ISSN 2278-3091
Volume 9, No.5, September - October 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse204952020.pdf
https://doi.org/10.30534/ijatcse/2020/204952020

Identifying the concept of Image and Captioning Using Deep Neural Networks

S.Sagar Imambi[1], S Naganjaneyulu [2], N.V. Nikhila[3]
[1] Associate Professor, Koneru Lakshmaiah Education Foundation, Green Fields, Vaddeswaram, Andhra Pradesh, India,simambi@gmail.com
[2] Professor, Dept. of Information Technology Lakireddy Bali Reddy College of Engineering, Mylavaram, Andhra Pradesh, India.
[3] Graduate student, Koneru Lakshmaiah Education Foundation, Green Fields, Vaddeswaram, Andhra Pradesh, India.

ABSTRACT

In recent years, Image captioning has become a challenging artificial intelligence problem. Many researchers have been interested in the field of AI and became an arduous and exciting task. Image captioning automatically generates the textual description consistent with the content observed in a picture, and it is the mixture of two methods, including computer vision and natural language processing. Computer vision is to understand the images' content and natural language processing to understand the image into words in the correct order. Recently, Deep learning methods are achieving better results on caption generation problems. 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 Neural Network (CNN) implicitly extract features from the image, and Recurrent Neural Network is used for sentence generation. The developed model was trained to capture the image concept and generate the textual description observed in an image.

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

1. INTRODUCTION

Recently, many researchers from the field of Artificial Intelligence are doing their research on this challenging problem to develop a model that can automatically generate the textual description by understanding the content of an image by using well-formed English sentences. [5]Image captioning is an arduous and challenging task that will have a significant impact on many applications. It helps visually handicapped people in understanding the content of the images, and in medical applications like Skin vision in confirming whether a skin condition is skin cancer or not. It can also be applied in many areas, including the military, education, web searching, commerce, and social media [2]. We will see lots of images of various sources like news articles, internet, advertisements, and document diagrams every day. The viewers have to demonstrate the images themselves because most of the images do not have the image's narration. Even though a human can understand the images without detailed captions, it is difficult for a machine to demonstrate or interpret the captions for an image. Image captioning is a model build to interpret the detailed captions from an image. This process is as fast and as accurate as human by the machine [4]. Early image caption generation combines the information by some static object class libraries using statistical language models. Gaizauskas and Aker used dependency models for automatically tagging the images. Li et al. proposed an n-gram model. Yang et al. proposed a model using the hidden Markov model parameters, and may indirect methods are also proposed earlier for image captioning [6]. All the methods described or proposed by the researchers have their characteristics, and they are brainstorming. However, they all share a common disadvantage, i.e., they cannot make an instinctive feature observation on actions or objects in an image or did not give an end-to-end model to solve this problem. The advances of deep learning methods have made many breakthroughs had new hopes in image captioning.

2. UNITS LITERATURE SURVEY

2.1 Image Captioning

If anyone asks what you see in the above picture, some will say that there is a dog in a grassy area, some will say a dog with brown spots in grassy areas like others give some other captions.

Figure 1: Image of dog

All the captions are related to this image only. Nevertheless, the point is that it is easy for us as humans to see the image and describe the image in an appropriate language but write a computer program that takes the image as the input and produces a relevant caption as the output is about image captioning [16].

Image captioning involves mainly two methods Computer vision and natural language processing. However, recognizing and describing the images and videos is the fundamental challenge of computer vision. Image captioning can be done using Deep learning models like supervised Convolutional Neural Network (CNN) [4]. Natural Language Processing helps in image captioning.[24]

2.2 Convolution neural network

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

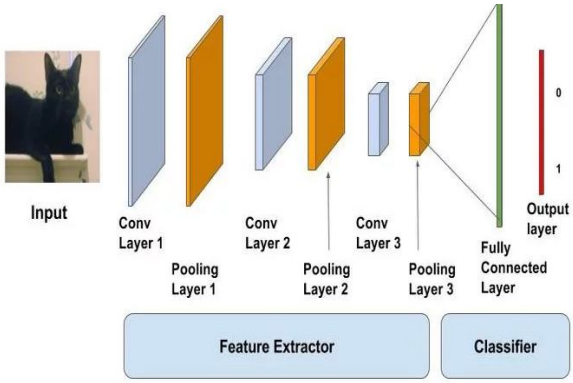

Figure 2: Convolutional Neural Network Architecture

- Convolutional layer
- Pooling layer
- Fully connected layer

A. 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 us consider an image of the type of jpg and size be 480 x 480. Then it has converted to 480 x 480 x 3 image where the represents the RBG values. To describe the pixel [15], they are given numbers from 0 to 255. Further, the array with numbers is given as input to the image classification.

B. Convolutional layer:
Convolutional Layer is an essential part of image classification. The main task in this layer is extracting features from the input image. A convolutional layer is a combination of feature maps and is used to extract regional characteristics of various positions in the former surface [14]. Nevertheless, its extraction is a 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 features are extracted from the image using a kernel.

Pooling layer:
A problem with the Convolution layer's output 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., downsampling. The pooling layer decreases the dimensions of the feature map. Two types of standard pooling techniques are used in reducing the dimensionality. They are max pooling and the average pooling. In max pooling, the max value of each patch in the feature map was calculated. Whereas average pooling, finding the average of each patch in the feature map. Fig 3 is an example of max pooling.
Fully Connected layer:
It is used to establish a connection between the output of the previous layer. There is no spatial arrangement in this layer. There can be many fully connected layers, and the last layer is connected to the output layer. One of the frequently used methods is soft regression because of its performance. Other methods like SVM can also be used with CNN to solve the more complex task.

2.3 Recurrent neural network

RNNs are the robust network architectures for processing the data and have been widely used in speech recognition, natural language processing, and handwritten recognition in recent years. Here in these networks, they allow the cyclical connection. The weights can be reused across various instances of neurons. 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. However, these traditional RNN cannot learn the long term dependencies present between the inputs and outputs.[7].

2.4 Other Long term recurrent Convolutional Network (LRCN)

LCRN combines the deep hierarchical visual feature extractor like CNN with a model that 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 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. The outputs are passed to a set of Recurrent Sequence models, i.e., middle right, and finally produce a variable-length prediction, i.e., right. Both LSTM and CNN weights are shared across time and resulting in a representation of long sentences. [9]

2.5 Multimodal Recurrent Neural Network(M-RNN)

The architecture of the 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, a 512-dimensional multimodal layer concatenates the vision of the M-RNN model and the language model part. It has three inputs recurrent layer, the embedding layer, and the image representation. A Softmax layer generating probability distribution of upcoming words is present in both the Simple RNN and M-RNN. The dimension of this layer is the size of the 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). Until the development of model VGG, AlexNet was the best model for image classification. AlexNet derivatives mainly focus on window size with smaller intervals and steps in the first convolutional layer. VGG is a predefined convolutional neural network model x that addresses the critical aspect of CNN, i.e., depth. It is considered an excellent vision model architecture to date and is used for object recognition.

A. Architecture of VGG

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

Convolutional Layers**:** In VGG, the convolutional layer uses a very small 3x3 receptive field even though it is the least possible size, but it still captures up, own, left, right, center. There are also 1x1 convolutional filters, which are seen as a linear transformation for the input, and it is followed by the RELU (rectification) unit. There is a convolutional stride that 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 that follow some convolutional layers.

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

Hidden layers: All the hidden layers in VGG use ReLU(rectification) unit. 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 that calculates the largest or maximum value in each patch of every feature map, and the results are pooled or down-sampled feature maps that give the most available feature in the patch.

SoftMax layer: The activation function applied to the last layer's output is used mainly in multi-class classification because it returns the discrete probability.

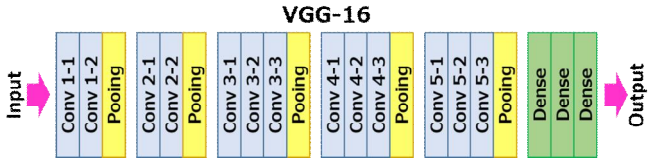

Fig 3:VGG-16 Architecture

3.2 LSTM Networks

Long Short Term Memory Networks are usually called as LSTMs. These are the particular kind of RNN which are capable of Long-term dependencies. Hochreiter and Schmidhuber introduced LSTM in 1997 and were popularized and refined by many people. These networks are widely in many areas, and they work tremendously on different varieties of problems. Long-term dependency can be avoided by these networks, 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 straightforward, containing a single tanh layer. However, for the LSTMs the repeating module has a different structure. Instead of containing a single Neural Network layer, it consists of four interacting significantly.[12].

The key to the LSTM is a cell State, and it has the capability of removing and adding the information to the cell state. It is regulated carefully by the structures called gates, and these are used to let the information pass through them(composed of a sigmoid layer and a pointwise multiplicative operation). The sigmoid layer gives the output range between 0 and 1, describing how much should be let through. A value of one indicates "let everything through" zero indicates "let nothing through" the first step is to decide what information is going to pass from the cell state

. 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). Next, we 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. The status is updated after combining these two.

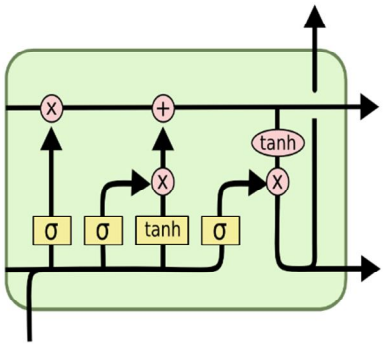

Figure 4: Bi-LSTM

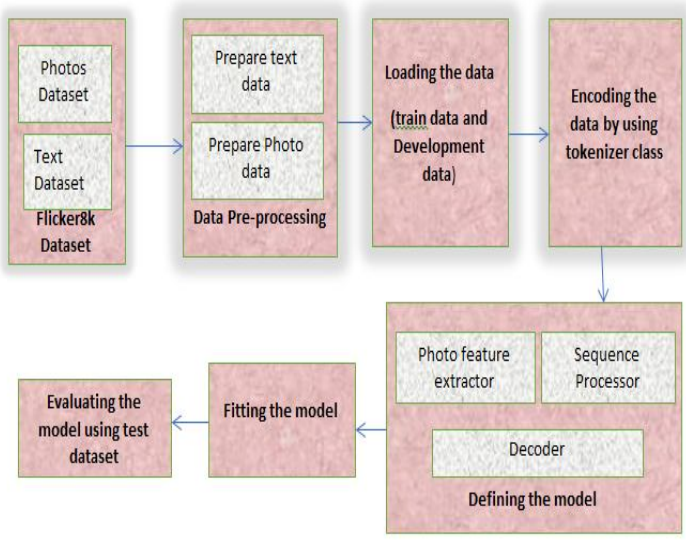

Figure 5: Block Diagram

5.ALGORITHM:

Step1:

Prepare text data by performing data cleaning:

Load_doc(): load the document and read the contents inside the file into 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 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. : This function creates a dictionary that contains captions for every photo from the list of photos.

Load _features: Get a 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 a unique index value so that the computer can understand

Step 5: Create a data generator

Step 6: Defining CNN-RNN model

Step 7: Training the model

Step 8: Testing the model

Step 9: Evaluating the model using the BLEU score

Step10: Generating the Caption for the image

The Above algorithm explains the steps involved in implementing image captioning.

6. RESULTS

The dataset used in this paper is the Flickr8k dataset. It consists of two different zip files.

1. Flickr8K_Dataset: The dataset contains a total of 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 every image, it contains five captions, i.e., a total of 40460 captions

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 predicts the above caption for the image. After predicting the caption, the model will compare the caption by using BLEU Score.

6.3 Image Captioning Model

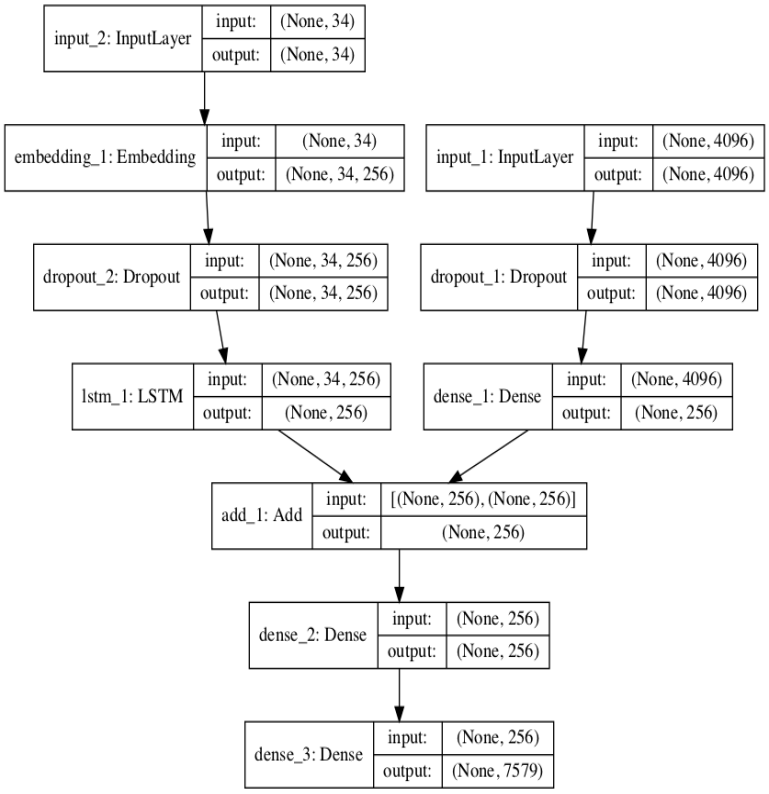

Figure 6: Plot of the image captioning deep learning mode

6.4 Performance Measures

BLEU-Score: BLEU-Bilingual Evaluation Understudy proposed by Kishore Papineni. 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. !.0 score indicated a perfect match, and a perfect mismatch results in a score of 0.0. This score is developed for evaluating the predictions made by the automatic translation systems.
In this project, we used the BLEU score for evaluating the model and used BLEU-1, BLEU-2, BLEU-3, BLEU-4.

6.5 Evaluation Results

Table 1: Evaluation Score

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

The above table shows the results after evaluating the model. Earlier traditional methods are used for image captioning. However, 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 neural nets. By using these technologies, we increased our accuracy.
If we give the image as input by using the above model, it will automatically display the textural description by observing the image's content.

7. Conclusion

:

Image captioning is essential in many applications. We implemented the caption generator model by using CNN and LSTM. This model predicts an image's caption and evaluates the descriptions using the BLEU(Bilingual Evaluation Understudy) Score. This model captioned the image with 70% accuracy. To improve accuracy, this model may be further extended.

REFERENCES

1.. Chetan Amritkar, Vaishali Jabade, "**Image Caption Generation using Deep Learning Technique**", 978-1-5386-5257-2/18/$31.00 ©2018 IEEE

2. Rahul Shahne, Mohammed Ismail, CSR Prabhu "**Survey on Deep Learning Techniques for Prognosis and Diagnosis of cancer from Microarray Gene Expression Data**" Journal of computational and theoretical Nanoscience 16 (12), 5078-5088, Dec 2019.

3.Long Chen, Hanwang Zhang, Jun Xiao "**SCA-CNN: Spatial and Channel-wise Attention in**

Convolutional Networks for Image Captioning", 12 April 2017

4. 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.
5. Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan, "**Show and Tell: Lessons learned from the 2015 MSCOCO Image Captioning Challenge**" IEEE transaction on pattern analysis and machine intelligence, vol.xx, no. xx, month 2016
6. Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan, "**Show and Tell: A Neural Image Caption Generator"** 20 Apr 2015
7. Andrej Karpathy, Li Fei-Fei, "**Deep Visual-Semantic Alignments for Generating Image Descriptions**" 14 Apr 2015
8. Lakshmi Praneetha, Vidyullatha P, "**Automated leaf disease detection in corn species through image analysis,**" International Journal of Advanced Trends in Computer Science and Engineering, 2019.
9. Yansong Feng and Mirella Lapata, **"How many words is a Picture Worth? Automatic Caption Generation for News Images**" Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp 1239–1249, July 2010
10. 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) pp:20-22, 2017.
11. Marc Tanti, Albert Gatt, Kenneth P. Camilleri, **"What is the Role of Recurrent Neural Networks (RNNs) in an Image Caption Generator?"** 25 Aug 2017
12. Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo **"Image Captioning with Semantic Attention"** 12 Mar 2016
13. 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
14. Junhua Mao, Wei Xu & Yi Yang & JiangWang & Zhiheng Huang, Alan Yuille, "**Deep captioning with multimodal recurrent neural networks (m-rnn)**" 11 Jun 2015
15. Phani Madhuri, N., Meghana, A., Prasada Rao, PVRD, Prem Kumar, P., **"Ailment prognosis and propose antidote for skin using deep learning** ", International Journal of Innovative Technology and Exploring Engineering,, 2019
16. Shah, P., Bakrola, V., & Pati, S. "**Image captioning using deep neural architectures**". International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS).
17. Bangare, S. L., Pradeepini, G., & Patil, S. T. (2017). **Neuroendoscopy adapter module development for better brain tumor image visualization**., International Journal of Electrical and Computer Engineering, 7(6), 3643-3654. doi:10.11591/ijece.v7i6.pp3643-3654.
18. Banerjee D, S.Sagar Imambi. " **Opinion mining for drug reviews "** International Journal of Innovative Technology and Exploring Engineering. Vol 8 No.7, pp. 1314- pp.13-18
19. S. Rizwana, S.Sagar Imambi "**Enhanced biomedical data modeling using unsupervised probabilistic machine learning technique**" International Journal of Recent Technology and Engineering,,2019, vol7.No 6. Pp 579-582.
20. Greeshma, L., and G. Pradeepini. "**Mining Maximal Efficient Closed Itemsets without Any Redundancy**." Information Systems Design and Intelligent Applications. Springer, New Delhi, 2016. 339-347
21. R. Bhimanpallewar, M. R. Narasinagrao, "**A Machine Learning Approach to Assess Crop Specific Suitability for Small /Marginal Scale Croplands"** International Journal of Applied Engineering Research 2017 Vol.12, NO.23, ppno.13966-13973.
22. Anitha, R., Jyothi, S., Mandhala, V. N., Bhattacharyya, D., & Kim, T. -. (2017). **Deep learning image processing technique for early detection of alzheimer's** disease. International Journal of Advanced Science and Technology, 107, 85-104.
23. Ahammad, S.H., Rajesh, V., Neetha, A., Sai Jeesmitha, B. & Srikanth, A. 2019, "**Automatic segmentation of spinal cord diffusion MR images for disease location finding**", *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 15, no. 3, pp. 1313-1321.
24. J Koteswaa Rao, L.Rohini, P & Narayana .M 2019,"Lemp, **A robust feature descriptor for retrieval applications**" International Journal of Engineering and Advance Technology Vol9.,no,1pppp4569-
25. Koteswara Rao, L., Rohni, P. & Narayana, M. 2019, "**Lemp: A robust image feature descriptor for retrieval applications**", International Journal of Engineering and Advanced Technology, vol. 9, no. 1, pp. 4565-4569.,2019