IJFANS INTERNATIONAL JOURNAL OF FOOD AND NUTRITIONAL SCIENCES

ISSN PRINT 2319 1775 Online 2320 7876

 Volume 11, Sp.Iss 6 , 2022

Advancements in Image Captioning: Bridging Computer Vision and Natural Language Processing with Deep Learning

S.Sagar Imambi[1], N.V. Nikhila[2]

[1]Professor, Koneru Lakshmaiah Education Foundation (KLEF), Vaddeswaram, Green fields, Guntur, Andhra Pradesh, India -522302
[2] Graduate ,Koneru Lakshmaiah Education Foundation (KLEF), Vaddeswaram, Green fields, Guntur, Andhra Pradesh, India -522302

DOI : 10.48047/IJFANS/11/S6/036

ABSTRACT

Over the past few years, image captioning has emerged as a complex and demanding task within the field of artificial intelligence. It has attracted many researchers in the field of AI and became an arduous and an interesting task. Image captioning, automatically generates the textual description according to the content observed in an image and it is the combination of two methods including computer vision and natural language processing. Computer vision is to realize the content of the image and natural speech processing is to understand the image into words in the correct order. Recently, Deep learning methods are achieving better results on the problem of caption generation and they can define a single end-to-end model to predict a caption when a photograph is given, instead of requiring a pipeline of specifically designed models or sophisticated data preparation. By using deep learning techniques like CNN, RNN accurate descriptions can be predicted. Convolutional Deep Neural Network (CNN) is used for feature extraction from image and Recurrent Neural Network is used for sentence generation. the model is trained in such a way that if an image is given to the model it generates the textual description observed in an image. Recurrent neural network can be trained on a dataset of images and text descriptions, and then used to generate new text descriptions for new images.

Key words:CNN, Bi-LSTM, BLEU, Captioning

1. INTRODUCTION

Many researchers making significant contributions.by doing their research on the challenging problem to develop a model which can automatically generate the textual description by understanding the content of an image by using well-formed English sentences. [5]Creating image captions is a demanding and difficult undertakingand had great impact on many applications like helping the visually handicapped people in understanding the content of the images. It is used in medical applications like Skin vision in confirming whether a skin condition is skin cancer or not and also applied in many areas including the military, education, web searching, commerce, social media platforms etc., [2] Every day we will see lots of images from various sources like news articles, internet, advertisements and document diagrams which contains the images where the viewers have to demonstrate the images themselves because most of the images do not have the description. Even though human can understand the images without any detailed captions but it is difficult for a machine to demonstrate or to interpret the captions for an image. Image captioning is a model build to interpret the detailed captions from an image as fast and as accurate as human by the machine [4]. Early image caption generation combine the information by some static object class libraries using statistical

ISSN PRINT 2319 1775 Online 2320 7876
Research paper **Volume 11, Sp.Iss 6 , 2022**

language models, Gaizauskas and Aker used dependency models for automatically tagging the images, Li et al proposed a n-gram model, yang et al proposed a language model by using the parameters of hidden markov model and may indirect methods are also proposed earlier for image captioning[6]. All the methods described or proposed by the researchers have their own characteristics and they are brainstorming but they all share a common disadvantage that they can't make an instinctive feature observations on actions or objects in an image or they didn't gave an end to end general model to solve this problem. The initiation of deep learning methods have made many breakthroughs had new hopes in creating captions for the images.

2. LITERATURE SURVEY

2.1 Image Captioning.

Figure 1:Image

If anyone asks what you see in the above picture some may assert that there is a dogin a grassy area, some other may verbalizethat a dog with brown spots in grassy area. While others may provide different descriptions, generating image captions remains a strenuous and formidable task. All the captions are related to this image only. But the point is it is easy for us as humans we can just see the image and we can describe the image in an appropriate language but writing a computer program that takes the image as the input and producing a relevant caption as the output is about image captioning [16].

Image captioning involves mainly two methods Computer vision models and NLP where recognising and describing the images as well as videos is the fundamental challenge of computer vision and this can be done by using Deep learning models like supervised convolutional neural network (CNN). [4] Natural Language Processing is employed to comprehend an image and convert it into words in the accurate sequence.

2.2 Convolutional Neural Networks :

ISSN PRINT 2319 1775 Online 2320 7876
Research paper **© 2012 IJFANS. All Rights Reserved,** Volume 11, Sp.Iss 6 , 2022

Neural network's main idea is taken from the Cognitive science [13] where many simple processing units are connected for intelligent behaviors. But due to its disadvantage of not having large computational power CNN is introduced. Image analysis is most common use of CNN. Convolutional neural Network consists of hidden layers known as Convolutional layers which make CNN more effective for image analysis.[4]

CNN layer types mainly include three types [9] as shown in the fig 2:

- Convolutional layer
- Pooling layer
- Fully connected layer

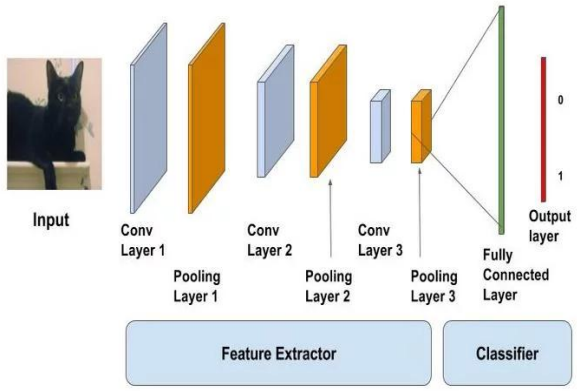


Figure 2:Convolutional Neural Network Archotecture

Input Layer:

When a computer sees image, it converts the image into an array of pixel values depending on the image resolution and size. Let's consider an image of type of jpg and size be 480 x 480. Then its converted to 480 x 480 x 3 image where the represents the RBG values. To describe the intensity of the pixel [15], they are given numbering from 0 to 255. Further the array with numbers are given as input to the image classification.

Convolutional layer:

Convolutional Layer is most important part of image classification. The main task in this layer is extracting attributes from the input image. Conv layer consists of many feature maps. The neuron of same feature map is applied in extracting regional characteristics of various positions in the former surface[14]. But for single neuron, its extraction is regional feature of the same positions in the former separate feature map [14]. The results in the Conv layers are passed to nonlinear Activation function like sigmoid, tanh, ReLu. Fig 2 shows how high-level image characteristics details are obtained from image using a kernel.

ISSN PRINT 2319 1775 Online 2320 7876
Research paper **Volume 11, Sp.Iss 6 , 2022**

Pooling layer:

A problem with the output from the Convolutional layer is that they are sensitive to the location of the features in the input . One idea to reduce the sensitivity is that we can decrease its dimensionality i.e., down sampling. The pooling layer is employed to reduce the size of the feature map.There are two types of common pooling techniques that may be relayed to decrease the dimensionality.They are max pooling and the average pooling. In max pooling, calculating the max value of each patch in the character map. Whereas average pooling, finding the average of each patch in the feature map. Fig 3 is an example of max pooling.

Fully Connected layer:

The role of the fully connected layer is to establish connections between the output and the former layer. There is no spatial arrangement in this layer. There can be many fully connected layers where the last layer is connected to the output layer. Soft regression is considered one of the most effective approaches due to its strong performance. Other methods like SVM can also be used with CNN to solve more complex task.

2.3 Recurrent Neural Networks (RNN)

RNNs are the powerful network architectures for processing the data and have been widely used in speech recognition, hand written recognition andnatural language processing in modern years. Here in these networks they allow the cyclical connection and the weights can be reused across various instances of neurons and each of them will have different timestamps so that the network can learn the history of the previous states and map them to the current state. But these traditional RNN can not learn the long term dependencies present in between the inputs and outputs.[7].

2.4 Long term recurrent Convolutional Network(LRCN).

LCRN is the combination of the deep hierarchical visual feature extractor like CNN with a model which will synthesize and recognize temporal dynamics for the tasks involving data (i/p or o/p), linguistic, visual or otherwise. LRCNs are a class of architectures supporting the strength of rapid progress in Convolutional networks for the visual recognition problems and also the growing desire to apply such models to inputs(time varying) and outputs. It processes the possibly variable length visual input i.e. left and a CNN i.e. middle left and the outputs are fed into a stack of Recurrent Sequence models i.e. middle right and finally produce a variable length prediction i.e. right.[9]

ISSN PRINT 2319 1775 Online 2320 7876
Research paper **Volume 11, Sp.Iss 6 , 2022**

2.5 Multimodal Recurrent Neural Network(M-RNN)

The architecture of multimodal Recurrent Neural Network(M-RNN) is shown below. It consists of five layers in every frame, the recurrent layer, the multimodal layer, two word embedding layers and the softmax layer. The embedding layers will embed the one hot input to a word dense representation and it encodes both syntactic and semantic meaning of the words. Semantically relevant words can be found by calculating the Euclidean distance between the dense word vectors in the embedding layers. After the embedding layers there will be a recurrent layer with 256 dimensions and the calculation of this layer is different from the calculation done for the traditional RNN. After the recurrent layer there will be a 512 dimensional multimodal layer that concatenates the vision part of the M-RNN model and language model part. It has three inputs recurrent layer, embedding layer and the image representation. Both the Simple RNN and M-RNN both have a softmax layer which generated the probability distribution of the next word but the dimension of this layer is the size of vocabulary M.[14].

3. THEORETICAL ANALYSIS

3.1 VGG Neural Networks:

These networks are developed by the researchers Simonyan and Zisserman from the Oxford visual geometry group (VGG) for the competition ILSVRC 2014. Before the development of VGG, AlexNet was used which is a revolutionary advancement, it improved the traditional Convolutional Networks (CNN) and until the development of model VGG, AlexNet was the best model for image classification. AlexNet derivatives mainly focus on the smaller window size and strides in the first convolutional layer. VGG is a convolutional neural network model which addresses the important of aspect of CNN i.e. depth and it is considered as one of the excellent vision model architecture till date and used for the object recognition.

Architecture of VGG:

Input: The input for the VGG model is 224x224 pixel RGB image

*Convolutional Layers***:** In VGG the convolutional layer use a very small 3x3 receptive field even though it is the least possible size but it still captures up, own, left, right, centre. There are also 1x1 convolutional filters which are seen as linear transformation for the input and it is followed by the RELU (rectification) unit. There is a convolutional stride which is fixed to 1 pixel so that after convolution the spatial resolution is preserved. Spatial pooling is carried out by the layers known as max-pooling layers which follow some of the convolutional layers.

ISSN PRINT 2319 1775 Online 2320 7876
Research paper **Volume 11, Sp.Iss 6 , 2022**

Fully Connected layers: There will be three fully connected layers in which the first two layers have 4096 channels, the third layers has 1000 channels for each class and the final layer is the soft max layer

Hidden layers: All the hidden layers in VGG uses ReLU(rectification) unit and none of the layers contain LRN(Local response Normalisation) because LRN will not improve the performance but It increases the computation time and memory consumption.

Max Pooling Layer: Max pooling or maximum pooling is a calculation which calculates the largest or maximum value in each patch of every feature map and the results are pooled or down sampled feature maps that gives the most available feature in the patch.

Soft Max layer: It is the activation function which is applied on the output of the last layer and it is used particularly in multi class classification because it returns the discrete probability.

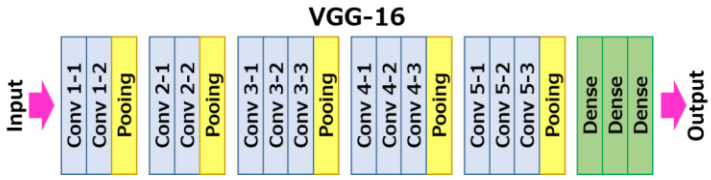


Fig 3:VGG-16 Architecture

3.2 LSTM Networks:

Long Short-Term Memory learning model isthe special kind of RNN which isadept of Long-term dependencies. In 1997, Hochreiter and Schmidhuber introduced this model.It became very popularized and refined by many people. Now these networks are widely in many areas and they work tremendously on different varieties of problems. These networks are explicitly designed to avoid the problem of long term dependency i.e. remembering the information for longer time intervals. All the RNNs will have the chain of repeating models of neural networks. The repeating module for the RNN is very simple containing a single tanh layer but for the LSTMs the repeating module has a different structure, Instead of containing a single Neural Network layer, it consists of four and those are interacting in a very special way.[12]

The key to the LS-TM is a State of cell and it has the capability of removing and adding the information to the cell state and it is regulated carefully by the components called gates and these are used to let the

ISSN PRINT 2319 1775 Online 2320 7876
Research paper **Volume 11, Sp.Iss 6 , 2022**

information pass through them.Thesigmoid layer gives the output range in between 0 and 1 that describes how much should be let through. A value of one indicates “let everything through” zero indicates “let nothing through” the initial step is to decide what information is going to pass from the cell state and this decision is taken by a sigmoid layer called as forget gate layer” and outputs the number between the range 0(get rid of this) and 1(keep this). The subsequent step is to decide what information is to be stored in the cell state in two parts, 1st part sigmoid layer(input gate layer) decides which values to be updated and 2nd part tanh layer creates new values that can be added to the state and these two are combined to create an update to the state.

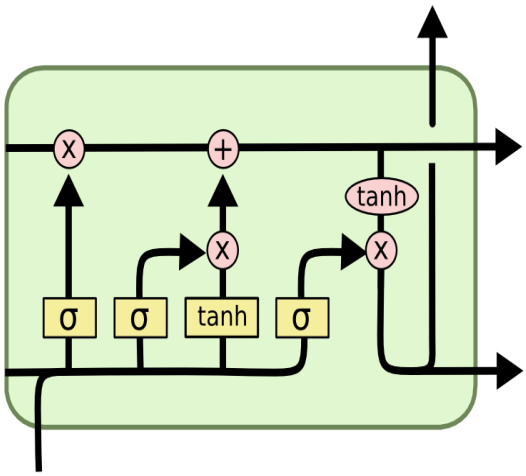
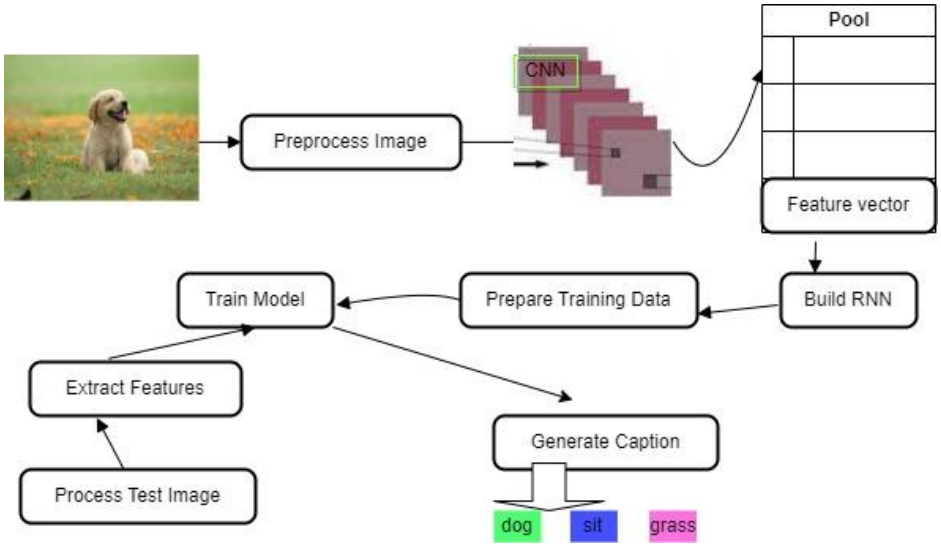


Figure 4:Bi-LSTM

4. Block Diagram:



ISSN PRINT 2319 1775 Online 2320 7876
Research paper **Volume 11, Sp.Iss 6 , 2022**

Figure 5: Block Diagram of the caption generation

Figure 5 depict the various layers of the caption creation process indetail.

5. ALGORITHM:

ISSN PRINT 2319 1775 Online 2320 7876

Research paper Volume 11, Sp.Iss 6 , 2022

Step1: Prepare text data by performing data cleaning:

Load_doc(): load the document and read the contents in side the file in to a string.

Load _description():Get dictionary of photo identifiers to descriptions.

Clean_description: Reduce the size of the vocabulary of words

To_Vocabulary():Separate all the unique words and create vocabulary from the all descriptions

Save_description: Create a list of descriptions that have been preprocessed and store them into a file filedescriptions.txt and store all the captions.

Step 2: Extracting the_feature vector from all images:

Extract_features: Extract features for all images and map image names with their respective feature array. Then these features are dump into features dictionary features.pkl, pickle file.

Step 3: Loading the dataset for Training Model

Load_photos(): Load the text file in a string and will return the list of image names

Load_clean_descriptions(). Create dictionary that contains captions for each photo from the list of photos

Load _features(): Get dictionary for image names and their feature vector which we have previously extracted from the VGG16().

Step 4.: Tokenizing the vocabulary: will map each word of the vocabulary with unique index value so that the Computer can understand

Step 5: Create data generator

Step6: Defining CNN-RNN model

Step7 : Training the model

Step 8 : Testing the model

Step 9 : Evaluating the model using BLEU score

Step10: Generating the Caption for the image

The above algorithm explains about the steps involved in implementing image captioning.

6. RESULTS

The dataset handled in this work is Flickr8k dataset. It consists of two different zip files.

1. *Flickr8K_Dataset*: The dataset contains totally 8092 images of different sizes in jpeg format. 6K Images of 8K images are used for training and the remaining 2000 images, 1000 images for testing and the other 000 images for development.
2. *Flickr8K_text*: This file contains the text files describing the training set, test set, token.txt and for each image it contains 5 captions i.e. total 40460 captions

ISSN PRINT 2319 1775 Online 2320 7876
Research paper **Volume 11, Sp.Iss 6 , 2022**

6.1 SAMPLE INPUT

Figure 5:Input Image

It will take the image as the input and predict the caption automatically using the model

***6.2* SAMPLE OUTPUT**

startseq the white cat is walking in the road endseq

The model will take the image as input and it predict the above caption for the image. After predicting the caption the model will compare the catpion by using BLEU Score.

6.3 Image Captioning Model

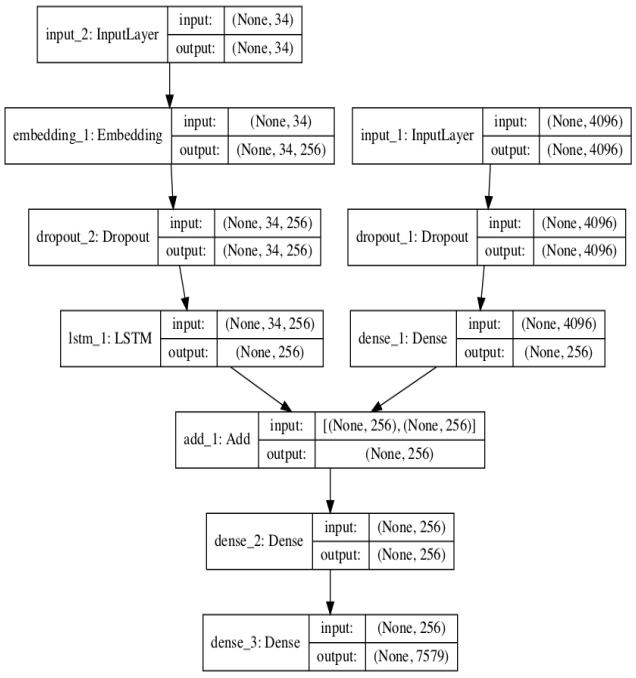


Figure 6: Plot of the image captioning deep learning mode

ISSN PRINT 2319 1775 Online 2320 7876
Research paper **Volume 11, Sp.Iss 6 , 2022**

6.4 PERFORMANCE MEASURES

BLEU-Score: BLEU-Bilingual Evaluation Understudy proposed by Kishore Papineni and it is a score used for comparing the candidate translation(generated sentence) of text to one or more reference translations and is used to evaluate the text generated for the natural language processing tasks. A perfect match results the score of 1.0 and a perfect mismatch results a score of 0.0. This score is developed for evaluating the predictions made by the automatic translation systems.

In this project we used BLEU score for evaluating the model and used BLEU-1, BLEU-2, BLEU-3, BLEU-4

6.5 EVALUATION RESULTS

BLEU SCORE	Results
BLEU-1	0.614035
BLEU-2	0.371077
BLEU-3	0.210103
BLEU-4	0.107481

Table 1: Evaluation Score

This is the table of results after evaluating the model. Earlier traditional methods are used for Image captioning. But because of drawbacks like less accuracy, Neural Networks methodologies came into existence. Neural Network methods are Convolutional Neural Networks (CNN), Bi Directional Long Short Term Memory NeuralNets, By using these technologies we increased the accuracy.

If we give the image as input by using the above model it will display automatically the textural description by observing the content of the image

7. CONCLUSON

In this paper we implemented the caption generator model by using CNN and LSTM and by using this model we can predict the caption for the provided image and also it evaluates the descriptions by using BLEU(Bilingual Evaluation Understudy) Score and also increased the accuracy by using the encoder and decoder model.

ISSN PRINT 2319 1775 Online 2320 7876
Research paper Volume 11, Sp.Iss 6 , 2022

8. REFERENCES

1. Chetan Amritkar, Vaishali Jabade, (2018)"Image Caption Generation using Deep Learning Technique", 978-1-5386-5257-2/ IEEE
2. Long Chen, Hanwang Zhang, Jun Xiao "SCA-CNN: Spatial and Channel-wise Attention in Convolutional Networks for Image Captioning", 12 , 2017
3. Jeff Donahue, Lisa Anne Hendricks, Marcus Rohrbach, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, Trevor Darrell, "Long-term Recurrent Convolutional Networks for Visual Recognition and Description", Manuscript received November 30, 2015.
4. Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan, (2016)"Show and Tell: Lessons learned from the 2015 MSCOCO Image Captioning Challenge" IEEE transaction on pattern analysis and machine intelligence,
5. Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan, "Show and Tell: A Neural Image Caption Generator" 20 apr 2015
6. Andrej Karpathy, Li Fei-Fei, "Deep Visual-Semantic Alignments for Generating Image Descriptions" 14 apr 2015
7. Ralf Gerber and Hans-Hellmut Nagel, "knowledge representation for the generation of quantified natural language descriptions of vehicle tra.ffic in image sequences" 0-7803-3258-X/96/$5.00 *0* 1996 IEEE
8. Yansong Feng and Mirella Lapata,(2010) "How ManyWords is a Picture Worth? Automatic Caption Generation for News Images" Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 1239–1249
9. Ali El Housseini, Abdelmalek Toumi, Ali Khenchaf, "Deep Learning for Target recognition from SAR images" 7th seminar on detection systems: architectures and technologies (dat'2017) february 20-22, 2017, algiers, algeria.
10. Marc Tanti, Albert Gatt, Kenneth P. Camilleri, "What is the Role of Recurrent Neural Networks (RNNs) in an Image Caption Generator?" 25 Aug 2017
11. Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo "Image Captioning with Semantic Attention" 12 Mar 2016
12. Cheng Wang, Haojin Yang, And Christoph Meinel," Image Captioning with Deep Bidirectional LSTMs and Multi-Task Learning" ACM Trans. Multimedia Comput. Commun. Appl., Vol. 14, No. 2s, Article 40. Publication date: April 2016
13. Junhua Mao, Wei Xu & Yi Yang &JiangWang& Zhiheng Huang, Alan Yuille, "deep captioning with multimodal recurrent neural networks (m-rnn)" 11 jun 2015