International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 04 | Apr 2021 www.irjet.net p-ISSN: 2395-0072

Image Caption Generator

Vivek Kamble[1], Sayam Koul[2], Abhishek Chaudhari[3] ,Rajashri Sonawale[4]

[1,2,3] *Student, Computer Engineering, Mahatma Gandhi Mission's College of Engineering and Technology*

[4]*Assistant Professor, Dept. of Computer Engineering, Mahatma Gandhi Mission's College of Engineering and Technology, Maharashtra, India*

Abstract - *Image captioning is a fundamental task which requires semantic understanding of images and the ability of generating description sentences with proper and correct structure. In this study, the authors propose a hybrid system employing the use of multilayer Convolutional Neural Network (CNN) to generate vocabulary describing the images and a Long Short Term Memory (LSTM) to accurately structure meaningful sentences using the generated keywords. The convolutional neural network compares the target image to a large dataset of training images, then generates an accurate description using the trained captions. We showcase the efficiency of our proposed model using the Flickr8K and Flickr30K datasets and show that their model gives superior results compared with the state-of-the-art models utilising the Bleu metric. The Bleu metric is an algorithm for evaluating the performance of a machine translation system by grading the quality of text translated from one natural language to another. The performance of the proposed model is evaluated using standard evaluation matrices, which outperform previous benchmark models.*

1.INTRODUCTION

Caption Generation involves the dual techniques from computer vision to understand the content of the image and a language model from the field of natural language processing to turn the understanding of the image into words in the right order. Image captioning has various applications such as recommendations in editing applications, usage in virtual assistants, for image indexing, for visually impaired persons, for social media, and several other natural language processing applications. Recently, deep learning methods have achieved state-of the-art results on examples of this problem. It has been demonstrated that deep learning models are able to achieve optimum results in the field of caption generation problems. Instead of requiring complex data preparation or a pipeline of specifically designed models, a single end-to-end model can be defined to predict a caption, given a photo. In order to evaluate our model, we measure its performance on the Flickr8K dataset using the BLEU standard metric. These results show that our proposed model performs better than standard models regarding image captioning in performance evaluation

2. DATASET AND EVALUATION METRICS

For task of image captioning there are several annotated images dataset are available. Most common of them are Pascal VOC dataset, Flickr 8K and MSCOCO Dataset. Flickr 8K Image captioning dataset [9] is used in the proposed model. Flickr 8K is a dataset consisting of 8,092 images from the Flickr.com website. This dataset contains collection of day-to-day activity with their related captions. First each object in image is labeled and after that description is added based on objects in an image. We split 8,000 images from this corpus into three disjoint sets. The training data (DTrain) has 6000 images whereas the development and test dataset consist of 1000 images each. In order to evaluate the image-caption pairs, we need to evaluate their ability to associate previously unseen images and captions with each other. The evaluation of model that generates natural language sentence can be done by the BLEU (Bilingual Evaluation Understudy) Score. It describes how natural sentence is compared to human generated sentence. It is widely used to evaluate performance of Machine translation.

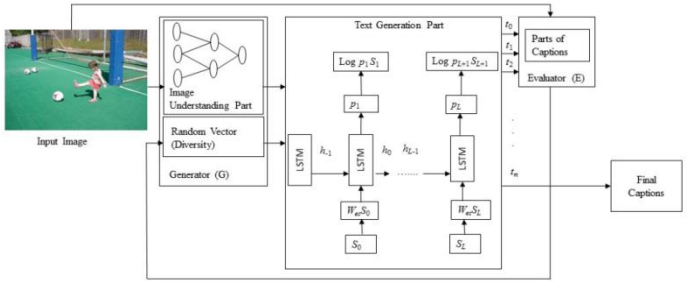
3. TRAINING PHASE

During training phase we provide pair of input image and its appropriate captions to the image captioning model. The VGG model is trained to identify all possible objects in an image. While LSTM part of model is trained to predict every word in the sentence after it has seen image as well as all previous words. While LSTM part of model is trained to predict every word in the sentence after it has seen image as well as all previous words. For each caption we add two additional symbols to denote the starting and ending of the sequence.

4. IMPLEMENTATION

To generate captions, first, Create a caption generator. This caption generator utilizes beam search to improve the quality of sentences generated. At each iteration, the generator passes the previous state of the LSTM (initial state is the image embedding) and previous sequence to generate the next softmax vector. At each iteration, the generator passes the previous state of the LSTM (initial state is the image embedding) and previous sequence to generate the next softmax vector. Next, you'll load the show and tell model and use it with the above caption generator to create candidate sentences. These sentences

will be printed along with their log probability. Once graph is defined it can be executed on any supported devices. The photo features are pre-computed using the pretrained model and saved. These features are then loaded and them into our model as the interpretation of a given photo in the dataset to reduce the redundancy of running each photo through the network every time we want to test a new language model configuration. The preloading of the image features is also done for real. Keras 2.0 was used to implement the deep learning model because of the presence of the VGG net which was used for the object identification. Tensorflow library is installed as a backend for the Keras framework for creating and training deep neural networks. TensorFlow is a deep learning library developed by Google. It provides heterogeneous platform for execution of algorithms i.e. it can be run on low power devices like mobile as well as large scale distributed system containing thousands of GPUs. The neural network was trained on the Nvidia Geforce 1050 graphics processing unit which has 640 Cuda cores. In order to define structure of our network TensorFlow uses graph definition. Once graph is defined it can be executed on any supported devices. The photo features are pre-computed using the pretrained model and saved. These features are then loaded and them into our model as the interpretation of a given photo in the dataset to reduce the redundancy of running each photo through the network every time we want to test a new language model configuration. The preloading of the image features is also done for real time implementation of the image captioning model. The architecture of the model is as shown in figure below.



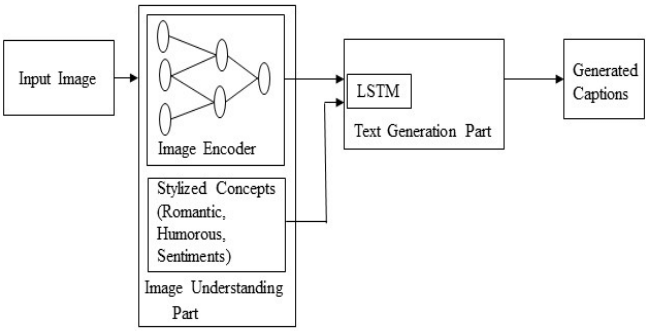
5. GENERAL OVERVIEW

Generally, a captioning model is a combination of two separate architecture that is CNN (Convolution Neural Networks)& RNN (Recurrent Neural Networks) and in this case LSTM (Long Short Term Memory), which is a special kind of RNN that includes a memory cell, in order to

maintain the information for a longer period of time .Basically, CNN is used to generate feature vectors from the spatial data in the images and the vectors are fed through the fully connected linear layer into the RNN architecture in order to generate the sequential data or sequence of words that in the end generate description of an image by applying various image processing techniques to find the patterns in an image.

6. SOLUTION APPROACH

- User friendly system. No need to give extra Instructions for the use of System.
- Main Aim Is To Automatically Describe an Image With One Or More Natural Languages Sentences
- System is Open source. So anyone can Use it & change it according to User Demand.
- It generates Caption about related input images. There is no extra Noise will be generated by User as well as System



11. CONCLUSIONS

In this paper, the authors have implemented a deep learning approach for the captioning of images. The sequential API of Keras was used with Tensorflow as a backend to implement the deep learning architecture to achieve a effective BLEU score of 0.683 for our model. The Bilingual Evaluation Understudy Score, or BLEU for short, is a metric for evaluating a generated sentence to a reference sentence. A perfect match results in a score of 1.0, whereas a perfect mismatch results in a score of 0.0. In the future, the authors are working on alternating PreTrained Photo Models to improve the feature extraction of the modelAlso, the authors are planning to improve achieve better performance by using word vectors on a much larger corpus of data such as news articles and other online sources of data.The configuration of the model was tuned, but other alternate configurations can be trained to see for improvement in the performance of the image captioning model.

12. REFERENCES

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, [Online] Available: https://papers.nips.cc/paper/4824-imagenetclassificationwith-deep-convolutionalneural-networks.pdf

[2] Jyoti Aneja, Aditya Deshpande, Alexander Schwing, Convolutional Image Captioning, [Online] Available: https://arxiv.org/pdf/1711.09151.pdf

[3] Jia-Yu Pan, Hyung-Jeong Yang, Pinar Duygulu, Automatic Image Captioning, Conference: Conference: Multimedia and Expo, 2004. ICME '04. 2004 IEEE International Conference on, Volume: 3

[4] Andrej Karpathy, Li Fei-Fei, Deep VisualSemantic Alignments for Generating Image Descriptions, [Online] Available: https://cs.stanford.edu/people/karpathy/cvpr2015.pdf

[5] BLEU: a Method for Automatic Evaluation of Machine Translation Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu IBM T. J. Watson Research Center Yorktown Heights, NY 10598, USA