1. LITERATURE SURVEY

In the literature review, relevant references from previous projects that are related to the current project are taken into account.

[1] The Convolutional Neural Network (CNN), one of the most well-known deep neural networks, is explained in this study. Convolutional, nonlinear, pooling, and fully connected layers are only a few of the many layers that make up CNN. The CNN is one of the most used algorithms and performs exceptionally well in machine learning issues.

[2] Sepp Hochreiter provides an explanation of the deep neural network algorithm in this publication (LSTM). The computational complexity is per step and also the representation of the weight pattern, and LSTM is local in both space and time. LSTM learns significantly more quickly and produces many more successful runs than other algorithms. It even completes difficult, fake, long-duration tasks that no prior recurrent network has ever been able to complete.

[3] Automatically describing the content of a picture is the central issue in artificial intelligence that links computer vision with natural language processing. A.L thoroughly examines a deep neural network-based image caption generating approach in this research. In this case, an image is the input, and the method's output takes the form of an English sentence that describes the image's content. Convolutional neural networks (CNN), recurrent neural networks (RNN), and sentence generation are the three parts of the method that they examine. This model analyses the image and generates additional unimportant but pertinent words for the picture.

[4] Current methods for creating image captions produce descriptions that lack particular details, like identified elements that are visible in the photos. Here, Di Lu and Spencer Whitehead suggested a brand-new task that creates evocative image descriptions from input photographs. We will train a CNN-LSTM model to be able to generate a caption based on the image as our straightforward answer to this issue.

[5] It is a difficult task to automatically describe the content of a picture using properly constructed English words and then translate it into voice, but it is something that is highly important for assisting those who are visually impaired. Modern cellphones have cameras, which can help vision impaired people capture images of their surroundings. Here, captions can be generated using photographs as input, and the captions are then read aloud loudly enough for people who are blind to hear.

2. PROPOSED METHODOLOGY

1. Object detection
CNN Encoder is used to identify objects in the image.

2. Creation of Sentences
Sentences are created using LSTM. The next word is obtained using each predicted word. The right sentence is created using these words.

3.Translation from Text to Speech
Further voice synthesis is done with the resulting captions.

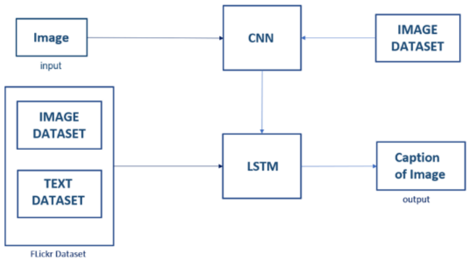
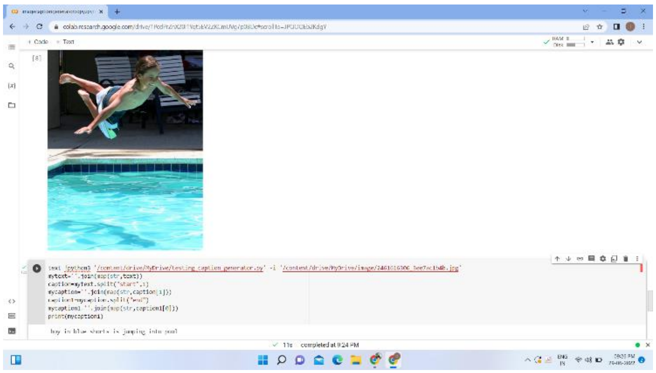


Figure 2.1

System Architecture of Image Caption Generator

For an image caption generator, this embedding becomes a representation of the image and used as the initial state of the LSTM for generating meaningful captions, for the image.

3.RESULT



```
text !python3 '/content/drive/MyDrive/testing_caption_generator.py' -i '/content/drive/MyDrive/image/2461616306_3ee7ac1b4b.jpg'
mytext=''.join(map(str,text))
caption=mytext.split("start",1)
mycaption=''.join(map(str,caption[1]))
caption1=mycaption.split("end")
mycaption1=''.join(map(str,caption1[0]))
print(mycaption1)
```

```
boy in blue shorts is jumping into pool
```

11s completed at 9:24 PM

Figure 6.1

Generated Caption: Boy in blue shorts is jumping into pool.

4.CONCLUSION

A CNN-LSTM model has been used to create an image caption generator in this sophisticated Python project.

Additionally, we have summarised the attention mechanism's use, discussed the model framework for solving the description task, and compiled all components of the image caption production task.

The datasets have been condensed, the CNN and LSTM models have been trained, captions have been produced, and speech has been generated.

5.FUTURE SCOPE

Our model depends on the data, thus it cannot anticipate terms that are not in its lexicon. This is one of the most important parts of our project to keep in mind. Here, an 8000-image dataset is utilised. But in order to construct production-level models, or greater accuracy models, we need to train the model on datasets that are larger than 100,000 photos.

6.REFERENCE

[1] S. ALBAWI , "Understanding of a CNN," 2017.

[2] S. Hochreiter, Neural Computation, December 1997.

[3]O. Vinyals, "A Neural Image Caption Generator," 17 November 2014.

[4]L. Huang, "Entityaware Image Caption Generation" 2018.

[5]D. Sesok, " Literature Review on Image Captioning," 16 March 2019.

[6] S. Yan, F.wu, "Image Captioning" 11 January 2019.

[7]A. Karpathy "Alignments for Generating Image Descriptions," December 2014.

[8]J. Donahue, "Visual Recognition and Description," 31 May 2016