INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

IMAGE CAPTION GENERATING DEEP LEARNING MODEL

[1]Mr.V.Sundara ratnam, [2]D.Keerthi, [3]D.Sushma patel, [4]G.Sri dharani, [5]D.Shravika

[1]Assistant Professor, Department of CSE(DS), Malla Reddy Engineering College for Women (Autonomous Institution – UGC, Govt. of India), Hyderabad, INDIA.

[2345]UG,Department of CSE(DS), Malla Reddy Engineering College for Women (Autonomous Institution – UGC, Govt. of India), Hyderabad , INDIA.

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—Deep Learning; Image Captioning;Convolutional Neural Networks; Recurrent Neural Networks;ResNet;Long Short Term Memory; insert (key words)

1. INTRODUCTION In earlier days Image Captioning was a tough task and the captions that are generated for the given image are not much relevant. With the advancement of Neural Networks of Deep Learning and also text processing techniques like Natural Language Processing, Many tasks that were challenging and difficult using Machine Learning became easy to implement with the help of Deep Learning and Neural Networks. These are very much useful in image recognition, Image classification, Image Captioning and many other Artificial Intelligence applications. Image Captioning is

ISSN: 2057-5688

basically generating descriptions about what is happening in the given input image. Basically ,this model takes image as input and gives caption for it. With the advancement of the technology the efficiency of image caption generation is also increasing. This Image Captioning is very much useful for many applications like Self driving cars which are now talk of the town. Image captioning can be used in many Machine Learning tasks for Recommendation Systems. There are many models proposed for image captioning like object detection model, visual attention-based image captioning and Image Captioning using Deep Learning. In Deep Learning also there are different deep learning models like Inception model, VGG Model , ResNet-LSTM model, traditional CNNRNN Model. In this paper we are going to explain about the model we have followed for captioning the images .i.e; ResNet-LSTM model

2. LITERATURE SURVEY

In method proposed by Liu, Shuang & Bai, Liang & Hu, Yanli & Wang, Haoran et al. [1], two models of deep learning namely, Convolutional Neural Network-Recurrent Neural Network(CNN-RNN) Based Image Captioning, Convolutional Neural Network-Convolutional Neural (CNNCNN)Based Image Captioning. In CNN-RNN Based frame work, Convolutional Neural Networks for encoding and Recurrent Neural Networks for the decoding process. Using CNN the images here are converted to vectors and these vectors are called image features these are passed into Recurrent neural networks as input. In RNN's se NLTK libraries are used to get the actual captions for the project. In the CNN-CNN based frame work only CNN is used for both encoding and decoding of the images. Here vocab dictionary is used and it is mapped with Image features to get the exact word for the given image using NLTK library. Thus generating the error free caption. Consisting of many models that are given at the same time of convolution techniques simultaneously is certainly quicker compared to the train the continuous flowing recurrently repetition of this techniques. CNN-CNN Model has less training time as compared to the CNN-RNN Model. The CNN-RNN Model has more training

ISSN: 2057-5688

time as it is sequential but it has less loss compared to the CNN-CNN Model.

In the method proposed by Ansari Hani et al[2] Here they have used encoding decoding model for image captioning. Here they have mentioned two more models for image captioning they are: Retrieval based captioning and template based captioning. Retrieval based captioning is the process where training images are placed in one space and their corresponding captions which are generated are placed in another scope now in the new scope the correlations are calculated for the test image and captions the highest valued correlation caption is retrieved as caption for the given image from the given set of captions dictionary. Prototype based descripting is the technique is done by them in this paper .Here they have used Inception V3 model as their encoder and they have used attention mechanism and GRU as their decoder to generate the captions.

In the method proposed by Subrata Das, Lalit Jain et al[3] This model is mainly based on how the deep learning models are used for Military Image captioning. It mainly uses CNNRNN based frame work.They have used Inception model for encoding the images and to decrease the gradient descent problem they have used Long 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 METHODOLOGY

As we have observed that using traditional CNN-RNN model there is vanishing gradient problem which hinders the Recurrent Neural Network to learn and get efficiently trained. So in order to reduce this gradient descent problem ,In this paper we are proposing this model so as to increase the efficiency of generating captions for the image and also to increase the accuracy of the captions. Given below is the architecture for our proposed model.

ISSN: 2057-5688

In this paper, We are going to explain Resnet-LSTM model for the image captioning process. Here Resnet Architecture is used for encoding and LSTM's are used for decoding .Once when the image is sent to Resnet (Residual Neural Network) it extracts the image features then with the help of vocabulary that is built using training captions data ,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.

There are many data sets which can be used for training the deep learning model for generating captions for the images like ImageNet, COCO,FLICKR 8K,FLICK 30K .In this paper, We are using FLICKR 8K data set for training the model. FLICKR 8K data set works efficiently for training the Image Caption Generating Deep Learning Model. The FLICKR 8K data set 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. Flickr Text data set consists of five captions for each given image which describes about the actions performed in the given images

With the introduction of transfer learning (using knowledge gained in training network on one type of problem and applying the knowledge in another problem of same pattern) using deep neural networks like RESNET(Residual Neural Network) which is a pretrained model for many image recognition and classification became easy. We use this ResNet model in place of Deep Convolutional Neural Network because ResNet is a pretrained model on ImageNet data set to classify the images. So by using the concept of transfer learning we are reducing the computation cost and training time. If we have used CNN which is not pretrained then the computation cost would have increased and the model takes more time to learn. By using ResNet pretrained model we are also increasing the accuracy of the model. Resnet50 consists of 50 deep convolutional neural network layers. ResNet50 is the architecture of Convolutional Neural Network that we are using in Image Caption Generation Deep Learning Model. The last layer of

ISSN: 2057-5688

Restnet50 is removed as it gives classification output and we are accessing the output of the o layer before the last one in order to get the image features as output single layered vector because we don't need classification output in this paper. The ResNet is preferred compared to traditional deep convolutional neural networks because the ResNet contains residual blocks which have skip connections that ultimately reduce the vanishing gradient problem in CNN and ResNet also decreases the loss of input features compared to CNN. ResNet is having better performance and accuracy in classification of images and extracting image features compared to traditional CNN ,VGG. Below is the figure representing the working of ResNet block and its importance compared to traditional CNN.

4 RESULTS

The output of ResNet(Image feature vector) and vocabulary built by using training data set captions are passed to Long Short Term Memory Networks to generate captions. When we pass image feature vector and vocabulary as input to first layer of LSTM, it generates the first word of the caption using training knowledge. The next words of a caption are generated with the help of image feature vector and previously generated words. Finally, all these words are concatenated to generate the caption for the given image. Long short term memory cells are the advanced RNN's which can remember data from long periods. This Long Short Term Memory Networks can overcome the problem of vanishing gradient which exists in Recurrent Neural Networks. In traditional RNN's they cannot remember long sequence of data due to vanishing gradient problem. So in the case of caption generation RNN's cannot remember important words that are generated previously and which are required for generation of future words. For example in the case of predicting the last word of this sentence,"I am from France. I speak very fluently in French". It is important to remember the starting word France which is not possible in case of traditional RNN but Long Short Term Memory Networks do not have this issue. So LSTM's are preferred for caption generation compared to traditional RNN's .

ISSN: 2057-5688

After defining and fitting the model. We trained our model for 50 epochs. It is observed that during the initial epochs of training the accuracy is very low and the captions generated are not much related to given test images. If we train the model for atleast 20 epochs then we have observed that the captions generated are some what related to the given test images. If the model is trained for 50 epochs we observe that the accuracy of the model increases and the captions generated are much related to the given test images as follows in the following figures

5 CONCLUSION

Image captioning deep learning model is proposed in this paper. We have used RESNET-LSTM model to generate captions for each of the given image. The Flickr 8k data set has been used for the purpose of training the model. RESNET is the architecture of convolution layer. This RESNET architecture is used for extracting the image features and this image features are given as input to Long Short Term Memory units and captions are generated with the help of 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 the model with the help of Graphic Processing Unit. This Image Captioning deep learning model is very much useful for analyzing the large amounts of unstructured and unlabeled data to find the patterns in those images for guiding the Self driving cars, for building the software to guide blind people.

6 REFERENCE

[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

ISSN: 2057-5688

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] GGeetha,T.Kirthigadevi,G GODWIN Ponsam,T.Karthik,M.Safa," Image Captioning Using Deep Convolutional Neural Networks(CNNs)" Published under licence 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. [5] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutionalnetworks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

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

[7] Lu, Jiasen, et al. "Knowing when to look: Adaptive attention via a visual sentinel for im-age captioning." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).Vol. 6. 2017.

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

[9] 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.

[10] Feng, Yansong, and Mirella Lapata. "How many words is a picture worth? automatic caption generation for news images." Proceedings of the 48th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2010.

[11] Rashtchian, Cyrus, et al. "Collecting image annotations using Amazon's Mechanical Turk."

ISSN: 2057-5688

Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk. Association for Computational
Linguistics, 2010