International Journal for Modern Trends in Science and Technology, 9(SI01): 99-102, 2023

ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTST09SI0119
Available online at: http://www.ijmtst.com/vol9si01.html

Constructing an Image Caption Generator with the use of CNN and LSTM

Dr.V.Suma Avani, P.Madhavi, J.Himabala, Y. Lakshmi Durga

Department of Computer Science and Engineering, Vijaya Institute of Technology for Women, Enikepadu, Vijayawada.

To Cite this Article
Dr.V.Suma Avani, P.Madhavi, J.Himabala, Y. Lakshmi Durga. Constructing an Image Caption Generator with the use of CNN and LSTM. International Journal for Modern Trends in Science and Technology 2023, 9(SI01), pp. 99-102. https://doi.org/10.46501/IJMTST09SI0119

Article Info
Received: 26 January 2023; Accepted: 22 February 2023; Published: 26 February 2023

ABSTRACT

Image captioning means, it is a process of creating a short description of an input image. It essentially involves writing a statement that describes the visual picture. Typically, the image may consist of many objects. Some objects are focused more than compared to others. Identifying such tasks is carried out manually. Hence, it needs a huge contribution of people and time to automate this process. The challenge is that the machine must deeply learn from the given datasets only in order to identify the objects, its actions, and their locations. The fact that people can do it readily for small sets but fail when there are more photos which makes it a challenging problem of deep learning. The image caption generation task can be shortened with the use of deep neural networks.

KEYWORDS: Captions, CNN,LSTM, RNN

1.INTRODUCTION

Image Caption Generator isaprocess that generates a caption about the given image in natural language like English. The traditional retrieval and template-based approaches to captioning began with the detection of the Subject, Verb, and Objectindependently and then joining them using a sentence template. However, the introduction of Deep Learning and significant advances in Natural Language Processing has had an equal impact on captioning.

[13] Image Caption Generator has two approaches. Bottom-up approaches means it, combines [1] [2] [3] the input from different objects which are identified in an original input image.Top-down approaches, means, it uses CNN [4] [5] [6] as encoder to extract the features from the image that are fed into decoders such as Long Short Term Memory (LSTM) and Recurrent Neural Networks (RNN).Our approach is based on top-down using CNN as encoder and LSTM as decoder. We use a deep Convolutional Neural Network (CNN) to extract important features of an image. Xception is used for image feature extraction. It is a CNN which has 71 layers deep.LSTM network uses this information and generate suitable captions. Figure 1 provides the model of CNN and LSTM.

CNN scans the image from top to bottom and left to right and extract some important features and combine the features. It is also responsible for image classification..It has three layers [7] they are Convolutional Layer, Pooling Layer and Fully Connected

Layer. Convolution Layer uses filter and strides to obtain the Feature Maps. These Feature Maps are the matrix that is obtained after the Convolution Layer. It can be simplified using ReLU (Rectified Linear Unit) that maps negative values to 0. The resulted Feature Map is reduced by sending it into the Pooling Layer where it is reduced to the smaller sized matrix. This is how the features are extracted. At the end of the CNN is the Fully Connected Layer where the actual Classification occurs.

[14] LSTM is an advanced RNN which is suitable for sequence prediction problems. It is also used in speech recognization. In RNN the output of previous step is fed into ongoing step. It is not suitable for larger sentences. The main advantage of LSTM over RNN is the LSTM keep the information in the memory for longer period of time.An LSTM recurrent unit remembers all the past information as far as network sees and it also forget the irrelevant information. This is done by introducing different layers called "gates" for different purposes.AnLSTM Network consists of three different gates for different purposes.They are input gate,output gate and forget gate. Input gate takes the information from the user and supplies to other gates. Output gate determines what output can be generated from current Internal State. Forget gate decides what information can be discarded based on previous data.

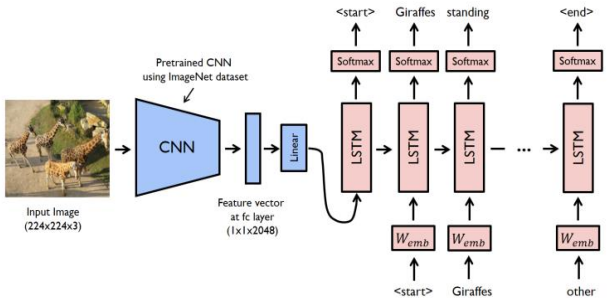


Figure1: CNN-LSTM architecture. Xception is responsible for image feature extraction.LSTM will decode the information from the CNN and generate suitable captions.

2. REALATED WORK

Serval methods have been proposed for image caption generator using deep learning concepts. [1] First, a method which is based on statistical probability to generate features and second, the neural network model based on encoder and decoder. Image Caption generator approaches [8] are divided into two i.e; top down and bottom up approaches. Top down approach means it uses encoder and decoder. CNN is used [6] as [9] encoder which extract the features from the image which is fed into the decoder such as LSTM and RNN which generates word one by one based on the features and weighted words that were end-to-end trainable.The main advantageof LSTM over RNN is LSTM stores information in the memory for longer period of time.

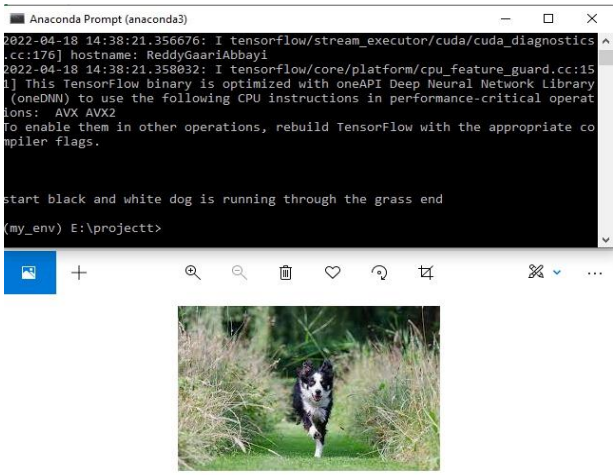
In Bottom up approach [11] the CNN and bi-directional RNN is trained to map captions to images and next they combine input from different object parts identified in original image.The RNN training is difficult [12] they had a general vanishing gradient problem. RNN stores data for shorter period.So, therefore it is favourable to store short data. To solve a particular problem it uses many steps which results in losing data when we backpropagate.With so many steps it has to store more and more data which leads to losing the information in this way vanishing gradient problem occurs. In this the images are not end to end trainable.

3. PROPOSED WORK

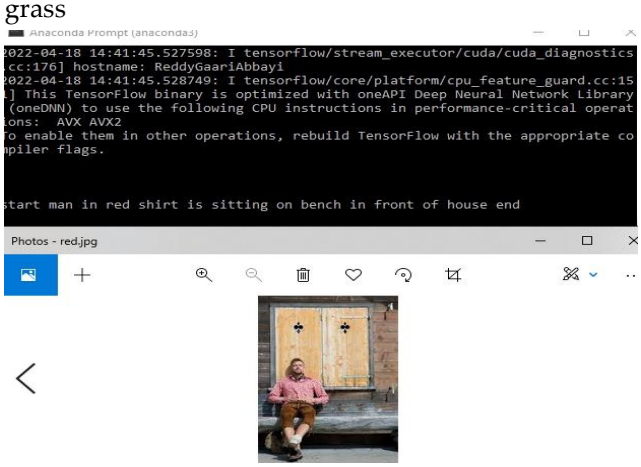
In this paper, the overview of image caption generator is first, the pre-trained model Xception is employed as encoder to extract the important features and the weight attached to each word in training captions is calculated. Second, the LSTM is trained using CNN features and weighted words of training images which adopted as decoder, which takes CNN features of the target image as input and generates description word one by one.
Target image->CNN->LSTM->CNN<-Training set

4. RESULTS

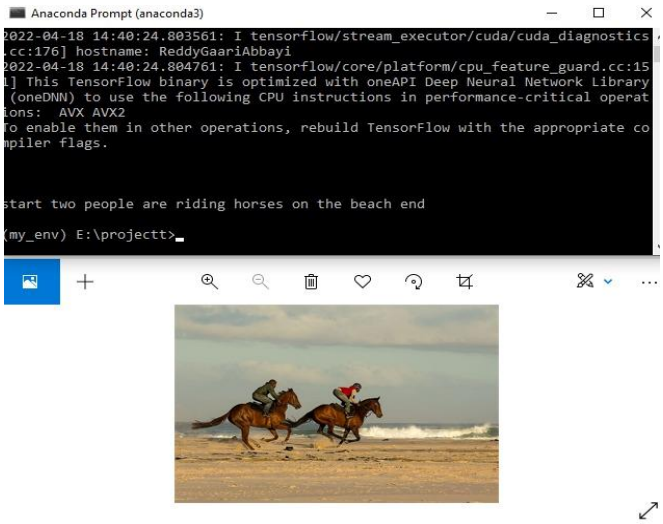
The project is executed in anaconda prompt using flickr8k dataset which consists of 8091 images and every image has 5 descriptions.It produce accurate results.



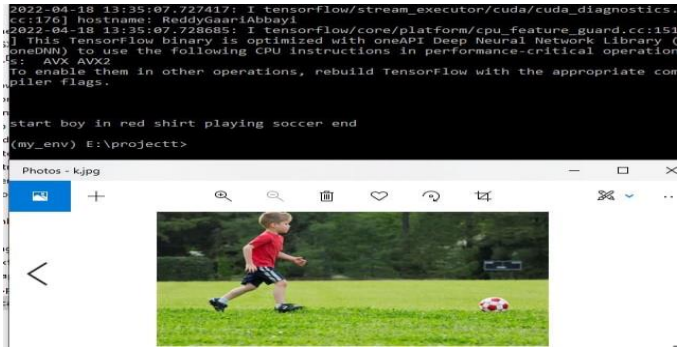
Caption: black and white dog is running through the grass



Caption: man in red shirt is sitting on bench in front of house



Caption: two people are riding horses on the beach



Caption:boy in red shirt playing soccer

5. CONCLUSION

In this paper,the main objective is to generate real-time captions for each input picture in a single pass while assuring that the result is accurate. In this project, the reference based CNN-LSTM model is making use of training images to generate a quality of captions. The words are weighted according to the relevance of image which leads the model to focus on key information of the captions.

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

REFERENCES

[1] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In Proceedings of the 11th European Conference on Computer Vision: Part IV, ECCV'10, pages 15–29, Berlin, Heidelberg, 2010. Springer-Verlag

[2] Polina Kuznetsova, Vicente Ordonez, Alexander C. Berg, Tamara L. Berg, and Yejin Choi. Collective generation of natural image descriptions. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1, ACL '12, pages 359–368, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[3] Siming Li, Girish Kulkarni, Tamara L. Berg, Alexander C. Berg, and Yejin Choi. Composing simple image descriptions using web-scale n-grams. In Proceedings of the Fifteenth Conference on Computational Natural Language Learning, CoNLL '11, pages 220–228, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[4] Xinlei Chen and C. Lawrence Zitnick. Learning a recurrent visual representation for image caption generation. CoRR, abs/1411.5654, 2014.

[5] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L. Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). CoRR, abs/1412.6632, 2014.

[6] Oriol Vinyals, Alexander Toshev, SamyBengio, and Dumitru Erhan. Show and tell: A neural image caption generator. CoRR, abs/1411.4555, 2014.

[7] Aneja, Jyoti, Aditya Deshpande, and Alexander G. Schwing. "Convolutional image captioning." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5561-5570. 2018.

[8] Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. CoRR, abs/1411.2539, 2014.

[9] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognitionand description. CoRR, abs/1411.4389, 2014.

[10] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." *Neural computation* 9, no. 8 (1997): 1735-1780.

[11] Andrej Karpathy and Fei-Fei Li. Deep visual-semantic alignments for generating image descriptions. CoRR, abs/1412.2306, 2014

[12] Hao Fang, Saurabh Gupta, Forrest N. Iandola, Rupesh Kumar Srivastava, Li Deng, Piotr Dollar, ´ Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. From captions to visual concepts and back. CoRR, abs/1411.4952, 2014.

[13] Parvathi, D. S. L., Leelavathi, N., Ravikumar, J. M. S. V., & Sujatha, B. (2020, July). Emotion Analysis Using Deep Learning. In 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC) (pp. 593-598). IEEE.

[14] Kumar, J. R., Sujatha, B., &Leelavathi, N. (2021, February). Automatic Vehicle Number Plate Recognition System Using Machine Learning. In IOP Conference Series: Materials Science and Engineering (Vol. 1074, No. 1, p. 012012). IOP Publishing."