IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 6, June 2022
Impact Factor: 6.252

Audio and Caption Generation using ML

R P Rajeshwari[1], Thammineni Pushyamitra[2], Madhuri G[3], B Pavani Kumari[4].
Assistant Professor, Department of Computer Science and Engineering[1]
Students, Department of Computer Science and Engineering[2,3,4]
Rao Bahadhur Y. Mahabaleshwarappa Engineering College, Bellary, Karnataka, India

Abstract: *Using deep learning and computer vision, image caption generation identifies the context of an image and comments on it with the most relevant captions. It entails an image's epithet combined with English keywords, as well as datasets provided during model training. We produce a vectorized representation of an image using a deep Convolutional neural network, which is then fed into a long-short-term memory (LSTM) network, which provides captions. LSTM (Long short-term memory) is also one of the RNN (Recurrent Neural Network) constructs, mainly utilized in the deep learning discipline. The vivid applications of image captioning are natural language processing applications, recommendations in editing applications, helpful for virtual assistants, for indexing the image, for the blind or partially-sighted people, for the internet community, etc.*

Keywords: Deep Learning and Computer Vision

I. INTRODUCTION

The captioning of a given image was one of the toughest tasks to perform and was unable to generate the accurate captions that are relevant to the given image. With the development of Deep Learning Neural Networks and Natural Language Processing, many formerly challenging tasks employing Machine Learning have become easier to accomplish. Many of the Deep Learning techniques which are helpful in various image activities such as image detection, classification of images, description of images, and other vivid AI applications. Image captioning has become one of the most extensively used technologies in the modern period. An ability to organically illustrate the English sentences that are properly framed for a given image is an unusual experimental fault. Although it has the potential to have a huge impact, such as by clearly assisting visually impaired individuals in gaining a better understanding of the images on the internet, a PyAudio function is used.

Furthermore, there are built-in applications that produce and provide a caption for a certain image, all of which are accomplished using deep learning models. Natural-language processing and computer vision are coupled by the challenge of automatically summarising the content of a photo using the pyaudio module. Image captioning refers to the procedure of creating textual descriptions based on the components of an image. In particular, the given model accepts an image as an input and provides the caption as an output. The applications of image captioning have increased tremendously. Some of them include web development, automatic self-driving cars, recommendation applications, and also use in search engines to easily identify the required resources or web content. The different techniques used for image captioning include image captioning with a focus on visual attention, object identification models, and image captioning with Deep Neural Networks. Some of the deep neural network learning models present for the image caption generation are the CNN:RNN, the InceptionV3, the VGG-LSTM, and finally, the ResNet-LSTM model. To get the most accurate results or the most relevant description for a given image, we have made use of the ResNet-LSTM model in the following paper.

Advancements in nlp and computer vision have recently been made and are being used to provide automatic image caption production [4]. This research develops a generative CNN-LSTM model that outperforms human baselines by 2.7 BLEU-4 points and comes close to matching human performance and the current state of the art (3.8 CIDEr points lower). Experiments on the MSCOCO dataset set demonstrate that in the vast majority of situations, it provides sensible and correct captions, and hyperparameter adjustment with offsetting the downsides of over fitting by reducing dropout and increasing the number of LSTM layers. We also show that despite varied prior contexts, semantically close emitted words modify the LSTM hidden state in identical ways, and that hidden state divergences occur exclusively when semantically distant words are emitted. The relationship between trained word embeddings and the LSTM hidden states now has semantic value. This is, as far as we know, a new contribution to the literature.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 6, June 2022
Impact Factor: 6.252

II. LITREATURE SRUVEY

Liu, Shuang and others [1] propose two deep neural network learning models: Convolutional_ Neural_Network-Recurrent_Neural_Network (CNN:RNN) based Image Captioning and Convolutional_Neural_Network-Convolutional_Neural_Network (CNN:CNN) based Captioning of image. The CNN:RNN framework employs Convolutional Neural Networks for encoding and Recurrent Neural Networks for decoding. The images are transformed to vectors using CNN, and these vectors are referred to as image features, and they are then fed into recurrent neural networks as input. The actual captions for the project are obtained using RNN's NLTK library. Just CNN is employed for picture encoding and decoding in the CNN: CNN-based framework. The NLTK library is utilised to get the exact word for the supplied image using a vocab dictionary that is mapped with image characteristics. As a result, the caption is error-free. The train, which is composed of several models that are given at the same time of convolution techniques concurrently, is unquestionably faster than the train, which is composed of a continuous flow of recurrently repeated of these techniques. In comparison to the CNN: RNN Model, the CNN: CNN Model requires less training time. The CNN: RNN Model takes longer to train since it is consecutive, but it has a lower loss than the CNN: CNN Model.

Ansari Hani [2] suggested a method for image captioning that used an encoding decoding approach. Using the CNN model for encoding, the initial steps were made, and feed forward networks, notably LSTM, were used in place of RNNs. The main contributions of LSTM were that it showed that it did not gain a representation of the picture vector at each step and revealed that it was capable of providing results based on the state of the art that were comparable to earlier work. The top layer of images was shown as a big CNN in their works, and therefore the top-down technique was termed. And, instead of detecting each object, we created models that could learn from one end to the other. They also mention two more image captioning approaches in this section: captioning based on retrieval and captioning based on templates. Retrieval-based captioning is a method of storing training photos and captions in separate locations. For the test image and captions in the new scope, correlations are determined and the highest-valued correlation caption is obtained from the caption dictionary as the caption for the provided image. They used prototype-based description as a technique in this paper. To construct the captions, they used the Inception V3 model as an encoder and attention mechanism, while GRU is used as a decoder.

Subrata Das, Lalit Jain, and others [3] proposed an approach. This methodology primarily explains how the deep neural network learning algorithm works when applied to captioning military images. It is primarily based on CNN:RNN. For image encoding, they employed the Inception model, and Long-Short-Term-Memory (LSTM) Networks to reduce the gradient descent difficulty. In this research, they propose an Inception-LSTM (ICLSTM) traffic classification approach for identifying encrypted traffic services. This approach turns traffic data into common grey pictures, then extracts critical features and performs successful traffic categorization using the ICLSTM neural network. To address the issue of category imbalance, distinct weight parameters are specified for each category independently in the training phase to make the identification effect more balanced and appropriate for different categories of encrypted data. The approach is validated on the public ISCX 2016 dataset, and the results of five classification trials reveal that the system's accuracy is greater than 98 percent for both conventional and VPN encrypted traffic service identification.

Moses Soh [4], Recent advancements in nlp and machine vision systems are used to provide automatic image caption production. This research develops a generative CNN-LSTM model that outperforms human baselines by 2.7 BLEU-4 points and is close to matching human performance and the current state of the art (3.8 CIDEr points lower). Experiments on the MSCOCO dataset set demonstrate that in the vast majority situations, it provides sensible and correct captions, and hyper parameter adjustment with withdraws and the many of LSTM layers allows us to mitigate the disadvantages of over fitting. We further show that, despite diverse prior contexts, semantically close emitted words modify the LSTM hidden state in identical ways, and that hidden state divergences only occur when semantically distant words are emitted. The relationship between trained word embeddings and the LSTM hidden states now has semantic value. This is, as far as we know, a new contribution to the literature.

III. DISADVANTAGES OF EXISTING SYSTEM

From the literature survey, it is observed that there are many flaws in the existing system, which result in the failure of the model to generate accurate and efficient captions for a given image. Some of the flaws that can be observed in the existing models, such as

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 6, June 2022
Impact Factor: 6.252

1) In the first conventional model, the CNN: CNN model, the CNN serves as both an encoder and a decoder. In CNN, there are many max pooling and padding layers, which results in huge data loss and makes the model a center-centric model in which CNN is employed for both encoding and decoding in this case. We find that the CNN: CNN model has a large forfeiture, which is unacceptable since the produced comments will be inaccurate and unrelated to the provided image.

2) To overcome the huge data loss observed in the previous CNN:CNN model, the CNN:RNN model was introduced, which resulted in less data forfeiture compared to the previous CNN:CNN model. As we encountered various other problems in this model also, the major problems include the training time and the vanishing gradient decent problem. To get accurate captions as a result, the model needs to be trained efficiently. The model should be trained with large datasets, requiring a large amount of time to train the model. The training time is directly proportional to the efficiency of the model. The training time has a huge impact on the overall model's efficiency, resulting in the training time issue. The Vanishing Gradient Descent problem is an example of this. The inputs are compared with the outputs to calculate the loss rate using a parameter called the gradient. Commonly, ANNs and RNNs face the gradient descent problem. The gradient is the proportion of changes in weights to changes in the neural network's output error. The slope of the gradient is the activation function of the neural network. The slope is inversely proportional to the training time, which means that the steeper the slope, the less the training time, and the model grasps at a higher speed. The increase in the inner layers also results in data loss. The Gradient Descent problem has a great impact on the long-term sequence learning in RNN, and the recall capacity is reduced, which results in the storage of words in concealed memory for a shorter duration. As a result, the RNN has difficulty interpreting the captions for the supplied images during training.

IV. PROPOSED MODEL

It becomes difficult for the intermittent Neural Network to learn and get trained efficiently. As we've shown, there's a vanishing gradient problem when exercising the typical CNN-RNN model, which makes it delicate for the intermittent neural net to study and be systematically trained. Hence, it is intended to palliate the shrinkage issue. We propose this model in this exploration in order to ameliorate the effectiveness of producing captions for images as well as the delicacy of the captions. The architecture for our proposed model is shown below.

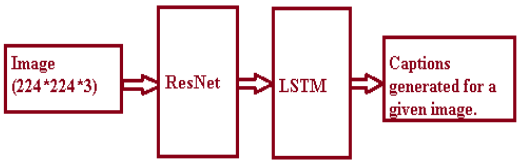

Fig 1: Architecture of ResNet-LSTM model

The ResNet-LSTM model for image captioning will be explained in this article. Encoding is done with the ResNet Architecture, and decoding is done with LSTMs. We'll train the model using these two parameters as input after passing the picture to ResNet (Residual Neural Network), which extracts visual characteristics with the aid of vocabulary developed from training data. We'll put the model to the test when it's been trained. The system overview for the proposed model is given below.

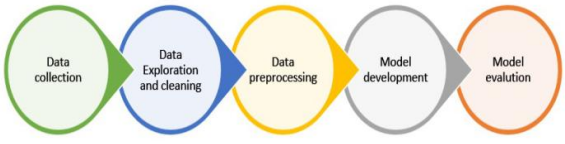

Fig 2: The system overview of image caption generator.

4.1 Dataset (Data Collection)

To generate captions, we need to train the model with a sample dataset. There are many datasets which are available, such as ImageNet, COCO, Flickr_8K, and Flickr_30K, and are used to train the deep learning models for caption generation. In this paper, we have made use of the Flickr_8K dataset to train the deep learning models for effective training of the models. The Flickr 8K dataset contains 8000 photographs, with 6000 images deployed for training, 1000

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 6, June 2022
Impact Factor: 6.252

for development, and 1000 for testing the deep learning model. The Flickr textual data dataset includes captions (set of 5) for every image. Each of them explains the behavioural patterns represented in the photo.

4.2 Data Exploration and Cleaning

We must preprocess the images after loading the data sets in order to use them as input to the ResNet. We must resize each image to the same size of 224X224X3; we can't use the convolution layer to input photos of varied sizes like we can with ResNet. We're also using the cv2 library to transform the pictures into RGB. To prevent ambiguity during vocabulary building and model training, we must transform all uppercase letters in the captions to lowercase letters. The <start seq> and <end seq> are appended at the front and last of every caption during training and testing because the model used in this paper produces textual descriptions one by one, with previously created words and picture attributes as inputs.

4.3 Data Preprocessing

It is not possible for any neural networks to process the strings as input; we must convert the string captions to numbers. Encoding captions is the term for this method. We must establish a new space where all words in each caption are taken after preprocessing the captions supplied in the training data set. Now we must assign numbers to the terms in the vocabulary. That location has now been referred to as the vocabulary library. We'll number each caption with the help of this lexical library by numbering their terms. Each word in a caption is numbered by reference to its value in a previously created vocabulary library.

4.4 Defining and Fitting the Model (Model Development)

After gathering the dataset, preparing the images and descriptions, as well as developing vocabulary, we must now define the model for caption generation. The model proposed in this paper is ResNet-LSTM. ResNet is utilised as an encoder in this model, extracting image characteristics from photos and converting them to a single-layered vector, which is then passed as input to LSTMs. Long Short Term Memory is employed in the form of a decoder, and the picture attributes are fed as load and a vocabulary library for progressively generating every word of the captions.

RESNET50

With the development of transfer learning, it became easy to use ResNet, a pre-trained model for a variety of picture recognition and classification tasks. It is an example of deep neural networks. ResNet is a pre-trained model for identifying photographs on the ImageNet data set, we utilize it instead of the Deep Convolutional Neural Network. As a result, we may reduce the cost of computation and training time by employing the concept of transfer learning. If we had used a CNN that had not been pre-trained, the calculation cost would have increased, and the model would have taken longer to learn. We can improve the model's accuracy by utilising a ResNet pre-trained model. Resnet50 is made up of 50 convolutional layers.We make use of ResNet50, which is one type of CNN and most commonly used for image caption generation. Usually, the ResNet50 is used to classify the images, but we don't need to classify the image output here. So, the last layer of Restnet50 is omitted. Instead, we make use of the pre-last layer as the output layer, intending to obtain the picture characteristics as a flattened vector with only one layer output. ResNet is mostly used instead of classic CNNs because it includes residual blocks with skip connections, which alleviates CNN's vanishing gradient issue. When compared to CNN, ResNet greatly reduces the loss of input data. ResNet outperforms and outperforms standard CNN and VGG when it comes to image categorization and extracting picture attributes. The operation and importance of the ResNet block compared to standard CNN are represented in the diagram below. The ResNet block's operation and importance in comparison to regular CNN are depicted in the diagram below.

The standard CNN has different combinations of layers such as the Convolutional Layer, an Activation Layer called the ReLU, and a Pooling Layer. The output of the typical CNN after processing the input is: H (x) = f (wx + b) or H (x) = f (x). Here H (x) refers to the output, x is the input, w is the multiplied weights, b is the added bias, and f () is the activation function. In CNN, the input is not equal to the output. Hence, this function used to get the photo features will result in erroneous output with less accuracy. In the ResNet model, skip connections are present at the base. The skip connections are the ones where the gradient takes to get to the output layer. The output is equivalent to the input when these skip connections are used. i.e., H(x)=x+ f(x), where f(x)=0, as shown in the diagram. As a result, we can see that when the

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 6, June 2022
Impact Factor: 6.252

photos are fed into the ResNet model, then the output will be the same as the input, without adding bias. Therefore, when ResNet is utilised to extract picture features, there is a small amount of data loss or loss of image features. As a result, ResNet outperforms the classic CNN model in retrieving visual information.

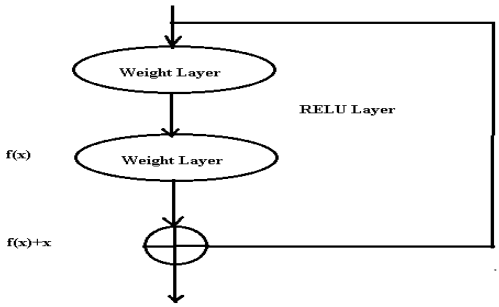

Fig 3: Block diagram of RNN.

Convolution, Activation layer, Normalization of batches, Pooling layer, and flattening layer are some of the layers of the Residual NN. The following is a description of the various layers of a Residual Neural Network and how they work: Layer of Convolution: After passing the image through the convolution layer, it is transformed into pixel values. On the image, feature maps (filters) are moved over the image, and convolution is conducted. The ReLU layer receives the output of this convolution layer. When the Rectified Linear Unit layer receives the convoluted image matrix, It alters the pixel values by using the ReLU activation function. The Batch Normalizing layer receives the output from the ReLU layer and performs normalisation and standardisation operations on it in the network by adding extra layers to scale the input to a common size, resulting in a speedier and more stable network. This layer's output is then transferred to the merging layer. The merging layer is commonly utilised for reducing the dimensions of a picture. A pooling layer sweeps a tiny matrix window across its input in general. It simply chooses the highest value in each sub matrix. As a result, the dimension of the input matrix is reduced, besides more forfeiture.

The layer should be flattened in this ResNet, the flattened layer's general function is to transform the picture featured matrix to a single-layered vector. Following the max pool function which is applied to the matrix, which is used to reduce the dimension of the matrix and transform it into a vector with one layer using picture features. After the matrix is transformed to a single layer, image features are extracted and given to the LSTM Unit, which generates every term in the captions using the vocabulary that is created. As a result, we use a single-layered vector to extract picture characteristics from the ResNet model, which is then transmitted to the LSTM Networks for caption production.

LSTM

LSTM refers to the Long-Short-Term-Memory (LSTM) that is utilised to build captions utilising the image feature vector, which is the output of ResNet and vocabulary created during the training of data set captions. The first layer of the LSTM creates the first word of the caption using training information when we give it the picture feature vector and vocabulary as inputs.

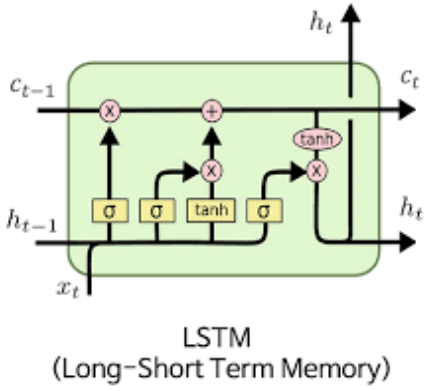

Fig 4: Cell Structure of LSTM

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 6, June 2022
Impact Factor: 6.252

The image feature vector and previously created words are used to generate the next words in a caption. Finally, the caption for the provided image is created by concatenating all of these terms. Advanced RNNs with long-short-term memory cells can remember data over lengthy periods of time. The LSTMs are used to overcome the problem of vanishing gradient in RNN. As a result, in the instance of caption generation, RNNs are unable to recall crucial words that have already been formed and are required for future word generation. LSTMs are most commonly used for caption generation over RNNs. Cells in LSTM Networks are made up of various gates, such as input, forget, and output gates, which help to store only the required data and forget the unwanted data.

V. RESULT

After defining and fitting the model, the next step is to train the model. We have trained our model for 20 epochs. As we have trained with a lower number of epochs, the accuracy of the captions generated is very low and is completely irrelevant to the images given during the test. But, when we train our model with 50 epochs, the accuracy has improved linearly, and the textual descriptions produced are more closely related to the provided photos.

VI. CONCLUSION

This research proposes a deep learning model for image captioning. To produce captions for each of the images, we employed the RESNET-LSTM model. For the purpose of training the model, the Flickr_8k data set was employed. The convolution layer's architecture is known as RESNET. The image features are extracted using the RESNET architecture, and these image features are fed into long-term short-term memory units, where captions are constructed using language acquired during the training phase. In comparison to CNN:RNN and VGG models, the ResNet:LSTM model has greater accuracy. To guide some of the applications like self-driving vehicles and some other applications to assist the visually impaired people the deep learning image caption model is used for effective analysis of huge amount of unlabeled data to discover the samples in pictures. An image might have a lot of information in it. Instead of only describing a single target item, the model should generate description sentences for many primary objects for photos containing numerous target objects. A global image description system capable of handling several languages should be designed for corpus description languages in various languages. It's tough to assess the results of natural language generation systems. Subjective review by linguists is the best way for determining the quality of machine-generated texts, but it is difficult to find. The evaluation indicators should be upgraded to bring them closer to human expert assessments in order to improve system performance. The speed at which the model is trained, tested, and generates phrases is a serious issue and one that is of great concern to increase performance.

REFERENCES

[1]. Liu, Shuang & Bai, Liang & Hu, Yanli & Wang, Haoran. (2018). Image Captioning Based on Deep Neural Networks. MATEC Web of Conferences. 232. 01052. 10.1051/matecconf/201823201052

[2]. A. Hani, N. Tagougui and M. Kherallah, "Image Caption Generation Using A Deep Architecture," 2019 International Arab Conference on Information Technology (ACIT), 2019, pp. 246-251, doi: 10.1109/ACIT47987.2019.8990998.

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

[4]. Learning CNN- LSTM Architectures For Image Caption Generation, Moses Soh,2016.

[5]. G Geetha, T. Kirthigadev, G GODWIN Ponsam, T. Karthik,M.Safa," Image Captioning Using Deep Convolutional Neural Networks(CNNs)" Published under license by IOP Publishing Ltd in Journal of Physics :Conference Series ,Volume 1712, International Conference On Computational Physics in Emerging Technologies(ICCPET) 2020 August 2020,Manglore India in 2015.

[6]. Simonyan, Karen, and Andrew Zisserman." Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[7]. Donahue, Jeffrey, et al." Long-term recurrent convolutional networks for visual recognition and description." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[8]. Lu, Jiasen, et al." Knowing when to look: Adaptive attention via a visual sentinel for im-age captioning."

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 6, June 2022
Impact Factor: 6.252

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Vol. 6. 2017.

[9]. Ordonez, Vicente, Girish Kulkarni, and Tamara L. Berg. ”Im2text: Describing images us-ing 1 million captioned photographs.” Advances in neural information processing systems. 2011.

[10]. Chen, Xinlei, and C. Lawrence Zitnick. ”Mind’s eye: A recurrent visual representation for image caption generation.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.S