International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 5 Issue 4, May-June 2021 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Image Captioning Generator using Deep Machine Learning

Sreejith S P[1], Vijayakumar A[2]

[1]Student, [2]Professor,
[1,2]Department of MCA, Jain Deemed-to-be University, Bengaluru, Karnataka, India

ABSTRACT

Technology's scope has evolved into one of the most powerful tools for human development in a variety of fields.AI and machine learning have become one of the most powerful tools for completing tasks quickly and accurately without the need for human intervention. This project demonstrates how deep machine learning can be used to create a caption or a sentence for a given picture. This can be used for visually impaired persons, as well as automobiles for self-identification, and for various applications to verify quickly and easily. The Convolutional Neural Network (CNN) is used to describe the alphabet, and the Long Short-Term Memory (LSTM) is used to organize the right meaningful sentences in this model. The flicker 8k and flicker 30k datasets were used to train this.

KEYWORDS: *Image captioning, Convolutional Neural Network, Long Short-Term Memory, FLICKER_8K dataset*

How to cite this paper: Sreejith S P | Vijayakumar A "Image Captioning Generator using Deep Machine Learning" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-5 | Issue-4, June 2021, pp.832-834, URL: www.ijtsrd.com/papers/ijtsrd42344.pdf

IJTSRD42344

I. INTRODUCTION

The method of creating a textual interpretation for a collection of images is known as image captioning. In the Deep Learning domain, it has been a critical and fundamental mission. Captioning images has a wide range of applications. NVIDIA is developing an app to assist people with poor or no vision using image captioning technology. Caption generation is a fascinating artificial intelligence problem that involves generating a descriptive sentence for a given image. It uses two computer vision techniques to understand the image's content, as well as a language model from the field of natural language processing to convert the image's understanding into words in the correct order. Image captioning can be used in a variety of ways. Image captioning has a variety of uses, including editing software recommendations, virtual assistants, image indexing, accessibility for visually disabled people, social media, and a variety of other natural language processing applications. On examples of this dilemma, deep learning methods have recently achieved state-of-the-art results. Deep learning models have been shown to be capable of achieving optimal results in the field of caption generation problems. Rather than requiring complex data preparation or a pipeline of custom-designed models, a single end-to-end model can be specified to predict a caption provided a photograph. To assess our model, we use the BLEU standard metric to assess its efficiency on the Flickr8K dataset. These findings demonstrate that our proposed model outperforms traditional models in terms of image captioning in performance evaluation. Image captioning can be thought of as an end-to-end Sequence to Sequence problem since it transforms images from a series of pixels to a series of words. Both the language or comments as well as the images must be processed for this reason. To obtain the feature vectors, we use recurrent Neural Networks for the Language part and Convolutional Neural Networks for the Image part respectively.

II. RELATED WORKS

[1]In this paper introduces a new image captioning problem: describing an image under a given topic. To solve this problem, a cross-modal embedding of image, caption, and topic is learned. The proposed method has achieved competitive results with the state-of-the-art methods on both the caption-image retrieval task and the caption generation task on the MS-COCO and Flickr30K datasets. This new framework provides users with controllability in generating intended captions for images, which may inspire exciting applications.

[2] In this paper, we proposed a novel image captioning model, called domain-specific image captioning generator, which generates a caption for given image using visual and semantic attention (referred as general caption in this paper), and produces a domain-specific caption with semantic on topology by replacing the specific words in the general caption with domain-specific words. In the experiments we evaluated our image caption generator qualitatively and quantitatively. The limitation of our model is our model is not end-to-end manner for semantic ontology. Therefore, as a future work, we plan to build a new image caption generator with semantic ontology embedding in end-to-end manner.

[3] Algorithm on the basis of the analysis of the line projection and the column projection, finally, realized the news video caption recognition using a similarity measure formula. Experimental results show that the algorithm can effectively segment the video caption and realized the caption efficiency, which can satisfy the need of video analysis and retrieval.

[4] In this work a Recurrent Neural Network (RNN) with Long Short-Term Memory (LSTM) cell and a Read-only Unit has been developed. An additional unit has been added to work that increases the model accuracy. Two models, one with the LSTM and the other with the LSTM and Read-only Unit have been trained on the same MSCOCO image train dataset. The best (average of minimums) loss values are 2.15 for LSTM and 1.85 for LSTM with Read-only Unit. MSCOCO image test dataset has been used for testing. Loss values for LSTM and LSTM with Read-only Unit model test are 2.05 and 1.90, accordingly. These metrics have shown that the new RNN model can generate the image caption more accurately.

III. PROPOSED SYSTEM

The proposed model used the two modalities to construct a multimodal embedding space to find alignments between sentence segments and their corresponding represented areas in the picture. The model used RCNNs (Regional Convolutional Neural Networks) to detect the object area, and CNN trained on Flickr8K to recognize the objects.

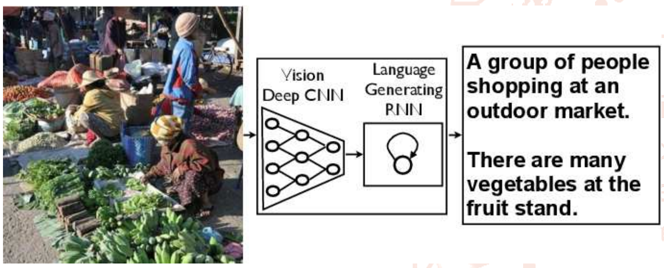

Fig1. Illustration of classical image captioning and topic-oriented image captioning processes

CNNs are designed mainly with the assumption that the input will be images. Three different layers make up CNNs. Convolutional, pooling, and fully-connected layers are the three types of layers. A CNN architecture is created when these layers are stacked together in Figure 2.

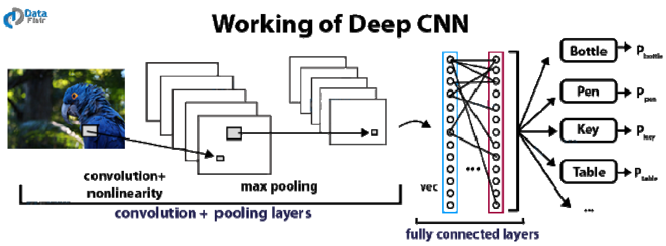

Fig 2 Deep CNN architecture

Multiple layers of artificial neurons make up convolutional neural networks. Artificial neurons are mathematical functions that measure the weighted number of multiple inputs and output an activation value, similar to their biological counterparts' Basic characteristics such as horizontal, vertical, and diagonal edges are normally detected by the CNN. The first layer's output is fed into the next layer, which removes more complex features including corners and edge combinations. Layers identifies the CNN it goes to another level like objects etc.

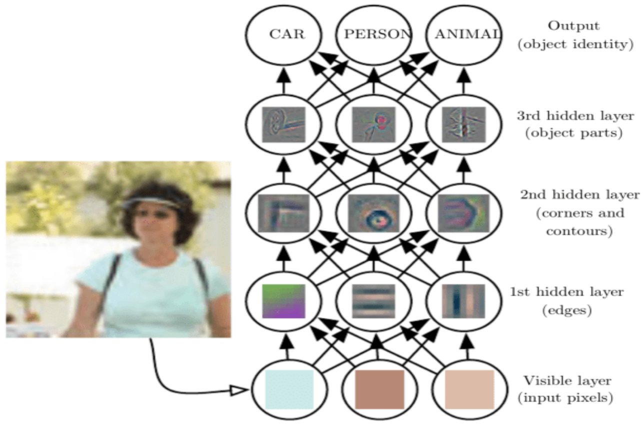

Fig 3 extracting hidden and visible layer of an image

Convolution is the process of doubling the pixels by its weight and adding it together .it is actually the 'C' in CNN.CNN is typically made up of many convolution layers, but it may also include other elements the layer of classification is the last layer of convolutional neural network and it will takes the data for output in convolution and gives input.

The CNN begins with a collection of random weights. During preparation, the developers include a large dataset of images annotated with their corresponding classes to the neural network (cat, dog, horse, etc.). Each image is processed with random values, and the output is compared to the image's correct mark. There is any network connection problem it will not fit label, it can occur in training. this can make small difference in weight of neuron. when the image is taken again output will near to correct output.

Conventional LSTM

Main Idea: A storage neuron (changeably block) with a memory recall (aka the cell conditional distribution) and locking modules that govern data flow into and out of storage will sustain its structure across period. The previous output or hidden states are fed into recurrent neural networks. The background information over what occurred at period T is maintained in the integrated data at period t.

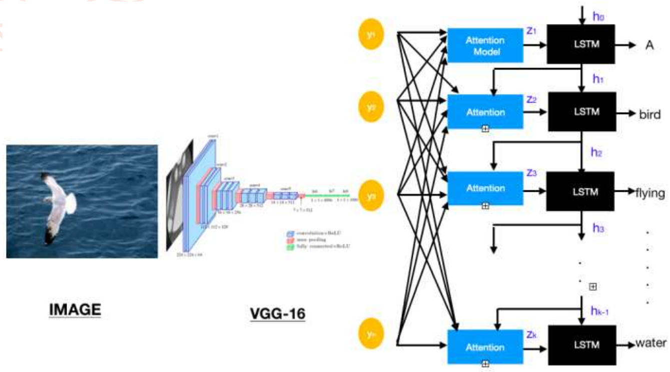

Fig 4 deep working of vgg-16 and LSTM.

RNNs are advantageous as their input variables (situation) can store data from previous values for an unspecified period of time. During the training process, we provide the image captioning model with a pair of input images and their corresponding captions. The VGG model has been trained torecognize all possible objects in an image. While LSTM portion of framework is developed to predict each piece of text once it has noticed picture as well as all previous words. We add two additional symbols to each caption to indicate the beginning and end of the series. When a stop word is detected, the sentence generator stops and the end of the string is marked. The model's loss function is calculated as, where I is the input image and S is the generated caption. The length of the produced sentence is N. At time t, pt and St stand for probability and expected expression, respectively. We attempted to mitigate this loss mechanism during the training phase.

IV. IMPLEMENTATION

The project is implemented using the Python Jupyter environment. The inclusion of the VGG net, and was used for image processing, Keras 2.0 was used to apply the deep convolutional neural network. For building and training deep neural networks, the Tensor flow library is built as a backend for the Keras system. We need to test a new language model configuration every time. The preloading of image features is also performed in order to apply the image captioning model in real time.

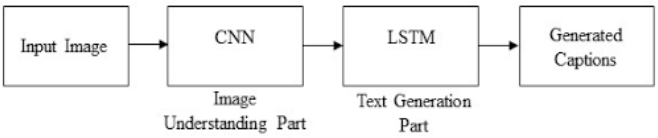

Fig 5 overall view of the implementation flow

V. RESULTS AND COMPARISION

The image caption generator is used, and the model's output is compared to the actual human sentence. This comparison is made using the model's results as well as an analysis of human-collected captions for the image. Figure 6 is used to measure the precision by using a human caption.

fig 6 test case figure.

In this image the test result by the model is 'the dog is seating in the seashore'. And the human given caption is' the dog is seating and watching the sea waves. So that by this comparison the caption that is generated by the model and the human almost the same caption and the accuracy is about 75% percent through the study. which shows that our generated sentence was very similar to the human given sentences.

VI. CONCLUSION

In this paper explains the pre trained deep machine learning to generate image to caption. And this can be implemented in different areas like theassistance for the visually impaired peoples. And also, for caption generation for the different applications. In future this model can used with larger data sets and the with the selected datasets so that the model can predict the accurate caption with the same as the human. And the alternate method training for accurate results in the feature in the image caption model.

REFERENCES

[1] Soheyla Amirian, Khaled Rasheed, Thiab R. Taha, Hamid R. Arabnia§," *Image Captioning with Generative Adversarial Network*", 2019 (CSCI).

[2] Miao Ma, Xi'an," *A Grey Relational Analysis based Evaluation Metric for Image Captioning and Video Captioning*". 2017 (IEEE).

[3] Niange Yu, *Xiaolin Hu, Binheng Song, Jian Yang, and Jianwei Zhang." Topic-Oriented Image Captioning Based on Order-Embedding*", (IEEE), 2019.

[4] Seung-Ho Han and Ho-Jin Choi" *Domain-Specific Image Caption Generator with Semantic Ontology*". 2020 (IEEE).

[5] Aghasi Poghosyan," *Long Short-Term Memory with Read-only Unit in Neural Image Caption Generator*". 2017 (IEEE).

[6] Wei Huang, Qi Wang, and Xuelong Li," *Denoising-Based Multiscale Feature Fusion for Remote Sensing Image Captioning*", 2020 (IEEE).

[7] Shiqi Wu Xiangrong Zhang, Xin Wang, Licheng Jiao" *Scene Attention Mechanism for Remote Sensing Image Caption Generation*". 2020 IEEE.

[8] He-Yen Hsieh, Jenq-Shiou Leu, Sheng-An Huang," *Implementing a Real-Time Image Captioning Service for Scene Identification Using Embedded System*", 2019 IEEE.

[9] Yang Xian, Member, IEEE, and Yingli Tian, Fellow," *Self-Guiding Multimodal LSTM - when we do not have a perfect training dataset for image captioning*", 2019 IEEE.

[10] Qiang Yang." *An improved algorithm of news video caption detection and recognition*". International Conference on Computer Science and Network Technology.

[11] Lakshminarasimhan Srinivasan1, Dinesh Sreekanthan2, Amutha A.L3 1,2." *Image Captioning - A Deep Learning Approach*". (2018)

[12] Yongzhuang Wang, Yangmei Shen, HongkaiXiong, Weiyao Lin." *ADAPTIVE HARD EXAMPLE MINING FOR IMAGE CAPTIONING*". 2019 IEEE.