International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

Generating Image Captions Using Deeplearning And Natural Language Processing

Author1:Bhanuchander Vootla
Computer science engineering department
Sphoorthy engineering college
Hyderabad,india
19n81a05a3bhanu@gmail.com

Author2:Sathyasai Madadi
Computer science engineering department
Sphoorthy engineering college
Hyderabad,india
Sathya@sphn.ac.in

Author3:Manish Reddy Tummala
Computer science engineering department
Sphoorthy engineering college
Hyderabad,india
Manish@sphn.ac.in

Autho4r:Mr.T.Balakrishna(phD)
Associate Professor
Computer science engineering department
Sphoorthy engineering college
Hyderabad,india
balakrishna@sphoorthyengg.ac.in

Abstract—Image captioning is the process of generating descriptions about what is going on in the image. By the help of Image Captioning descriptions are built which explain about the images. Image Captioning is basically very much useful in many applications like analyzing large amounts of unlabeled images and finding hidden patterns for Machine Learning Applications for guiding Self driving cars and for building software that guides blind people. This Image Captioning can be done by using Deep Learning Models. With the advancement of deep learning and Natural Language Processing now it has become easy to generate captions for the given images. In this paper we will be using Neural Networks for the image captioning. Convolution Neural Network (ResNet) is used as encoder which access the image features and Recurrent Neural Network (Long Short Term Memory) is used as decoder which generates the captions for the images with the help of image features and vocabulary that is built

Keywords—Image Captioning; Natural Language Processing; Deep Learning; Convolutional Neural Network; Recurrent Neural Network; Long Short Term Memory (LSTM).

1.INTRODUCTION

In earlier times, describing images was a difficult task the captions that are generated for the given image are not much relevant. With the development of Deep Neural Networks Learning and also word processing techniques like Natural Language processing, many tasks that were challenging and The difficult use of machine learning has become easy to implement with the help of Deep Learning and Neural Networks.These are very useful in image recognition, Image classification, image description and many other artificial News apps. Image captions are basically generating descriptions about what is happening in the given input image. Basically, this model takes an image as input and gives a caption to it. With the advancement of technology the efficiency of image caption generation also increases. This description of images is very useful for many applications like the self-driving cars that are being talked about now city. Image captions can be used in many machines Learning tasks for Recommendation systems.They exist many models proposed for describing images as an object detection model, image description based on visual attention and Describing images using Deep Learning. In Deep Learning there are also various deep learning models such as Inception model, VGG model, ResNet-LSTM model, traditional CNN RNN model. In this document we will explain about the model we used to describe the images, i.e. ResNet-LSTM model.

2.LITERATURE SURVEY

By the method proposed by Liu, Shuang & Bai, Liang & Hu,Yanli & Wang, Haoran et al. [1], two deep learning models namely convolutional neural network-recurrent neuron Network Based Image Captioning (CNN-RNN), Convolutional Neural Network-Convolutional Neural (CNNCNN) Image Captioning. In a CNN-RNN based framework work, convolutional neural networks for coding and Recurrent neural networks for the decoding process. By using CNN images here are converted to vectors and these vectors are called the image elements they are passed into Recurrent neural networks as input. There is NLTK in RNN libraries are used to get the actual captions for the project. In only CNN-CNN is used for both encoding and decoding images. Here is a dictionary of words is used and is mapped using Image functions to obtain an accurate word for a given image using the NLTK library. Thus error free caption generation. Consisting of many models which are simultaneously given by convolutional techniques at the same time it is certainly faster compared to the train continuous smooth repetition of this technique. The CNNCNN model has less training time compared to the CNN model CNN-RNN model. The CNN-RNN model has more training time because it is sequential but has less loss compared to CNN-CNN model.

In the method proposed by Ansari Hani et al[2] Here they used an encoding-decoding model for describing images. Here they mentioned two other models for the picture subtitles are: subtitles and search based template based subtitles. Search-based captioning is a process where the training images are placed in one space and their corresponding generated captions are placed the next range now in the new correlation range are calculated for the test image and captions with the highest value the correlation caption is loaded as the caption for the given image from the given subtitle dictionary file. Based on the prototype description is the technique they have done in this document. Here they used the Inception V3 model as their encoder and used as their mechanism of attention and GRU decoder for generating subtitles.

In the method proposed by Subrata Das, Lalit Jain et al[3] This model is mainly based on how deep learning models are

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

used to describe military images. It mainly uses a framework based on CNNRNN. They used the Inception model for this image encoding and to reduce gradient descent the problem is that they used long-term short-term memory (LSTM'S) networks.

In the method proposed by G Geetha et al[4] they have used CNN-LSTM model for image captioning. The entire flow of the model was explained from data set collection to caption generation. Here Convolutional Neural Networks was used as encoder and LSTM's was used as decoder for generating the captions

3.PROBLEM STATEMENT

As we have seen in the literature survey, there are many disadvantages of the existing model. Each existing model has its own an inherent disadvantage that makes the model less efficient and less effective accurate when the results are generated. Observed The disadvantages of all existing models are as follows:

1)In CNN-CNN based model where CNN is used for both encoding and decoding purpose we observe that CNN-CNN the model has a high loss that is not acceptable as generated the subtitles will not be accurate and the subtitles generated here will be irrelevant to the given test image.

2) While in the case of CNN-RNN based captions, yes be less loss compared to CNN-CNN based model but there is more training time. Training time affects the whole effectiveness of the model, and here we also encountered others problem. i.e.; The vanishing transition problem. The gradient is a parameter that is used to calculate the loss rate per given input parameter comparing both inputs and outputs. This gradient descent problem occurs mainly in artificial Neural networks and recurrent neural networks. The gradient is the ratio of the change in masses relative to the change in the error in neural network output. This gradient is also considered the slope of the activation function neural network. If the slope is high, then training on the model is faster and the neural network model learns faster. As hidden layers increase the loss gain, while the gradient decreases and eventually the gradient becomes zero. This gradient the problem hinders the learning of long-term sequences in Recurrent neural networks. This gradient descent problem it hinders the RNN in the process of learning and remembering. The words cannot be cached for long-term use That's why it's hard for RNN to parse the captions for a given image during training. RNN cannot save the words of larger subtitles for a longer time gradient descent problem during training, As the number of hidden layers increases, the gradient begins decrease and eventually reaches zero where the hidden key is the words in the subtitles are sent to forget the RNN gate. Therefore CNN-RNN model efficiently trained for generation image captions. Finally, we can conclude that as an RNN have a gradient descent problem generating captions for images using CNN-RNN model are not efficient and accurate.

4.PROPOSED MODEL

As we observed, using the traditional CNN-RNN model there is a vanishing gradient problem that prevents Recurrent Neural Network so you can learn and train efficiently.So to reduce this gradient descent problem, in this paper we propose this model to increase the efficiency of generating image captions and also increase the accuracy of the captions.Given below is the architecture for our proposed model.

A. System design

This framework is used to develop a model,labels for pictures according to the picture in the pictures. The basic flow illustration of the proposed system is created further, all components of this diagram are analyzed correctly.

B. Block Diagram of Proposed System

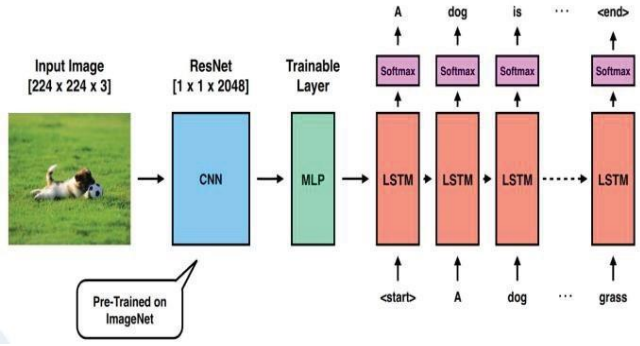

Fig. 1. Diagrammatic representation of the proposed model

C. Workflow of the Proposed System

Here the intended idea is to engender captions or descscripts for various images used as sample inputs. For to achieve this, a convolutional neural network (CNN) and Recurrent Neural Network (RNN) are two types deep learning algorithms are implemented.

Image or picture is provided by CNN recognize the objectsand scenario present in a particular image.Several other concepts have been used through CNN, such as pooling, padding and using filters etc. Basically CNN is used highlight all the features present in the sample image.

In addition, in the transfer of learning, where two major projects are considered: Glove and Inception V3. Glove defines a set of NLP vectors for regular words and Inception V3 is used to derive important functions from picture.

Further through natural language processing (NLP). images are processed into simpler formats, which makes it easier so that we can communicate with computers. NLP techniques like tokenization, derivation is done on words before passing these words to the RNN.

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

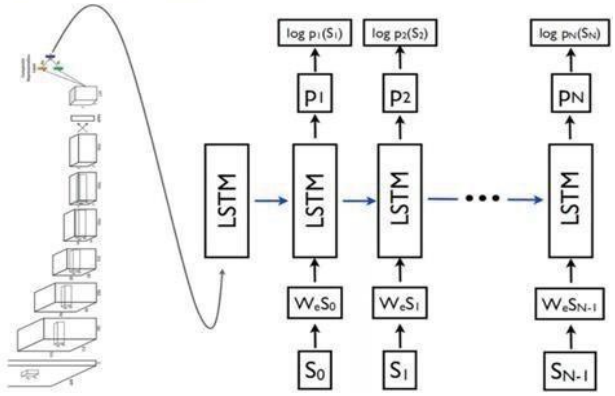

Fig. 2. Basic Workflow of the model

Finally, the set of words formed through CNN are send to the RNN. RNN model provides a meaningful description or sentence from the input provided by CNN.

D.Load Pre-Trained Network

Pre-trained networks are used to provide labels to the pictures. These pre-trained networks are present in hugeamount.Around 1k classes of images are trained by these pre-trained networks. Few pre-trained networks are: Google Net, Alex net, VGG16.

E. Use of transfer learning

Transfer of learning is done for training and classifying images into different classes. Over transfer learning, unknown classes could be trained and classified with higher accuracy. As already mentioned, transfer learning consists of two important modules are Glove and Inception V3. Here The initial V3 model with pre-trained weights is discarded for recognition of objects or figures in the input image.

F. Flickr8k Dataset

Flickr8k dataset comprises of images containing appropriate captions. Authors concentrated on Flickr8k dataset in their work because they found this dataset very suitable according to their problem statements. This dataset has suitable 1-line caption for every image which makes it suitable for the model.

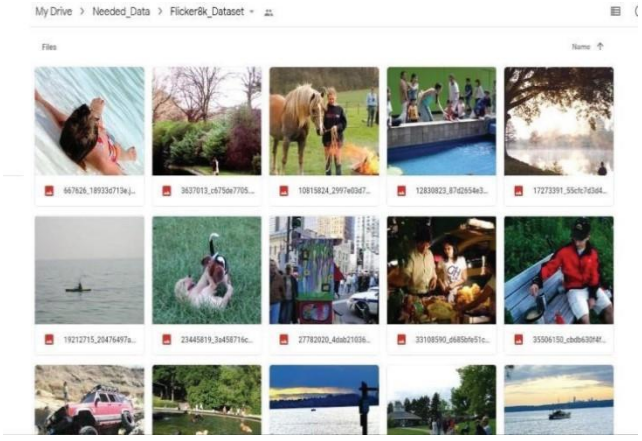

Fig.3. Images available in our dataset

G. Platform Used

In this work authors used Google Colab for coding and executing purpose. Usually, it provides a platform for machine learning and data analysis. Colab notebooks can be loaded from GitHub and these notebooks are stored in Google Drive.

5.SIMULATION AND EXPERIMENTATION

A. System Implementation

From a database maintenance perspective, it consists ofdifferent images of different categories and each aredesignated of them are provided with a meaningful label. Firststep will be to identify the object; this could be achieved image via CNN. Next, RNN applications and Long ShortTerm Memory (LSTM) helps to assign a meaningful description for images.

We will now train the model with these two parameters as input .After training ,We will test the model. Given below is the flow diagram of our proposed model in this paper.[Fig]

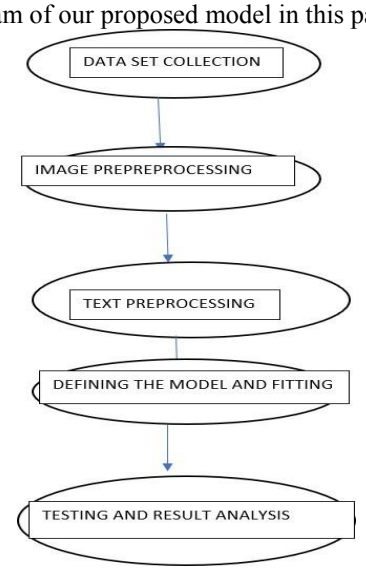

Fig 4:Model Implementation Flow Chart

A. Data File Collection

There are many datasets that can be used to train a deep learning model to generate captions for images, such as ImageNet, COCO, FLICKR 8K, FLICK 30K. In this paper, we use the FLICKR 8K dataset to train the model. The FLICKR 8K dataset works effectively in training a deep learning caption generation model. The FLICKR 8K dataset consists of 8000 images in which 6000 images can be used for training the deep learning model and 1000 images for development and 1000 images for testing the model.

This Flickr8k dataset has 8000 images in PNG format and are of high quality in nature.The resolution for the images is1024x1024.

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

B. Image PreProcessing

After loading the data sets we need to preprocess the images in order to give this images as input to the ResNet. As we cannot pass different sized images through the Convolution layer like ResNet we need to resize every image so that they are in same size i.e;224X224X3 .We are also converting the images to RGB by using inbuilt functions of cv2 library.

C. . Text PreProcessing

After fetching image captions using FLICKR's text dataset, we need to preprocess these captions so that there is no ambiguity or difficulty when generating a vocabulary from captions as well as training a deep learning model. We need to check if the captions contain any numbers, if found they must be removed and then we need to remove the white space as well as the missing captions in that dataset. We need to change all capital letters in the captions to lower case to eliminate ambiguity during vocabulary building and model training. Since this model will generate captions one word at a time and previously generated words are used as inputs along with image features as input, <start seq> and <end seq> are appended to the start and end of each caption to signal the neural network to start and end subtitles when training and testing the model

C. Defining And Training The Model

After collecting the dataset and preprocessing the images and captions and building the vocabulary. Now we need to define a model for generating captions. Our proposed model is ResNet (Residual Neural Network)-LSTM (Long Short Term Memory) model. In this model, Resnet is used as an encoder that extracts image features from the images and converts them into a single-layer vector and passes them as input to the LSTM. Long short-term memory is used as a decoder that takes image features as input as well as a vocabulary dictionary to generate each subtitle word in turn.

D. Defining And Training The Model

After collecting the dataset and preprocessing the images and captions and building the vocabulary. Now we need to define a model for generating captions. Our proposed model is ResNet (Residual Neural Network)-LSTM (Long Short Term Memory) model. In this model, Resnet is used as an encoder that extracts image features from the images and converts them into a single-layer vector and passes them as input to the LSTM. Long short-term memory is used as a decoder that takes image features as input as well as a vocabulary dictionary to generate each subtitle word in turn. with caption squentially.

E. Epoch

Epoch is simply the number of model counts behind the dates, that is, the epoch refers to one Passover via the full dataset. More about the number of epochs, next will be an improvement of the model, but only up to a certain point beyond which the model does not improve and beyond that it does the model starts to take longer to run.

F. Iteration

The term iteration refers to the number of times a batch is repeated data passes through the algorithm. They say that, when one iteration is completed when a batch of data passed through a neural network. For example, dataset 10 sample images, batch size 2 and specified number of epochs is 3, then the number of iterations will be (10/2=5, 5*3=15) 15 iterations.

Below diagram shows an example of caption generation.

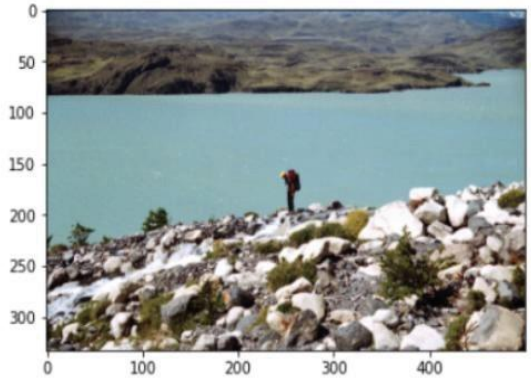

Fig. 4. Man standing at edge of Water

G. Predictions

The model will be ready to generate captions after the training and testing is completed. For this sample pictures will be given as an input and as output captions for the input picture will be generated.

6. CONCLUSION

In this paper, a deep learning model for image description is proposed. We used a RESNET-LSTM model to generate captions for each of the given images. The Flickr 8k dataset was used for model training purposes. RESNET is a convolutional layer architecture. This RESNET architecture is used to extract image features and these image features are given as input to Long Short Term Memory units and captions are generated using the vocabulary generated during the training process. We can conclude that this ResNet-LSTM model has higher accuracy compared to CNN-RNN and VGG model. This model works efficiently when we run it using a

GPU. This image captioning deep learning model is very useful for analyzing large amounts of unstructured and unlabeled data to find patterns in these images for guiding self-driving cars, for creating software to guide blind people.

FUTURE SCOPE

Generating image captions is considered essential tool how it can be applied to dissimilar meadows for their different

purposes. By generating captions for more images of the same file, these files can be organized or organized easy and fast. People who are blind or those who visually impaired people can understand the pictures according to their captions or description provided by the image captioning process.The images added to the website could be well understood has a valid description. Hence the process website generation can be done quickly by just adding images and captions for them can be found in the application form image caption generator. Hence the description of the images gains its popularity and importance in the field of image processing know-how.

In our article, we explained how to generate captions for images. Although deep learning is advanced, until now accurate caption generation is not possible due to many reasons such as hardware requirement problem, there is no proper programming logic or model to generate accurate captions because machines cannot think or make decisions as accurately as humans. So we hope to generate captions with higher accuracy in the future with advances in hardware and deep learning models. It is also envisaged that we will extend this model and create a complete image-to-speech conversion by converting image captions to speech. This is very useful for blind people.

REFERENCES

[1] Hossain, M. Z., Sohel, F., Shiratuddin, M. F., & Laga, H. (2019), "A comprehensive survey of deep learning for image captioning", ACM Computing Surveys (CSUR), 51(6), 1-36.

[2] Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2016). "Show and tell: Lessons learned from the 2015 mscoco image captioning challenge", IEEE transactions on pattern analysis and machine intelligence, 39(4), 652-663.

[3] Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Li, S., Choi, Y., & Berg, T. L. (2013). "Baby talk: Understanding and generating simple image descriptions", IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(12), 2891-2903.

[4] Chen, X., & Zitnick, C. L. (2014). "Learning a recurrent visual representation for image caption generation", . arXiv preprint arXiv:1411.5654.

[5] Kalchbrenner, N., Grefenstette, E., Blunsom, P. (2014)." A convolutional neural network for modelling sentences.", arXiv preprint arXiv:1404.2188.

[6] Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J., & Goel, V. (2017). "Self-critical sequence training for image captioning", IEEE Conference on Computer Vision and Pattern Recognition (pp. 7008-7024).

[7] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., & Bengio, Y. (2015, June). "Show, attend and tell: Neural image caption generation with visual attention",InInternational conference on machine learning(pp.2048-2057)

[8] You, Q., Jin, H., Wang, Z., Fang, C., Luo, J. (2016). "Image captioning with semantic attention", In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4651- 4659).

[9] Nguyen, D. K., & Okatani, T. (2019). "Multi-task learning of hierarchical vision-language representation", IEEE Conference on Computer Vision and Pattern Recognition (pp. 10492-10501).

[10] Sharma, S., Suhubdy, D., Michalski, V., Kahou, S. E.,Bengio,Y. (2018). "Chat painter: Improving text to image generation using dialogue",.arXiv preprint arXiv:1802.08216.

[11] Sehgal, S., Sharma, J.,Chaudhary, N. (2020, June). Generating Image Captions based on Deep Learning and Natural Language Processing. In 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO) (pp. 165-169). IEEE.

[12] N. S. Ghosh, R. Majumdar, B. Giri and A. Ghosh, "Detection of Human Activity by Widget,"2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 2020, pp. 1330-1334, doi: 10.1109/ICRITO48877.2020.9197982

,

[13] P. Das, A. Ghosh and R. Majumdar, "Determining Attention Mechanism for Visual Sentiment Analysis of an Image using SVM Classifier in Deep learning based Architecture," 2020 8th International Confe rence on Reliability, Infoco m Tech nologies and Optimization (Trends and Future Directions) (ICRITO), 2020, pp. 339 - 343 , doi : 10. 110 9/I CR I T O488 77. 20 20. 91 978 99.

[14] S. S. Khan, R. Majumdar, P. P. Maut, A. Ghosh and V. P. Mishra, "Analyzing and Applying Captured Object with Machine Learning Techniques," 2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE), 2019, pp. 287-290, doi: 10.1109/ICCIKE47802.2019.9004382.

[15] Navaney, P., Dubey, G., Rana, A., "SMS Spam Filtering Using Supervised Machine Learning Algorithms", Proceedings of the 8th International Conference Confluence 2018 on Cloud Computing, Data Science and Engineering, Confluence 2018, 2018, pp. 43–48, 8442564

[16] Tyagi, N., Rana, A., Kansal, V., "Creating Elasticity with Enhanced Weighted Optimization Load Balancing Algorithm in Cloud Computing", Proceedings - 2019 Amity International Conference on Artificial Intelligence, AICAI 2019, 2019, pp. 600–604, 8701375