Generating Captions for Images Using Neural Networks

ISHAAN TANEJA[1], SUNIL MAGGU[2]
[1, 2] *Maharaja Agrasen Institute of Technology, Delhi, India*

Abstract- Different concepts in the field of Artificial Intelligence are on the rise these days, generating captions from given images, being one of them. The ability to train a machine to be provided with an image and then it being able to describe the details around the same can be used in various applications, be it robotics, or other businesses. The primary purpose of this paper is to recommend a model which describes images and provides its captions using concepts of Deep Learning and Machine translation. The model aims to detect different types of objects around an image, recognize the relationships between them and then generate the desired captions. The model, developed in Python, is trained using the Flickr 8K dataset in order to accomplish the same. The model was developed using Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM). In addition to discussing VGG16, a variant of CNN that has proven useful in our use case, the paper delves deeply into the fundamental notions of CNN. The wider use of this model is to help the masses, wherein, it can be used in image indexing to help those with visual impairments, also it can be implemented on some social networks and can be used in other applications as well.

I. INTRODUCTION

Artificial Intelligence has had its fair share of challenges, generating captions for images being one of them. With the advancements in image categorization and object identification, the issue of picture captioning—which entails automatically producing one or more phrases to comprehend an image's visual content—has become viable.Due to its superior performance in a type of learning that has the potential to be extremely beneficial for real-world applications, deep learning has garnered a lot of interest. This model involves the usage of Deep Learning and Natural Language Processing concepts in order to recognize the image context and its features in order to generate a description for the same. While the current applications of this technology might be limited to images, it can be used to study and create a platform which can also provide the generation of captions for videos, to help security systems in the future. We have examined a few approaches to achieving successful outcomes to produce improved results. And the approach used in our model is the one which involves the use of Convolutional Neural Networks as well as Recurrent Neural Networks. The process is divided into two phases: wherein, the first phase involves the use of CNNs to detect the features around the image which will help us in getting the most minute of details required to generate the caption text. Here, VGG16, the 16 layered CNN model is used for recognition of features. The second phase involves the use of RNN, which is trained with the captions for each feature of the images.

The purpose of this paper is to generate a highly efficient model which generates captions to given images in an accurate way and to maximize the efficiency of the model, the BLEU (Bilingual Evaluation Understudy) Scores are to be calculated for the model. It is an algorithm that has been used to gauge how well machine translations have performed. To assess the caliber of our automatically produced caption, we employed BLEU.

The following are the sections that the paper has been split into:
Section 1: This section introduces the topic of the study and discusses the goals of the research. The purpose of the research is briefly described in this section.
Section 2: This part informs the reader about the relevant work that has already been done in this area.
Section 3: This part outlines the technique that was used to develop the model in depth. It also includes a number of flowcharts and diagrams that may be used to quickly, clearly, and simply grasp how things function.
Section 4: This section examines the outcomes presented by the created model.

Section 5: This part offers the research's conclusion, aids in determining the research's potential future directions, and lists all the references used while writing the paper.

II. RELATED WORK

Research in the problem of generating captions for images has been going on for a while now.

An image is interpreted by computer vision as a two-dimensional array. Therefore, picture captioning has been characterized as a language translation issue by Venugopalan (et al) [2]. Work on this has mostly focused on Recurrent Neural Networks (RNNs) [3,4]. Which performs machine translation. By creating an image caption word-for-word, the same application is expanded by the image caption generator.

Utilizing ranking metrics and BLEU, evaluation is done similarly to [1][5][6], with the result that performance improves as the size of the picture collection grows. To comprehend how dataset size affects generalization, about five datasets have been used.

Finding the right model for caption generation had many obstacles in its way.

Rennie et al. (2017) [7] suggested a strategy based on reinforcement learning for training image caption generators. It made use of a self-critical sequence training technique to outperform conventional maximum likelihood training by directly optimizing the captioning measure.

III. METHODOLOGY

In addition to Keras, another neural network library, Tensorflow, the most popular deep learning library, has been employed.

I. Dataset Used:

The Flickr 8k dataset has been used in order to train the model. There are 5 captions per image in the dataset, which contains 8000 photos. The five descriptions for a single image aid in comprehending all the many possibilities that may occur. The dataset comes with three prepared datasets: Flickr_8k.trainImages (6,000 images), Flickr_8k.devImages (1,000 images), and Flickr_8k.testImages (1,000 images).

Fig 1- Sample images from Dataset

Image Dataset:
https://github.com/jbrownlee/Datasets/releases/download/Flickr8k/Flickr8k_Dataset.zip

Text Dataset:
https://github.com/jbrownlee/Datasets/releases/download/Flickr8k/Flickr8k_text.zip

II. Preprocessing and Model Training

The photos and their related captions are cleaned and pre-processed independently as part of the two-part process of data preparation.

A. Convolutional Neural Networks (CNN):

Convolutional Neural Networks are specialized deep neural networks that can handle data with input shapes like a 2D matrix. A few of the layers used by CNN, a deep learning neural network, to perform the required categorization include convolutional layers, pooling layers, flattening layers, and fully connected layers. Working with images requires CNN. It accepts an image as its input, gives various elements and objects in the picture weights and biases, and then distinguishes between them. A brief description of each layer is:

Convolutional layer: Uses learnable filters to perform convolutions on the input picture in order to extract regional characteristics and produce feature maps.

The network may learn complicated representations thanks to the activation layer, which introduces non-linearity by applying an activation function (like ReLU) to the feature maps.

Pooling Layer: Reduces the spatial dimensions of the feature maps while maintaining their key characteristics.

Fully Connected Layer: High-level characteristics are converted into predictions by the fully connected layer, which links every neuron in the previous layer to every neuron in the current layer.[8]

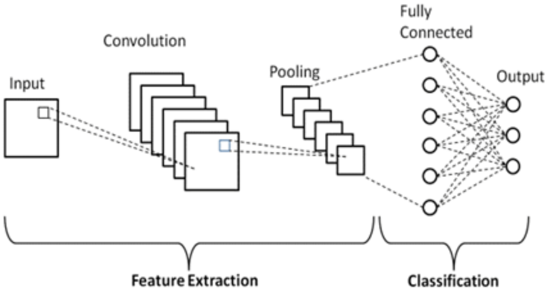


Fig 2- CNN Architecture

VGG16, a pre-trained CNN model which consists of 16 layers has been used to train the model. The ImageNet dataset was used to train the CNN-based VGG16 model. It was developed by the Visual Geometry Group (VGG) at Oxford University. 16 layers make up VGG16, 13 of which are convolutional layers and 3 of which are fully coupled layers. It is regularly employed as the initial step in picture recognition tasks. The network's fully connected layers use the characteristics retrieved by the convolutional layers to generate predictions. These layers employ the retrieved characteristics as input to categorize the input pictures into a preset set of classes.

B. Recurrent Neural Networks(RNN):

The human brain has this special capacity to make sense of previous words, learn from it and join further words in order to create a sentence. However, basic neural networks didn't have the capability to perform the same. This is where RNNs come into play. They are networks with loops that, by using their internal states to store information for a time, create feedback loops.[9]

Long Short-Term Memory (LSTM) is a form of recurrent neural network (RNN) architecture made to handle sequential input and efficiently capture long-range relationships. Their tendency to retain information for extended periods of time is basically their default habit, and "gates" are used to manage it. There are three main types of gates in LSTM, which are the input gate, output gate, and forget gate. These gates determine whether to output the cell's value, read a value into the cell, or ignore the current cell value.

The previous concealed states serve an important purpose since they are conveyed to the subsequent stage in the sequence.

The hidden state, which serves as the neural network's memory, contains the data that it has previously seen. It enables the neural network to function similarly to a human brain trying to form sentences as a consequence.

- Model Implementation:

We have implemented CNN along with LSTM to generate the desired model for generating the captions, which will take the image as input and provide text as output.

We'll use the Keras Model from Functional API to stack the model. There will be three sections to the structure:

1. Feature Extractor: The first tool will be used to shrink the dimensions from 2048 to 256. We're going to employ a Dropout Layer. CNN and LSTM will add one of them. The extracted characteristics predicted by this model will be used as input after the photographs have been pre-processed using the Xception model (minus the output layer).
2. Sequence Processor: The Embedding layer, followed by the LSTM layer, will handle the text input.
3. Decoder: To create the final predictions, we will combine the output from the previous two layers using a dense layer. An output vector with a fixed length is produced by the feature extractor and sequence processor. A Dense layer processes them after merging them. The final layer will have the same number of nodes to our vocabulary size. [9,10].

Taking into consideration previously created words and the visual context stored in the picture characteristics, the LSTM learns to construct words one at a time. The model has attention features that enable it to concentrate on pertinent picture areas while creating captions. [11] The model refines its parameters by training on paired image-caption data in order to provide captions that properly reflect the content of the input pictures. The CNN + LSTM model generates relevant and coherent picture captions by integrating visual feature extraction with sequential caption creation.

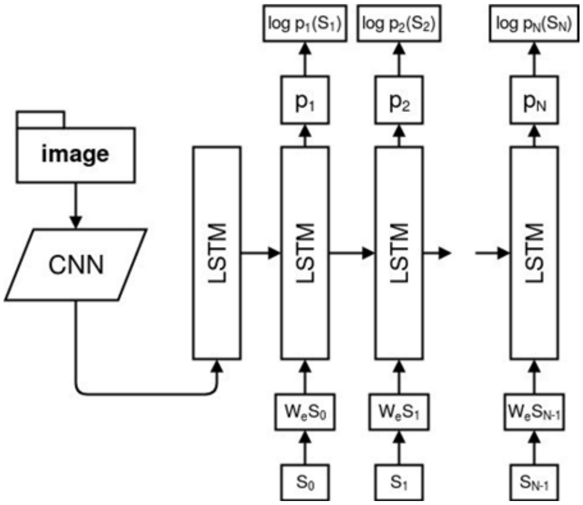


Fig- 3 CNN+LSTM Model implementation

IV. MODEL EVALUATION AND ANALYSIS

The CNN + LSTM model for creating picture captions was evaluated, and the findings are presented in the study report. The produced captions' quality, relevance, and fluency are the main areas of attention in the evaluation.

The metric used by us for checking the model performance is BLEU here.

It makes it possible to perform comparative analysis, fine-tune, monitor progress, assess datasets, and understand models. The BLEU score enables researchers to unbiasedly compare various models, follow advancement over time, optimize hyperparameters, evaluate datasets, and obtain insights into the model's strengths and limits by quantitatively quantifying the alignment between generated captions and reference captions. [12] It contributes in enhancing the caliber and efficiency of our picture caption generator model by acting as a performance benchmark.

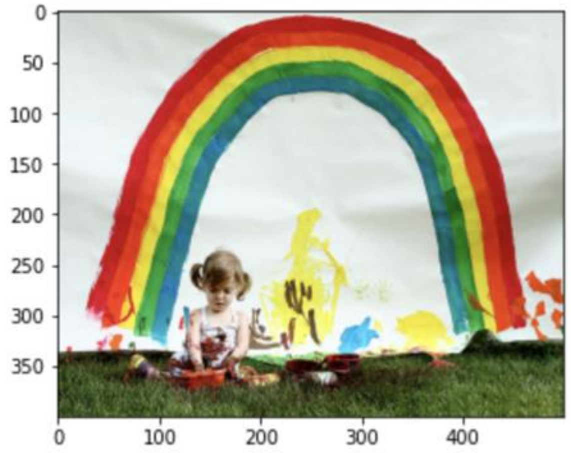
There were two kinds of BLEU scores calculated with them being:

BLEU-1(1.0, 0, 0, 0)	0.616880
BLEU-2 (0.5, 0.5, 0, 0)	0.593009

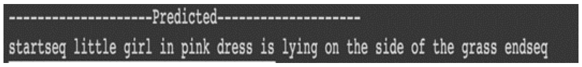
Table 2- Showcasing different BLEU scores

We provided some images for the model to provide captions for after getting trained, and they showed the following results:

Input Image 1:



Output Caption:



```
--------------------Predicted--------------------
startseq little girl in pink dress is lying on the side of the grass endseq
```

Fig- 4,5 - Sample Input and Output for the model

Image	Original Description	Predicted Description
101669240_b2d3e7f17b.jpg	man skis past another man displaying paintings in the snow	two people are hiking up snowy mountain
1002674143_1b742ab4b8.jpg	small girl in the grass plays with finger paints in front of white canvas with rainbow on it	little girl in pink dress is lying on the side of the grass

Table 2- Comparison between original and predicted description of images

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 pictures by synchronizing their operations.

We may draw the conclusion from the results that the deep learning technique used produced successful results.

The association between items in photos could be determined since the CNN and LSTM were synchronized.

Utilizing a larger dataset and training on more photos is also anticipated to improve performance. The text-to-speech technology that we have also implemented can be quite helpful to visually impaired persons and help them obtain a better feel of their surroundings because of the significant accuracy of the generated image captions.

Our model was developed using the relatively modest and homogeneous Flickr 8K dataset. The Flickr30K and MSCOCO datasets may be used to train our model, which will improve our ability to forecast the future.

REFERENCES

[1] Haoran Wang ,Yue Zhang and Xiaosheng Yu, An Overview of Image Caption Generation Methods, ,2020

[2] Sequence to sequence -video to text by Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko,2017.

[3] " Every Picture Tells a Story: Generating Sentences from Images." Computer Vision ECCV (2016) by Farhadi, Ali, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hocken-maier, and David Forsyth

[4] Show and Tell: A Neural Image Caption Generator by Oriol Vinyal, Alexander Toshev, Samy Bengio, Dumitru Erhan, IEEE (2015)

[5] Krishnakumar, K.Kousalya, S.Gokul, R.Karthikeyan D.Kaviyarasu , Image Caption Generator using Deep Learning, International Journal of Advanced Science and Technology, Vol. 29, No. 3s, (2020), pp. 975-980ISSN: 2005-4238 IJAST

[6] P. Aishwarya Naidu1, Satvik Vats,Gehna Anand, Nalina V, A Deep Learning Model for Image Caption Generation, Published: 30/June/2020

[7] Rennie, Steven & Marcheret, E. & Mroueh, Youssef & Ross, Jarret & Goel, Vaibhava. (2017). Self-Critical Sequence Training for Image Captioning.

[8] Alzubaidi, L., Zhang, J., Humaidi, A.J. *et al.* Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *J Big Data* 8, 53 (2021). https://doi.org/10.1186/s40537-021-00444-8

[9] Pranay Mathur, Aman Gill, Aayush Yadav, Anurag Mishra, and Nand Kumar Bansode,"Camera2Caption: A Real-Time Image Caption Generator", International Conference on Computational Intelligence in Data Science(ICCIDS) - 2017

[10] Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan, Show and Tell: A Neural Image Caption Generator. CVPR2015

[11] D. Bahdanau, K. Cho, and Y. Bengio. "Neural machine translation by jointly learning to align and translate. arXiv:1409.0473", 2014.

[12] BLEU: A method for automatic evaluation of machine translation. InACL, 2002 by K. Papineni, S. Roukos, T. Ward, and W. J. Zhu.

[13] Priyanka Kalena, Nishi Malde, Aromal Nair, Saurabh Parkar, and Grishma Sharma, "Visual Image Caption Generator Using Deep Learning", (ICAST-2019)