International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VIII Aug 2021- Available at www.ijraset.com

Image Captioning using CNN and LSTM

Anish Banda[1], Harshavardhan Manne[2], Rohan Garakurthi[3]
[1, 2, 3]*Department of Computer Science and Engineering, Mahatma Gandhi Institute of Technology, Hyderabad, India*

Abstract: ***In the model we proposed, we examine the deep neural networks-based image caption generation technique. We give image as input to the model, the technique give output in three different forms i.e., sentence in three different languages describing the image, mp3 audio file and an image file is also generated. In this model, we use the techniques of both computer vision and natural language processing. We are aiming to develop a model using the techniques of Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) to build a model to generate a Caption. Target image is compared with the training images, we have a large dataset containing the training images, this is done by convolutional neural network. This model generates a decent description utilizing the trained data. To extract features from images we need encoder, we use CNN as encoder. To decode the description of image generated we use LSTM. To evaluate the accuracy of generated caption we use BLEU metric algorithm. It grades the quality of content generated. Performance is calculated by the standard calculation matrices.***

Keywords: CNN, RNN, LSTM, BLEU score, encoder, decoder, captions, image description.

I. INTRODUCTION

Image captioning is one of the most innovative model in the field of neural networks and artificial intelligence. Image captioning process generates a description of the input image. Image Captioning has different applications like suggestions, in altering applications, in virtual assistant, for faster images retrieval, in providing assistance to visually disabled persons etc.

In the past few years a new AI field named Deep Learning, emerged a lot gaining popularity in machine learning tasks due to its better performance in terms of speed and accuracy in comparison to its peer machine and neural network learning algorithms. The ability of generating a meaningful description from a input image is a difficult task but it also has a much effect in improving current algorithms, for instance in helping the visually disabled persons to have a better understanding of images.

First, we classify images separately We classify the images of CIFAR10 training dataset using various extractors. We initially trained our created model using the KNN methodology. Then after we tried applying some popular linear classifiers. The accuracy obtained with all these models is observed to be much less than expected since a high loss factor at the time of classification task will increase the loss even further at the time of generating captions. We then try to create and also train a simple CNN and achieve decent results within few hours of training. We used the BLEU assessment score to compare the accuracy of our created model with already existing ones. Thus, at the completion of this work we can conclude that the CNN are a good fit and is very useful in the process of encoding the images for the captioning model

II. RELATED WORK

In generating the caption for the image the central part is the scene understanding present in the image, which is significant in many of the applications (eg. Searching using pictures, dictating the stories from collections, helping visually disabled persons in understanding during browsing the internet and so forth). Over many decades, variety of image captioning models are developed.

The winners of ILSVRC have used some models, which contributed a ton to the field of image captioning. VGG16 is one of those architectural designs which was proposed by He et. al. in 2014 [1]. Analysts at Microsoft's AI utilized a CNN to create significant level highlights for every possible article in the picture. At that point they utilized Multiple Instance Learning (MIL)[6] to sort out what word is best suited to each area. On MSCOCO dataset, a total of 29% of BLEU metric was yielded. This is called pipeline approach. After this approach scientists from Google designed an end-to-end approach that is trainable. These scientists are motivated by the RNN architecture which is used for interpretation in machines.

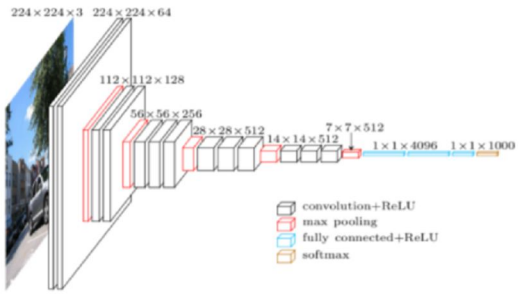
Vinyals et al. used the CNN since CNN could work as a better encoder compared to RNN[2]. They named this approach as the Neural Image Caption (NIC). Later to this, some scientists at the Stanford altered the neural image caption. They utilized a methodology that influences datasets of pictures and their description portrayals to find out the best correspondencies among linguistic and semantic information. In the past few years, progress has been made not only in image captioning models but also in various assessment metrics. The accuracy metric used by us was the BLEU score [3]. BLEU - which was a standard assessment metric adopted by many of the groups.

III. PROPOSED SYSTEM

Convolutional Neural Network (CNN) which acts as an encoder helps to encode the images into vectors. We use the VGG16 design proposed by K. Simonyan for certain alterations. We might utilize some part of the recent and advanced classification algorithms however that would increase the training time significantly. These image encodings are passed as an output to LSTM networks, a specialized version of RNN. The LSTM internal architecture utilized is in comparable design similar to the model utilized in machine translations. The input to this network is a picture that is encoded to a 224*224 size. Next we used the Flickr-8k dataset in the creation of the model. The model then generates a caption dependent on the word reference and structures present in words of captions in the preparation data. The produced caption is compared with the originally generated description through BLEU metric.

A. Convolutional Neural Networks

Convolutional Neural Networks (also called Convolutional nets or CNN) are a kind of Neural Networks which were demonstrated to be powerful in the area of picture acknowledgment and grouping. The complete design of a convolution net is demonstrated by using some important functions is as mentioned below[4]



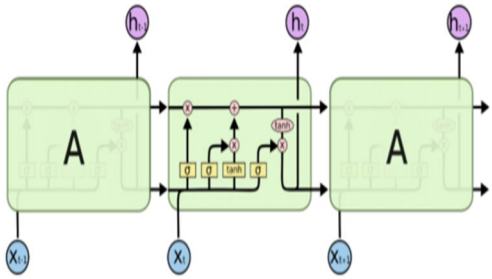
1) Sub Sampling or poolin
2) Classification (Completely Connected Layer)
3) Rectified Linear Unit (RELU)
4) Convolution

The picture taken for order should be in the dimension of 224*224 picture. This preprocessing is done by deducting the average of RGB esteems from every pixel decided from preparation dataset.

The layer of convolution comprises of 3*3 channels and the iteration size is made as 1. A ReLU (Rectified Linear Unit) actuation follows each layer of convolution. A rectified linear unit determines its capability using the equation f(x) = max(0,x). The output of CNN encoder would accordingly be a 1*1*4096 vector which is encoded using the training images, which is next passed to RNN for language generation. There were seen to be more effective. CNN[5] architectures like Resnet are found to be extravagant in computation because the number of layers in Resnet was 152 whereas we use VGG16 which is only a 16 layer organization.

B. Recurrent Neural Net Decoder Architecture

Recurrent neural networks are some kind of artificial nets where association among the units, structure a coordinated phase. The benefit of utilizing RNN over traditional type is that the RNN deals with self-assertive arrangement of information sources utilizing its memory. One of the issues with RNNs is that they don't consider long term dependencies into account. For Example, consider a machine that attempts to produce words all alone. For example, the sentence is "We experienced childhood in America, I talk familiar Spanish", if the system is attempting to anticipate the final say regarding the words for example English, the system has to realize that the language to trail by familiar of subject to the setting of the word America. It is conceivable that the difference among the pertinent data and places in which it is required turns out to be enormous in some cases the traditional recurrent neural networks fail. To eliminate the previously specified issue of "remembering memory for a much time", Hochreiter and Schmiber formulated the Long Short-Term Memory (LSTM) networks. From that point forward these networks, altered the areas of discourse acknowledgment, machine interpretation and so on Like the conventional RNNs, LSTMs likewise have a chain like design, however the mostly used libraries have a variety of structures in the event of long short term memory layer organization. A basic LSTM network is appeared beneath. We utilize this long short term network with a some modified variety.



Four Interacting layers in a LSTM layer

C. *Architectural Diagram of LSTM*

The following are the equations and their descriptions used in the architecture of LSTM.

D. *Equations*

The entire network is governed by the following equations

$$i_t = \sigma(W_{ix} x_{i+} W_{im} m_{t-1})$$

where i_t is the input gate at time t, W represents the trained parameters. The variable m_{t-1} denotes the output of the module at time t-1 and σ represents the sigmoid operation that includes numbers between 0 and 1, demonstrating how much of each output of every component should be passed on to the next component[8].

$$f_t = \sigma(W_{fx} x_{i+} W_{fm} m_{t-1})$$

where f_t represents the forget gate which indicates that the cell value should be forgotten or remembered.

$$O_t = \sigma(W_{ox} x_t + W_{om} m_{t-1})$$

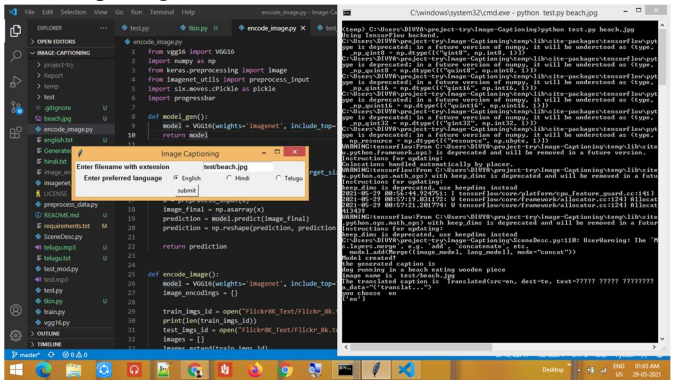
where o_t represents the gate of the output which determines whether to pass the new value of the cell or not.

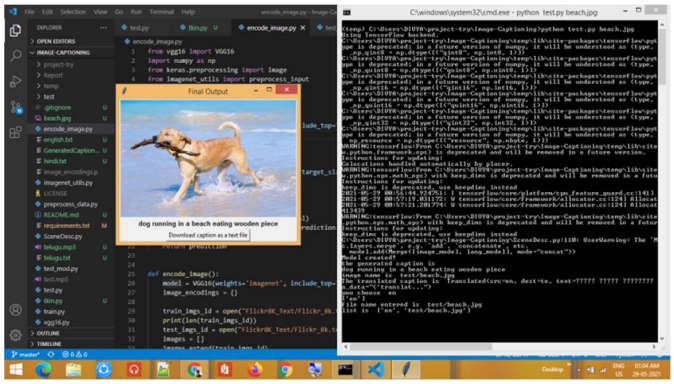
$$P_{t+1} = \text{Softmax}(m_t)$$

The output p_{t+1} of every gives the prediction of the words to be generated. The entire of this LSTM network is continuously repeated until an end sequence (.) is encountered. The series of these predicted words contribute to the entire description for a input image. The complete training process for the combined model (CNN encoder + RNN. language generator) and the LSTM[8] network in unravelled form is given in the image. The LSTM is designed such that is it predicts each words only after it sees the total image as well as the previously generated words as defined by $p(S_t \mid I, S_0, \ldots, S_{t-1})$

IV. RESULT ANALYSIS

Users can give input image as an JPG or PNG format. They will be prompted to choose a language so that the caption will be generated in their required language. The caption generated can be downloadable in their chosen language.





V. CONCLUSION

Automatically generating a caption for an image is a long way from develop and there are a ton of continuous examination projects focusing on more precise picture include extraction and semantically better sentence formation. We effectively finished what we referenced in the undertaking proposition, we utilized a more modest dataset (Flickr8k) because of restricted computational force. Our model is equipped to see a picture and producing a semantically correct output in user chosen languages, depending on the available words in its dictionary which is created based on various words in existing captions present in training dataset. This technique contains a encoder as convolutional neural network and decoder as long short term memory that aids in generation of descriptions. The motivation behind the created model is to expand the probability of the sentence generated according to the given picture. We tested the model using the training images present in Flickr8K and it showed good outcomes. We assess the precision of the created model based on a metric called Bilingual Evaluation Understudy. The precision can be made larger if a similar model is worked on a higher dataset i.e containing more images. Moreover, it would be attractive to recognize how unsupervised data be used from isolated context and isolated pictures, to enhance image explanation method

A. Future Scope

Captioning a image is very useful for the people who cannot visualize the picture. Model which we proposed can be very useful for them and this model can be coordinated with the applications like android or iOS to describe the picture in real time. Efficiency of our model can be upgraded for better results by the parameters hyper tuning. Efficiency can be upgraded by installing it on bigger database the words in dictionary can be incremented efficiently. By using the latest architecture, like ResNet and GoogleNet will also improve the efficiency in the separation task. This reduces the fault rate in generation of english sentence. Usage of Recurrent unit besides usage of bidirectional LSTM network can assist in enhancing the efficiency of model proposed.

REFERENCES

[1] Karen, simmyon and andrew askdnaf, "“Show and tell: A neural image caption generator.”," In: Very deep convolutional networks for large-scale image recognition., (2014).
[2] Vinyals, Oriol and et al, "Show and tell: A neural image caption generator.," Proceedings of the IEEE conference on computer vision and pattern",, 2015.
[3] Karpathy, Andrej and Li Fei-Fei, "Deep visual-semantic alignments for generating image descriptions.," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015.
[4] Johnson, Justin, Andrej Karpathy and Li Fei-Fei, "Densecap: Fully convolutional localization networks for dense captioning.," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.
[5] Wang, Cheng and et al, "Image captioning with deep bidirectional LSTMs.," Proceedings of the 2016 ACM on Multimedia Conference. ACM, 2016.
[6] Xu, Kelvin and et al, "Show, attend and tell: Neural image caption generation with visual attention," International Conference on Machine Learning, 2015.
[7] Papineni, Kishore and et al, "BLEU: a method for automatic evaluation of machine translation," Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002.
[8] Vedantam, Ramakrishna, C. Lawrence Zitnick and Devi Parikh, "Cider: Consensus-based image description evaluation," Proceedings of the IEEE conference on computer vision and pattern recognition, 2015.
[9] Fang , Hao and et al, "From captions to visual concepts and back.," Proceedings of the IEEE conference on computer vision and pattern recognition, 2015.