IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 11, May 2023

Image Captions Generator-A CNN-RNN Attention Model

Monika Agarwal[1], Avinash Yadav[2], Harsh Kesharwani[3]
Assistant Professor, Department of Computer Science and Engineering[1]
Students, Department of Computer Science and Engineering
Raj Kumar Goel Institute of Technology, Ghaziabad, Uttar Pradesh, India

Abstract: *The process of creating descriptions for what's happening in an image is known as image captioning. Image captioning is used to create explanations that provide context for the images. In general, image captioning is extremely helpful in a variety of applications, such as the analysis of enormous quantities of unlabelled photos and the discovery of hidden insights for ML applications that guide to create a software that provides guidance for the blind. Deep Learning Models can be used for this image captioning. Deep learning and Natural Language Processing advancements have made it simpler than ever to create descriptions for the provided visuals. Neural networks will be used in image captioning. CNN (Inception V3) is used for the encoding, which is supposed to retrieve the features from the depth of the image. RNN (Long-Short Term Memory) has been used for the decoder, and it helps to generatethe captions for the images using the features of the image*

Keywords: RNN, CNN

I. INTRODUCTION

Image captioning was a difficult task, and the captions that were generated for an image were sometimes not very appropriate. Many tasks which were difficult to perform using Machine Learning have become simpler to accomplish with the help of both Deep Learning and Neural Networks, thanks to the progress of both text processing techniques like Natural Language Processing and Neural Networks. In simplest terms, the model outputs a caption based on the input image. The potency of creating image captions is improving along with technological innovation. Variety of models have been proposed for image captioning, such as object identification models, visual attention-based image captioning, and deep learning-based image captioning. There are a number of deep learning models, such as the Res-Net model, the VGG model, the Inception-LSTM model, and the conventional CNN-RNN model. This paper explains the methodology we used to caption images (Inception-LSTM model).

In this project, we have converted an image to text description first; then, using a simple text-to-speech API, we have extracted the text description/caption and converted it to audio. So, the central part is focused on building the caption/text description whereas the second part, which is transforming the text into speech is relatively easy with the text-to-speech API.

Our model makes use of CNN-RNN to help blind people determine what an image in front of them is about. The model will convert the contents of an image and will provide the output in the form of audio.

II. RELATED WORK

We will be using CNN-RNN and Attention based model in this paper.

2.1. CNN

CNNs (Convolutional Neural Networks) are specialised deep neural networks that are used for image recognition and categorization. It is employed to process data that is represented as two dimensional matrices. It can handle images that have been resized, translated, and rotated. The visual imagery is analysed by scanning the image from right to left and from top to bottom, then taking the pertinent features out of it. The features for picture categorization are finally combined.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 11, May 2023

2.2.LSTM

Sequence prediction problems can be solved with LSTM (Long Short-Term Memory).The next word is mostly predicted using this technique. As inputs are processed using LSTM, the crucial information is carried out and the unnecessary information is deleted.We must combine CNN and LSTM in order to create an image caption generating model. Model for the Image Caption Generator: CNN with LSTM.

2.3 ATTENTION MECHANISM

The performance of the encoder-decoder model for machine translation was enhanced by the attention mechanism. By integrating all of the encoded input vectors in a weighted fashion, with the most pertinent vectors obtaining the highest weights, the attention mechanism's goal was to enable the decoder to employ the most pertinent parts of the input sequence in a flexible manner.

III. PROPOSED MODEL

As we have noticed, the vanishing gradient problem in the standard CNN-RNN model prevents the recurrent neural network from learning and being trained effectively. Therefore, in this paper, we suggest this approach to boost the effectiveness of producing captions, that are relevant, for the image and also to increase its accuracy in order to lessen this gradient descent problem. For the purpose of captioning images, we will discuss the Inception-LSTM model. In this case, LSTMs are employed for decoding, while the Inception V3 Architecture is used for encoding. When an image is delivered to Inception V3, it first extracts the image's features before employing a vocabulary that has been created using training caption data. With these two parameters as input, we will now train the model. We will test the model after training. The flow diagram for our suggested model in this research paper is shown below.

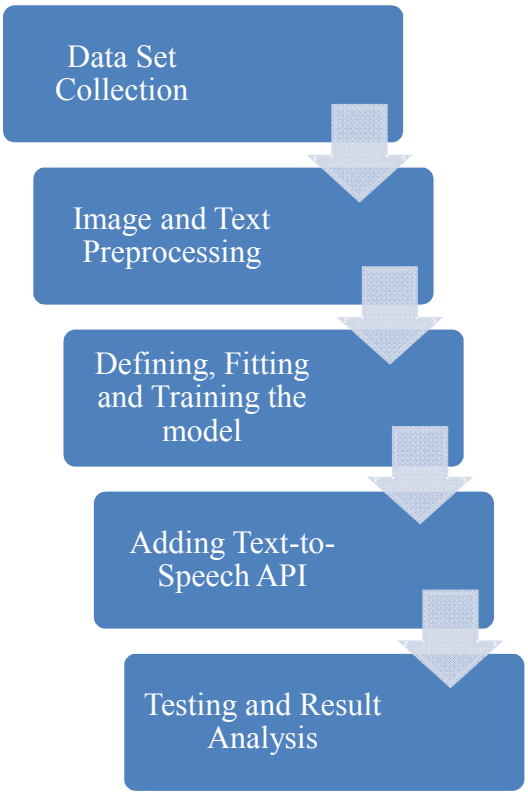

Fig.1: Model Flow Diagram

3.1.DATA SET COLLECTION

Many other data sets, including ImageNet, COCO, FLICKR 8K, and FLICK 30K, can be utilized to train the deep learning model that will provide captions for the photos. The FLICKR 8K data set is used in this paper to train the

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 11, May 2023

model. The FLICKR 8K data set is effective for training the deep learning model that generates image captions. 8000 photos make up the FLICKR 8K data set, of which 6000 images can be used to train a deep learning model, 1,000 images to develop it, and 1,000 images to test it. Each image in the Flickr Text data set has five captions, each of which describes an action that is taking place in the image.

3.2.IMAGE PREPROCESSING

To provide the images as input to the Inception V3 after importing the data sets, we must first preprocess the images. We must scale each image to the same size, which is 224X224X3, because we cannot feed various-sized images through the Convolution layer like Inception V3. Moreover, we are transforming the photos to RGB using the cv2 library's built-in capabilities.

3.3.TEXT PREPROCESSING

Once the captions have been added to the FLICKR text data set, we need to preprocess them so that there won't be any confusion or difficulties when building the deep learning model's vocabulary from the captions. The captions must be examined to determine whether any numbers are there; if so, they must be taken out. Next, the given data set's white spaces and missing captions must be eliminated. To avoid ambiguity while developing the vocabulary and training the model, we must convert all uppercase letters in the captions to lowercase. The features of the image are inputs as well as previously constructed words because this model generates captions one word at a time. These words are attached at the beginning and end of each caption to inform the neural network when the model is being trained and tested where each caption begins and ends.

3.4.DIFINING AND FITTING THE MODEL

Following the data collection, image and caption preparation, and vocabulary development. The model for caption generation must now be defined. Inception-LSTM (Long-Short Term Memory) is the model we've suggested. The encoder in this model, Inception V3, extracts the image characteristics from the images, turns them into single-layered vectors, and then passes them as input to LSTMs. Long-Short Term Memory is utilized as a decoder, which uses a vocabulary dictionary and visual attributes as input to generate each word of the caption in turn.

3.4.1. INCEPTION V3

On the ImageNet dataset, it has been shown that the picture recognition model InceptionV3 can reach more than 78.1 accuracy. The model is the outcome of multiple notions that different scholars have created throughout time. Several symmetric and asymmetric building elements, such as convolutions, average pooling, maximum pooling, concatenations, dropouts, and completely connected layers, make up the model. The model makes considerable use of batch normalization, which is also applied to the activation inputs. Softmax is used to calculate the losses.

3.4.2. LSTM

Long-Short Term Memory Networks are used to create captions. The first layer of an LSTM generates the first word of the caption using training data when we provide it an image feature vector and a vocabulary as input. The image feature vector and previously created words are used to generate the subsequent words in a caption. In order to create the caption for the provided image, all of these words are finally concatenated. The sophisticated RNNs that can retain data over extended periods of time are called long-short term memory cells. The vanishing gradient issue that Recurrent Neural Networks have can be resolved by these Long Short Term Memory Networks. Traditional RNNs have a vanishing gradient problem, which prevents them from remembering extended sequences of data. RNNs can't remember significant words that were generated earlier and are needed to generate new words in the context of caption production. For instance, suppose you had to guess the final phrase of the following sentence: "I'm from Germany. I have excellent German-speaking ability." It is crucial to remember the first word Germany, which is impossible with a regular RNN, but not with a long short-term memory network. Therefore, LSTMs rather than conventional RNNs are chosen for caption production.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 11, May 2023

3.5.CONVERTING CAPTIONS FROM TEXT TO AUDIO

Developers can produce speech that sounds like human voice using the Text-to-Speech API. The Text-to-Speech API can be used to turn a string into audio data. We can choose a particular voice or modify the output's pitch, volume, speaking rate, and sampling rate, among other configuration options, for speech synthesis.

3.6. MODEL TESTING

The outcome of the test set of captions is assessed using the BLEU (Bilingual Evaluation Understudy) tool. One method for evaluating how closely two reference sentences resemble each other is the BLEU metric. It provides a value between 0 and 1 in return for a single hypothesis sentence and numerous reference sentences.

IV. RESULT ANALYSIS

After model creation and fitting, 15 training epochs were used to train our model. Early training epochs are shown to have very low accuracy and generated captions that are not closely connected to the given test pictures. If the model is trained for at least 8 epochs, the captions generated are somewhat similar to the test images supplied. As seen in the accompanying images, the model's accuracy improves and its captions become more similar to the test images when it has been trained for 15 epochs.

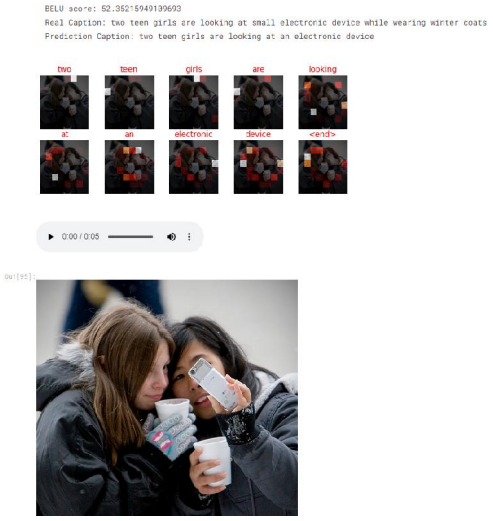

Fig.2: Caption generated for the given test image after 8 epochs.

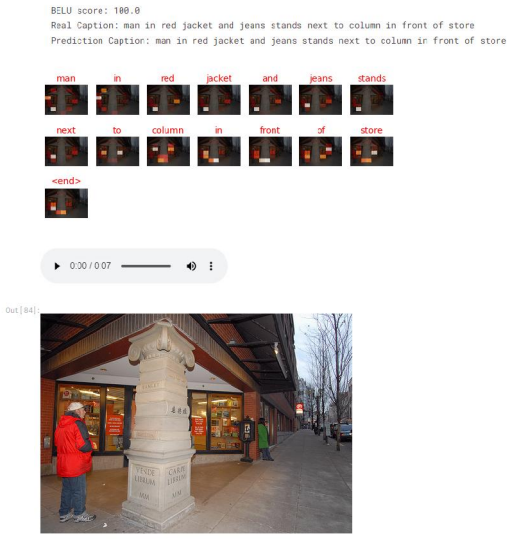

Fig. 3: Caption generated for the given test image after 15 epochs.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 11, May 2023

Here, we can see that the BLEU score for caption generation is only 52.35, which is not a very good score. Because of this, we decided that our model needed more training, so we increased the number of epochs. The result is shown in Fig. 3, where we can see an improvement in model accuracy and captions that are very similar to the test images that were provided.

V. CONCLUSION

In this paper, a deep learning model for image captioning is proposed. For each of the provided images, captions were created using the Inception-LSTM model. The model has been trained using data from the Flickr 8k dataset. The Long Short Term Memory units get the picture features once the visual characteristics are retrieved using the Inception V3 architecture. The language developed during the training phase is subsequently used to generate captions. Comparing this Inception-LSTM model to CNN-RNN and VGG Model, we can say that it is more accurate. This model works well when used with the Graphic Processing Unit. When analysing huge amounts of unstructured and unlabelled data to find patterns in the photos that will be used to direct self-driving cars and develop software that will assist the blind, our deep learning model for picture captioning is immensely beneficial.

FUTURE SCOPE

We discussed how to provide captions for the images in our paper. Even though deep learning is sophisticated, it is currently not able to generate correct captions due to a variety of issues, such as hardware needs issues, a lack of proper programming logic or models, and the fact that machines cannot think or make decisions as accurately as humans do. We therefore anticipate being able to produce captions with greater accuracy in the future thanks to technology advancements and deep learning models.

REFERENCES

[1]. Sreejith S P, Vijayakumar A "Image Captioning Generator using Deep Machine Learning", International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-5 | Issue-4, June 2021, pp.832-834

[2]. PrekshaKhant, Vishal Deshmukh, Aishwarya Kude,PrachiKiraula , "Image Caption Generator using CNN-LSTM", International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2395-0056, Volume: 09 Issue: 01 | Jan 2022.

[3]. Ali Ashraf Mohamed, "Image Caption Using CNN and LSTM",2020.

[4]. S. Amirian, K. Rasheed, T. R. Taha and H. R. Arabnia, "Image Captioning with Generative Adversarial Network," 2019 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 2019, pp. 272-275, doi: 10.1109/CSCI49370.2019.00055.

[5]. H. -Y. Hsieh, J. -S. Leu and S. -A. Huang, "Implementing a Real-Time Image Captioning Service for Scene Identification Using Embedded System," 2019 16th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON), Boston, MA, USA, 2019, pp. 1-2, doi: 10.1109/SAHCN.2019.8824961.

[6]. Aghasi Poghosyan, Aghasi Poghosyan, "Long Short-Term Memory With Read-Only Unit In Neural Image Caption Generator." 8312-163 ©2017 IEEE.

[7]. Dataset Link: https://www.kaggle.com/datasets/srbhshinde/flickr8k-sau

[8]. Mathur, Pranay, Aman Gill, Aayush Yadav, Anurag Mishra and Nand Kumar Bansode. "Camera2Caption: A real-time image caption generator." 2017 International Conference on Computational Intelligence in Data Science(ICCIDS) (2017): 1-6.

[9]. S. Das, L. Jain and A. Das, "Deep Learning for Military Image Captioning," 2018 21st International Conference on Information Fusion (FUSION), Cambridge, UK, 2018, pp. 2165-2171, doi: 10.23919/ICIF.2018.8455321.