Research Article

An Advanced Image Captioning using combination of CNN and LSTM

Priyanka Raut[1], Rushali A Deshmukh[2]

[1]Savitribai Phule Pune Univresity, PG research Scholar, Maharhatra/India
[2]Savitribai Phule Pune Univresity, faculty, Maharhatra/India

Article History: Received: 10 November 2020; Revised: 12 January 2021; Accepted: 27 January 2021; Published online: 05 April 2021

Abstract: The Captioning of Image now a days is gaining a lot of interest which generates an automated simple and short sentence describing the image content. Machines indeed are trained in a way that they can understand the Image content and generate captions which are almost accurate at a human level of knowledge is a very tedious and interesting task. There are various solutions used to solve this tedious task and generate simple sentences known as captions using neural network which still comes with problems such as inaccurate captions, generating captions only for the seen images, etc. In this paper, the proposed system model was able to generate more precise captions using a two staged model which consists of a combination of Deep Neural Network algorithms (Convolutional and Long Short-Term Memory). The proposed model was able to overcome the problems arise using Traditional CNN and RNN algorithms. The model is trained and tested using the Flicker8k Data set.

Keywords: Pre-processing, CNN, RNN, LSTM, Feature extraction, Feature Vector, Filters, Image Captions

1. Introduction

Image Caption generation is a task in which a machine model is trained using artificial Intelligence in a way that the machine can understand the Image scene at a same level as human beings do understand the visual world. Image Captioning is basically like a short description generated by just looking at the image visually. In this task, an machine is feed with an input image and based on the intelligence and training given, the model generates a simple caption which indeed explains the content of the image in a human readable form. This task is a Supervised learning algorithm example. Such task become more challenging when a machine has to generate a caption for unseen or not trained images. Generally, a model tries to break an image down into objects and classify these objects before generating sentence or caption. Captioning of image basically aims towards generating natural language and simple captions which describes the image content accurately. In this task, all objects and their relationship should be depicted precisely. A traditional algorithm which is a combination of Convolutional and Recurrent network used for generating captions has many problems such as gradient vanishing, not so accurate identification of objects and their relationship or generation of captions only for seen images, etc. An Automatic Image Captioning Model which is a combination of advanced Convolutional and Long Short-Term Memory Deep Neural Network algorithms (CNN and LSTM) is a variation of traditional method to overcome the problems that arises using traditional way of captioning. The Model is divided into two stages: First stage uses Convolutional algorithm and second stage uses Long Short-Term Memory. The input to the first stage is image/picture. The proposed system model also focuses on the informative captions that best describes the image scene. In the proposed system model, the first stage known as Encoder stage is feed with image vector where the image is already pre-processed and then gave as input to Stage 1. At this stage, various convolutional layers are applied on the vector which fetches appropriate features from the provided vector before sending it to next stage. After applying number of convolutional layers/operations on the image vector , it is then send to next stage which is Decoder stage . Stage 2 processes the image vector given by the Stage 1 in a linear way to generate captions. The methodology uses LSTM algorithm in Stage 2 which is advanced version of recurrent neural network (RNN) helps to overcome the gradient explosion problem. LSTM has an advantage as it various memory gates which decides the flow of the information which the Stage 2. It also has an advantage to retain the data for longer period of time and dependencies. This Stage 2 outputs an sequential decoded simple language sentence or captions for the given input image.

1.1. Problem Statement

To design a system to generate a Accurate caption based on the Input Image using Convolutional neural Network (CNN) and Long Short Term Memory algorithm (LSTM).

1.2. Literature Survey

Jie Wu et al. [1] discussed the various Image Captioning methods, the captions are composed of most frequent words used. They explained the proposed Method which included the Content Sensitive and Global Discriminative to generate accurate and concrete captions. The Content Sensitive in this method focuses on the

less frequently used words and more concrete phrases or words to justify the image content better. They also introduced Global Discriminative in this methodology which focused on pulling the caption which is more accurate and which described the image better than any other methodology. In this paper, they have introduced a new concept which uses the less frequent words such A young girl such as young = adjective's is an less frequent word and the girl is a frequently used. Another concept they have used is to pull the image which justifies the sentence and to push away the other images.

Mingxing Zhang et al. [2] have presented the Lack of insufficient concepts using the traditional methods of the Image Captioning. They have also explained various reasons for the lack of concepts problem in detail. The one of the reason they explained is the difference between the number of occurrences of the positive samples and the negative samples of the concept. The other reason they have mentioned is incomplete labeling of the captions during training the captions. They proposed a methodology known as the Online Positive Recall and the Missing of the Concept Missing to resolve the above problems. The methodology generated high accuracy Captions and also were able to identify more of the semantics. They also described the 2-stage optimization methodology for resolving the missing concept mining.

Kun Fu et al. [3] proposed a methodology, Image-Text Surgery for the Captioning of the Image which used its own generated pseudo image and text pairs method to generate a caption. In this paper, the explained the various efficient learning methodologies using pseudo pairs. The pseudo pair is the Image and text pair which were generated using a unique knowledge base which is subset of MSCOCO dataset and has it own syntax known as seed syntax. They have used the concept of pseudo pair to avoid any human labeled data to avoid the human intervention. They evaluated the model against the subset of MSCOCO dataset. The methodology shows significant improvement in the results. Their methodology is robust which showed better results than traditional methodologies. They have also given a brief on the knowledge base construction.

Vishwash Batra et al. [4] proposed a model for news images and caption generation for various news articles. The proposed system was trained using images, descriptions and captions. The idea of adding long description was an advancement which shows an great improvement the results. The method is different than other traditional methodologies. The input along with image was a long description which increases the chances of correct caption generation. The more the vocabulary better the results. They used RNN and CNN deep neural network algorithm. Example: If the image has a building in it. It is difficult to predict if its a school or a society. The long description along with an image helps to predict the correct caption. They have also explained the use of the proposed model at various news sites such BBC, etc. The have used BBC News Corpus to perform their experiment.

Min Yang et al. [5] proposed a methodology known as "MLADIC" algorithm for the Cross-Domain Image Captioning. They have introduced and explained the method to reduce the difference between cross-domain such as of the source and the target. They have trained the system to learn the alignment of the image and text pair. They tested the model which was feed with few selected pictures and its description which was not paired in the target.

Chetan Amritkar et al. [6] presented the method where the image caption i.e. The content of the image can be generated using CNN and RNN. They have explained the Computer Vision and the use of Natural Language Processing, NLP. The system model generated a vector of features using CNN algorithm and RNN algorithm to decode the vector into natural captions. The explained the methodology requires both the Image and text processing. They have highlighted the RNN problem which is vanishing of the gradient.

Parth Shah et al. [7] presented a methodology which is an advancement task in recognizing of the objects and machine translation which shows an improvement in the results of the captioning model. They introduced the deep neural network algorithms such as CNN which is used to generate the caption. The Convolutional neural network is explained in detailed. They have evaluated the performance of the proposed model using various standard ways of evaluation matrices.

Aghasi Poghosyan et al. [8] presented and highlighted the most important issues in the existing Captioning Models which is the model needs to predict the next word in the process where it is dependent on the last predicted word in the sequence. This is a very difficult task for a system. The model generated Caption using RNN algorithm which was modified and later LSTM was used to generate the sequence of words using LSTM cell gates which resulted in better results. They explained the process of the next word generation to be predicted after the previous word was generated. As LSTM can store the previous output along with the current input for a longer time in the sequence.

Karpathy at al. [9] explained the algorithm using deep neural network that conclude the alignment which is latent between the image region and the natural language sentence segments. They also explained the CNN and a RNN which is bi-directional. They have introduced an attention method which focuses on the important features. They start with all features from top to important features using CNN layers which is one of the best methods. They have mentioned the disadvantage of this method which is losing important features or the information which may affect the quality and correctness of the caption generated.

Xu et al. [10] presented a suggestion which summarizes the attention during the generation process of the words related from different objects by LSTM system Model. They explained the use of neural network models for generating caption like humans do. They have also mentioned about the Encoding and decoding Model, the combination of Visual Attention and Image Captioning.

Vinyals, Oriol, et al. [11] presented the model known as the Neural Image Caption Model (NIC). The NIC uses CNN and RNN as a encoder and decoder algorithm. For Image Classification, CNN is pre-trained and its output is provided to RNN which generates natural sentences as captions. They have also discussed about the advancement in RNN algorithm which is LSTM.

2. Proposed System

The Image Captioning Model generates accurate captions, C in a very simple language matching the human level of imagination for the given input image, I. The Captioning Model uses advance version of CNN and LSTM algorithm to generate natural language captions.

2.1. Architecture

The Proposed System Architecture shown in the Figure 1 is divided into four Modules. The Input Image is first given as the input to the Image Based Module which uses CNN algorithm's Convolutional and Pooling layer, to generate a vector which is known as feature vector of the input image. After every Convolutional layer, a ReLu layer is used. And then Pooling layer is used to reduce the size of feature vector before passing it to the next model. The last layer of CNN, Fully Connected Network is excluded from this Model as we just need the feature vector. Convolutional and Pooling layer are used as Feature extractors where as the Fully Connected Network is used as Classifier. Now, the output of previous model which is vector of features generated is given to next Module, Language Based Module where the encoded features vector is decoded into a natural language caption using LSTM, Long Short Term Memory algorithm which is advanced version of Recurrent Neural Network has an advantage of storing long sequence of data . LSTM has a memory cell which can store the data for a longer period of time. The Sequence of sentence/caption has 2 special token which are startseq and endseq token so that the algorithm knows when to start the sequence and stop the sequence of sentence. The Caption is generated at last. The Captioning Model focuses on objects, Color, actions and relationship between the objects.

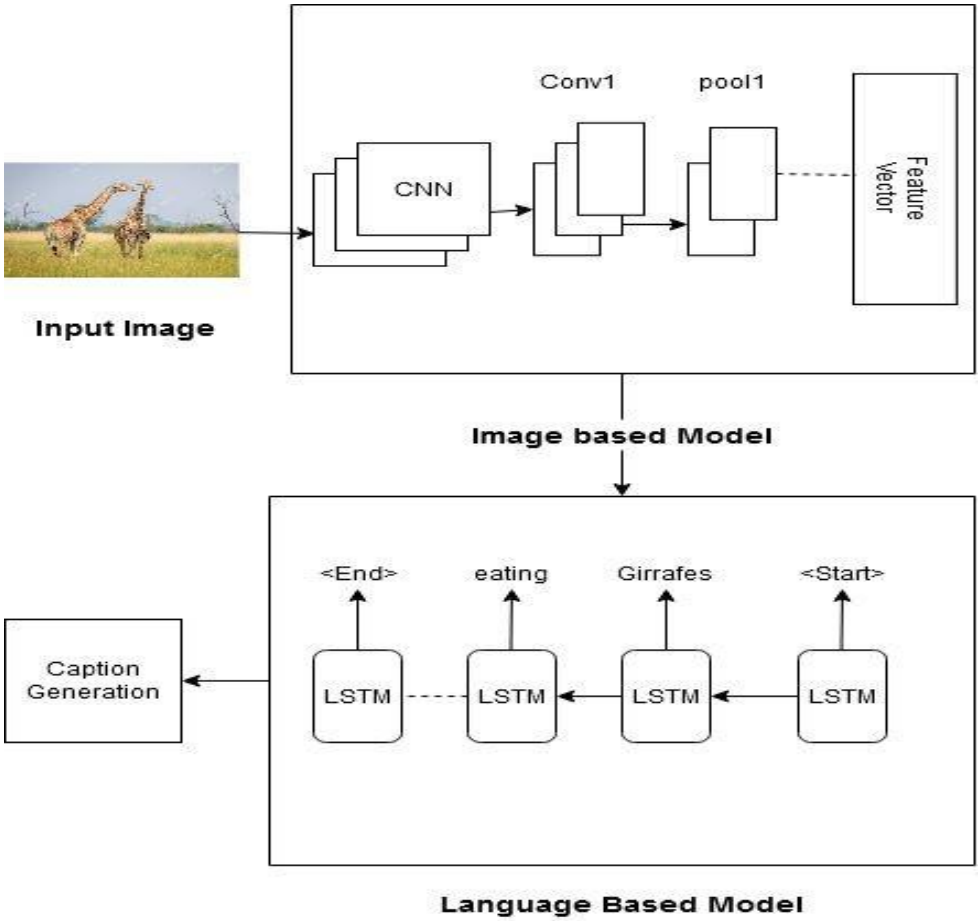

Figure 1. System Architecture

2.2. Modules

2.2.1. Image Preprocessing

The system/machine does not understand the Images. First the input image is converted into fixed sized pixel matrix (224x224x3) where the color code of each pixel is put at the respective locations. Then further every single image is pre-processed where the noise from the image is removed. The image later then is converted into the Gray-scale and then threshold value is set to divide the image into foreground and background. Edge detection is done on the every image objects. The final pixel matrix is the output of Image Pre-processing model and the input to the next Module.

2.2.2. Image based Model (CNN)

A modified version of CNN where Convolutional and Pooling layers act as the feature extractor in this Module. The input to the Image Based Module is the output of previous Pre-processing module which is matrix of pixels. This module extracts features from the image pixel matrix and store it into a feature vector. Convolutional layer is the first layer of CNN applied in this module to extract features. After every Convolutional, ReLU layer is applied. Then a pooling layer is applied to reduce the size of the feature vector without losing any features of the image. The Features such as, objects, the verbs that is the action of the object, the color of the object and the most important relationship between the object is fetched from this module and stored in a feature vector. The output of this module, the feature vector is the input of the next module. The size of the vector is linearly transformed to input size of the LSTM network which is used in next module.

2.2.3. Language Based Module (LSTM)

The Input to the Language Based Module is the linear feature vector for a given input image. The main aim of this module is to convert the encoded features into a simple language which can be understandable to the users using Long Short-Term Memory (LSTM). The module uses LSTM algorithm as it overcomes the variant

gradient issue of the RNN algorithm and also can store a long sequence of data without forgetting the sequence of the data. The LSTM uses its memory cells to store the data. For training The LB i.e. Language Based Model, we first have to pre-define our label and target text.The Label stores the data in a sequence starting with a start token and the Target stores the sequence of data with a end token at the end of its token so that the algorithm understand when to stop.

Example: Consider the caption is ”X and Z are playing basketball”.
Label: [start, X,and, Y,are, playing, basketball,.]
Target: [X, and, Y, are, playing, basketball, end].

2.2.4. Caption Generation

The Caption Generation module is the last module where the input to this module is the input from the previous Language Based Module. This Module’s aim towards generating caption in a linear sequence provided from previous module for given input image. At the end of this module, a caption in a simple human understandable form is generated.

3. Algorithm

3.1. Convolutional Neural Network (CNN)

Input: Pre-processed Pixel Matrix of the Input Image *I* to the Image Based Model System , *S*.
Output: Encoded Feature Vector, F_v . BEGIN
Initialize the number of filters = 32, stride = 5, epoch = 10 and weights required with random values.
for each Input Image *I*:
Step 1: Input the pre-processed Image pixel matrix to the Image Based Model.
Step 2: Apply convolutional layer to extract features from the image which performs depth, stride and zero-padding operations on the image pixel values.
Step 3: Apply ReLU operation after every convolutional layer. Step 4: Apply pooling layer for each feature in Feature Map. Step 5: Feature vector Fv of the input Image I.
end for END
Following layers of CNN are used in our proposed system as these two layers act as feature extractors for the given image.

3.3.1 Convolutional Layer

The Convolutional Layer is the first layer of CNN deep neural network algorithm. The input to this layer is a pixel matrix of input image which is pre-processed. In this layer, a feature vector is formed by sliding a selected filter over the input image and then computing a dot product of it. We first take a filter/feature. Then we multiply each image pixel with the corresponding feature pixel, We then add all the pixel values generated and finally divide them with the total number of pixels in it. Then we create a map and put the value of previous calculated filter in the map. These filters are extracting the features out of the input image forming a new feature vector. The number of filters is clearly equivalent to the increase in the number of features and better results.

The size of the final Map is controlled by 3 parameters of CNN which are decided before the convolutional layer is applied.

Depth: the number/quantity of filters or the features used for the convolutional operation on input image.

Stride: the quantity or the number of pixels values by which we slide on our matrix which is also called as filter matrix slided over the input matrix. When the stride = 2 then we shift the filter two pixel at a time.

Zero-padding: In some cases, it is necessary to pad the matrix with zero’s to be able to apply the filter on the borders of the image matrix.

Rectified Linear Unit, ReLU: After every convolutional layer, we apply ReLU operation in our proposed methodology. It is used to perform a nonlinear operation. In this operation, the negative values from the filtered images are replaced with Zero known as Activation function to avoid the summing up of all pixel values to zero.

3.3.2 Pooling Layer

The Pooling layer is used for Dimensionality Reduction of the matrix from previous layer. The Pooling layer shrinks the image feature vector into a smaller size without affecting the quality of features Various types of operations in pooling are Max, Min, Sum and Average on the matrix of pixels. We have set stride = 5 in our methodology. From each window, we select the maximum value and replace it in map. This dimensionality reduction helps to increase the computing speed inside the network. It also helps to overcome the issue of over fitting.

3.2. Long Short Term Memory (LSTM)

Long Short Term Memory is an advanced Version of Recurrent neural Network. The Recurrent Neural Network has a problem known as gradient vanishing problem. The gradient vanishes using back-propogation during the computations. LSTM has overcome the gradient vanishing problem. Long Short Term Memory algorithm used in the deep learning field has an feedback connections. LSTM is known to process an entire and long sequence of data. LSTM consists of a cell and gates (input, output and forget gate). The cell remembers the data in it for longer period of time and the use of the gates is to control the data flow inside the network. It decides what data needs to stored and sent to the network, what data is let out from the network and what data is forgot from the network. LSTM are used where there is a need of processing, predicting or classifying.

LSTM has a memory cell which stores the information for each time step. The LSTM basic architecture is shown in the Fig. 2 and contains the following variables:

Forget Gate "f_t"
Candidate layer "C" (NN with Tanh)
Input Gate "i_t" (input gate activation vector)
Output Gate "o_t" (output gate activation vector)
Hidden state "h_t" (hidden state vector)
Memory state (vector)
Input vector "x_t"

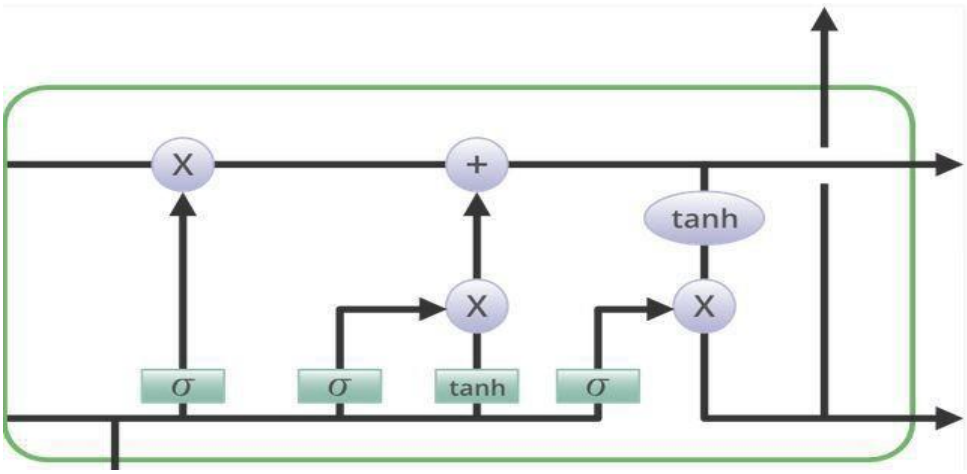

Figure 2. LSTM basic Architecture

The LSTM equation which takes the forward propagation at a time step t are shown below :
$it = \sigma (Wi [ht-1 , xt] + bi)$,
$ft = \sigma (Wf [ht-1 , xt] + bf)$,
$ot = \sigma (Wo [ht-1 , xt] + bo)$,
$gt = tanh (Wg [ht-1 , xt] + bg)$, $ct = f t \odot ct-1 + it \odot gt$,
$ht = ot \odot tanh (ct)$.

At a time step t, an LSTM cell has 2 inputs : x_t, input vector at given time step and h_{t-1}, hidden state vector of previous time step. The weight matrices is denoted by *W* and biases with *b*. The above equations are used to update the data in the network cell while the forward propagation process.

The LSTM algorithm generates the Caption C with two extra tokens in a sequential way,

$C = \{ w\text{-}1 , w0 , w1 , \ldots , wL \}$
where
w- 1 is the Start token,
wL is the End token,
L is the caption Length**.**

4. Experimental Setup

4.1. Data Set

We have used Flicker8k Data set for model experiments. Flicker8k data set can be easily downloaded and is suitable for small workstation such as laptops and Desktop. The model can be trained effectively using Flicker8k data set. The Flicker8k_dataset.zip file has:

Flickr8k_Dataset: This folder has 8092 images, each image with different sizes, shapes and colors. From 8092 images, 6000 images are used for training, 1000 images are used for development and the rest 1092 images are used for the testing the proposed model. The size of this dataset is 1 GB.

Flickr8k_text: The Flickr8k_text folder has a Flickr8k.token.txt file which has 5 captions per image for training the proposed model stored in the form of key- value pair where key = unique Image id and the value = Caption for the image. The size of this file is 2 MB.

5. Results

The proposed system is trained and tested using Flickr8k Data set. Figure 3. Allows user the select the image for which the caption needs to be generated using Automatic Captioning Model. Figure 4. displays the selected image and has a **Predict** button which once clicked, the image goes through pre-processing where every selected image is converted into grayscale image, then the threshold of the image is calculated and perform Edge detection of the input image and then the image is converted into fixed sized image vector which is 244x224x3. Then this vector is given as input to the CNN model to generate feature vector. The output of CNN model is the input of LSTM model. The LSTM model then generates the natural language Caption as shown in Figure 5.

Figure 3. Select an Input Image

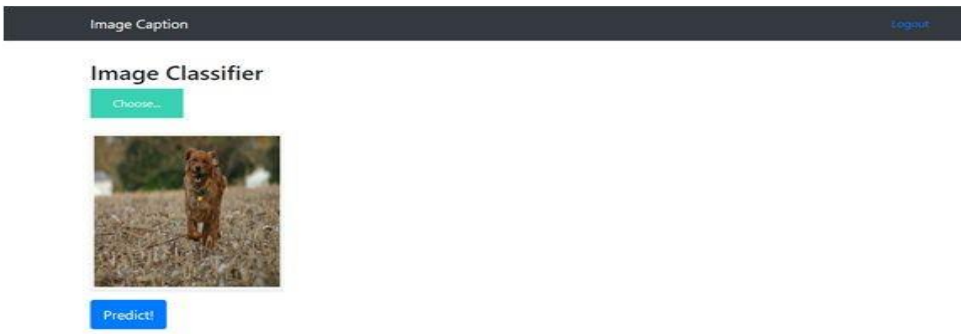

Figure 4. Click Predict to generate Caption of Input Image

Figure 5. Image Caption

6. Results

In this paper, we proposed a Automatic Image Captioning Model which is a combination of CNN and advanced RNN: LSTM for generating qualitative and accurate captions which could describe the image in natural and easy language. The Proposed Model is trained with 6000 Images using Flicker8k dataset which contained Images along with its captions. The proposed Convolutional deep neural network extracts the important features from image and stores it in feature vector. The feature vector is send to LSTM model to generate a sequential sentence combing the extracted features and their relationship to form a caption. The proposed Model generates precise captions for the Image. The model has reduced the error rate in the caption. The proposed system used LSTM algorithm to overcome the gradient vanishing problem of traditional RNN algorithm. The System is tested with 2000 Flicker8k dataset images. The System is accurately able to identify the objects in the images and their relationship. In Future, this proposed model can be extended where an system can be trained using images, its caption and also descriptions which helps improves the caption accuracy.

References

1. Jie Wu, Tianshui Chen, Hefeng Wu, Zhi Yang1, Qing Wang, and Liang Lin, (2019). Concrete Image Captioning By Integrating Content Sensitive And Global Discrimination Objective, in *IEEE International Conference on Multimedia and Expo (ICME).*
2. Mingxing Zhang, Yang Yang, Hanwang Zhang, Yanli Ji, Heng Tao Shen, Tat-Seng Chua (2018). More is Better: Precise and Detailed Image Captioning using Online Positive Recall and Missing Concepts Mining in *IEEET RANSACTIONS ON IMAGE PROCESSING*.
3. Kun Fu, Jin Li, Junqi Jin, and Changshui Zhang, Fellow (2018). Image-Text Surgery: Efficient Concept Learning in Image Captioning by Generating Pseudopairs *in IEEE Trans 2162-237X*.
4. Vishwash Batra, Yulan He, George Vogiatzis (2018). Neural Caption Generation for News Images in School of Engineering and Applied Science, *Aston University*.
5. Min Yang, Wei Zhao, Wei Xu, Yabing Feng, Zhou Zhao, Xiaojun Chen, Kai Lei, (2018). Multitask Learning for Cross-domain Image Captioning in *IEEE TRANSACTIONS ON MULTIMEDIA.*
6. Chetan Amritkar, Vaishali Jabade, (2018). Image Caption Generation using Deep Learning Technique in *IEEE, 978-1-5386- 5257-2/18/$31.00.*
7. Parth Shah, Vishvajit Bakrola, Supriya Pati, (2017). Image Captioning using Deep Neural Architectures in *IEEE International Conference on Innovations in information Embedded and Communication Systems (ICIIECS).*
8. Aghasi Poghosyan , Hakob Sarukhanyan, (2017). Long Short-Term Memory with Read-only Unit in Neural Image Caption Generator in *IEEE, 978-1-5386-2830-0/17/$31.00.*
9. Karpathy, A. and Fei-Fei, L., (2017). Deep visual-semantic alignments for generating image descriptions in *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(4):664–676, April*.
10. Xu and Kelvin, (2015). Show, attend and tell: Neural image caption generation with visual attention, in *International Conference on Machine Learning.*
11. Vinyals, Oriol, (2015). Show and tell: A neural image caption generator in Computer Vision and Pattern Recognition (CVPR), *IEEE Conference.*