International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 – 5161

Available at http://inpressco.com/category/ijcet

Research Article

Automatic Image Caption Generation using Neural Network

Ms. Priyanka Raut and Mrs. Rushali A Deshmukh

Department of Computer Engineering Rajarshi Shahu College of Engineering Pune , India

Received 10 Nov 2020, Accepted 10 Dec 2020, Available online 01 Feb 2021, **Special Issue-8 (Feb 2021)**

Abstract

Automatic Image Captioning has gained significant interest and aims towards generating Natural language sentences based on the given input image which is a very challenging task. It has been imagined that machines one day will understand the visual world at a human degree of knowledge using Artificial Intelligence. Although there are various solutions been generated by the Neural Networks in recent times which have given precise results, still a major problem arises is that the algorithm can only describe the concepts seen in the training data set with more error rate captions. In this paper, the proposed methodology consists of the Deep Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM) algorithm for generating natural language captions based on Images with more accurate results. We evaluate our methodology on Flickr8k data set.

Keywords: *Image Pre-processing, CNN Algorithm, Feature Extractors, Feature Vector, RNN, LSTM, Caption Generation*

Introduction

Image captioning is one such task in which the machine model learns to generate natural sentences on the given input image . In these tasks, we have to train a model to generate the caption for the images. This is an example of Supervised Learning

Algorithm. However, in such tasks like image classification, the content or description of an image contains object to be classified. The situation could be considerably more challenging when we expect machine model to understand complex scenes. Image captioning aims to generate natural language sentences to describe the objects and their relationship of input image or visual description generation, which aims towards generating simple text descriptions for an image often times capturing all the different objects depicted and their relationships. Automatic Caption Generation Model using Convolutional Neural Network (CNN) and Long Short Term memory (LSTM) algorithm is a bit different from traditional Image Caption model which uses CNN and RNN algorithm for caption Generation. The Input to the Model is Image. Rather than only showing and focusing on the relationship between the object , the proposed model also focuses outputs the informative text best describing the scenario in the Image. It is like a Story telling after observing a Image. Deep Convolutional Neural systems have as of late achieved state-of-the-craftsmanship execution on various picture acknowledgment benchmarks, including the Image Net. In the proposed algorithm , the model is trained with Image using Convolutional Neural Network (CNN) algorithm and Long Short Term Memory (LSTM) which helps in providing more precise results and to overcome the problems in traditional RNN algorithm. Standard RNN(Recurrent Neural Network) suffer from vanishing and exploding gradient problem. On the Other hand, LSTM solves the gradient problem by introducing new gates and maintains information in the memory for long dependencies. Convolutional Neural System provides Image Feature Vector which is then Given as input to the LSTM System to generate sequential Natural Sentences.

Literature Survey

Kun Fu et al. [1] proposed a novel method, Image-Text Surgery for the Image Captioning which synthesized the pseudo imagesentence pairs. The pseudo image-sentence pairs were generated with the knowledge base which used the syntax from a seed data set. They further introduced the adaptive visual replacement which filtered the unnecessary visual features in the pseudo data with this mechanism.

Parth Shah et al. [2] presented about how advancement in the task of the recognizing objects and machine translation has improved the Image Captioning Model. They have presented the implementation of the methodology using Deep Neural Network Algorithms such CNN. The presented methodology and its performance was evaluated using various standard evaluation matrices.

Mingxing Zhang et al. [3] presented on the Deficiency of insufficient concepts in Image Captioning using traditional methods for Captioning. They explained the reasons for the the problem such as the imbalance between the number of occurrence of positive and the negative samples of the concept. The another reason they highlighted is Incomplete Labeling in Training Captions. They proposed a Model called Online Positive Recall and Missing Concepts Mining to overcome the problem. The method re-weights the loss of different samples based on their predictions and used a two-stage optimization methodology for mining missing concepts. The proposed methodology gave a high accuracy Captions and was able to detect more semantics.

Chetan Amritkar et al. [4] presented on how the contents of an image can be generated using the Computer Vision and Natural language Processing (NLP). The proposed model used CNN and RNN to generate Captions. The Models used CNN algorithm for extracting Features from the Image and RNN further Generated Sentences based on the Features extracted.

Aghasi Poghosyan et al. [5] presented the most fundamental problem of the existing Image Captioning Models where the next word to be predicted in the captioning process depends on the last predicted word rather than Image Content. The proposed model which generated Image Description and used RNN with modified additional LSTM cell gate for Image Features which generated more accurate captions.

VishwashBatra et al. [6] proposed a methodology which focused on News images and generating captions automatically for news paper articles which is different from the traditional Methods due to the input given to the system not only contained Images but also Text paragraphs. They used several deep neural network architecture such as RNN. The Results shown are more accurate than other traditional models as additional Text Descriptions were used to describe the Image .

Jie Wu et al. [7] discussed the currents methods for Image captioning as they as the caption generated are composed of most frequently used words which leads to Caption generation. They Proposed a new Method including Content Sensitive and Global Discriminative which generated more concrete and discriminative captions. The Content Sensitive method focused on less frequent and more concrete words and phrases which described the image content better. They further used the Global Discriminative methods which pulled the generated sentence better related image than other methods.

Min Yang et al. [8] presented “MLADIC” algorithm for crossdomain Image Captioning. They explained the steps to reduce the gap between different domains such as source and target domains. They first pre-trained the Model to learn the alignment between the images and text data . Then the model was tune the learned model with limited image and text pairs and unpaired information in the target domain.

Proposed Methodology

Caption Generation Model aims towards generating simple and informative captions for the input Image. The Model formulates the given problem as follows: Given Input Image *I,* generate a caption that best describes the Image. The Training Data set Dconsists of the Image-Caption Data.

A. Architecture

The Proposed System shown in Fig. 1 first takes Input Images which are then passed to the Encoder Module (CNN) which uses it Convolutional and Pooling layer to generate the feature Vector. After Every Convolutional layer , a Pooling layer will be used. In this Model, the last layer of CNN which is Fully Connected Network is not used as we are only generating the Feature Vector of the Image and not classifying the Image. Once the Image Vector is generated, it is given to the Next Module, Decoder Module (LSTM) a special Recurrent Neural Network (RNN) algorithm which uses the feature vector as Input and has a “memory cell” that can hold information in the memory for long period of time which generates Sequential Sentences from the Vector. At last, the Accurate caption describing the Image is selected as Output Image Caption. The Encoder/ Decoder Model which first store the objects, features into feature vector such Color, action, objects , size, etc and then the decoder module forms a sentence based on the vector provided by the encoder module.

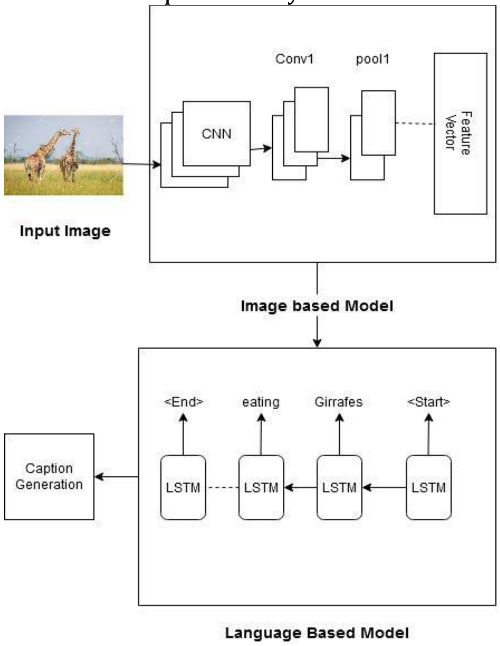

Fig. 1. System Architecture

B. Modules

Module 1: Image Pre-processing

As machine does not understand Images. The Image needs to be stored in the matrix of pixels and color code of each pixel at the respective locations. The Pre-processing would be the first task in the proposed system where input image is given as input and is transformed into matrix of pixels. In this Module the Noise from the Image is removed.

Module 2: Image based Model (CNN)

A Pre-trained CNN extracts the features from the input Image in the form feature vector by using various CNN layers. After Every Convolutional layer , a pooling layer is used to extracts features such Color of the Objects, Objects, Dimensionality Reduction, etc.The feature vector is linearly transformed to have same dimension as the input dimension of LSTM network. This model is known as the Encoder . **Module 3: Language Based Module (LSTM)**

The Decoder module translates the features and objects given by the previous Image based Module to natural sentence i.e. Caption based on the Image using Long Short Term Memory (LSTM). The system is trained as Language model on the feature vector .The Decoder Module generates sequential Sentences based on the features .For training The LSTM Model, we have to predefine our label and target text.

Example: Consider the caption is ”X and Y are drinking coffee”.

•Label: [start,X,and,Y,are,drinking,coffee,.]

•Target: [X,and,Y,are,drinking,coffee,end]

Module 4 : Caption Generation and comparative Result analysis

This Module’s important aim is generating caption for given input image. System will be given test images which should generate appropriate Results.

C. Algorithms

I. Convolutional Neural Network (CNN):

Following are the Layers in CNN Architecture:

The Convolutional and Pooling Layer behaves as the Features Extractors from the input Image.

1. Convolutional Layer:

The first layer in a CNN is a Convolutional Layer. The input to this layer is a array of pixel values. Each picture/image can be considered as a matrix of pixel values. Pixel values range from 0 to 255 for gray scale image. In CNN , the nbyn matrix called a filter or kernel or feature detector is formed by sliding the filter over the input image and it computes the dot product which is called as the Convolved Feature or the Feature map or Activation Map.The filters behave as the feature detectors or Highlight features from the input original Image. CNN model learns the estimated values of these filters on its own as self learning during the training process. The more number of filters we apply, the more the number of features from the images get extracted and identified, it helps our neural network to improve and learn better in identifying the features from the given Image as input. The network becomes good at recognizing pattern in the unseen new images.

The size of the Feature Map (Convolved Feature) is controlled by three parameters that we need to decide before the convolution step is performed:

Depth: Depth corresponds to the number/quantity of filters we use for the convolution operation on the image.

Stride: Stride is the quantity or number of pixels by which we slide our matrix which is called as filter matrix over the input matrix. When the value of stride is 1 then we shift the filter one pixel at a time.

Zero-padding: Sometimes,it is helpful to pad the input matrix with zeros 0’s around the border , so we can apply the filter to the borders or bordering elements of our input image matrix.

Non Linearity (ReLU): An operation is added called ReLU which has been used after every Convolution operation. ReLU (Rectified Linear Unit) is a nonlinear operation.The output is given by: Output Max(zero,Input). ReLU is an element wise operation (applied per pixel) which replaces all the negative pixel values in the feature map by zero. The main purpose of ReLU operation is to introduce nonlinearity in ConvNet network .

2. Pooling Layer

The Pooling layer reduces the dimensionality of each Features in the feature map which retains the important features and information. Pooling can be described into following types : Max, Min, Sum and Average. In Max Pooling, we define a Spatial neighborhood (2 X 2 window) and select the largest element from the feature map inside that window. The main Function of the Pooling Layer is to reduce the size of the input matrix which helps for computing faster in the network and also reduces the number of parameters . It prevents Over fitting.

The CNN Algorithm is as follows:

```
Input: Pixel Matrix of Image I to the System S.
Output: Feature Vector Generation, Fv .
BEGIN
Initialize the filters and weights required with random values.
for each Image I:
Step 1: Input the 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: After every convolutional layer apply ReLU operation.
Step 4: Apply pooling layer for each feature in Feature Map for dimensionality reduction.
Step 5: Feature vector Fv of Image I.
end for
END
```

When the network model is feed with unseen images to the ConvNet, Forward Propogation is used to extract features from the Image. And later are stored into feature vector generated from Convolutional Neural Network CNN (Encoder) Module which is later been given as input to the Long Short Term Memory LSTM (Decoder) Module of the proposed Model.

II. Long Short Term Memory (LSTM):

The Traditional RNNs can keep track of long-term dependencies information in the input sequence. The issue of RNNs is when training the RNN using back-propagation, the gradients can vanish due to the computations. LSTM units are modified version of RNN which solves the vanishing gradient problem. Long short-term memory (LSTM) is an artificial and modified recurrent neural system .LSTM has feedback connections input .It can process images as well as the sequence of data (ex: video). A typical LSTM system consists of a cell, a input gate, output gate and forget gate. The cell has the functionality to remember values over arbitrary time intervals and the three gates which controls the flow of the information inside or outside the system model.

We have additional piece of information which is called MEMORY in LSTM for each time step. The LSTM cell shown in Fig. 2 contains :

1. Forget Gate “f” (neural system with sigmoid)
2. Candidate layer “C”(NN with Tanh)
3. Input Gate “I”(NN with sigmoid)
4. Output Gate “O”(NN with sigmoid)
5. Hidden state “H” (vector) 6. Memory state (vector)

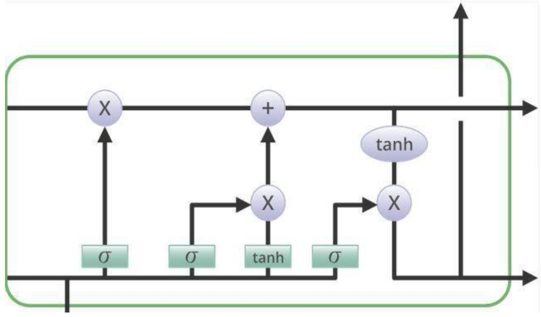

Fig. 2. LSTM Cell

The LSTM generates the Caption C,
$\boldsymbol{C = \{ w_{-1}, w_0, w_1, \dots, w_L \}}$ where
$\boldsymbol{w_{-1}}$ = Start token,
$\boldsymbol{w_L}$ = End token,
$\boldsymbol{L}$ = Length of the Caption **.**

The LSTM takes the forward propagation at time step t using equation as shown below : $\boldsymbol{it = \sigma\,(Wi\,[ht\text{-}1\,,\,xt] + bi\,),\, ft = \sigma\,(Wf\,[ht\text{-}1\,,\,xt] + bf\,),\, ot = \sigma\,(Wo\,[ht\text{-}1\,,\,xt] + bo\,),\, gt = tanh\,(Wg\,[ht\text{-}1\,,\,xt] + bg),\, ct = f\,t \odot ct\text{-}1 + it \odot gt\,,\, h_t = o_t \odot tanh\,(\,c_t\,)}$ **.**

At the time step t, an LSTM with two inputs, the input vector ($\boldsymbol{x_t}$) at the given time step and the hidden state vector ($\boldsymbol{h_{t-1}}$) of the previous time step. The W are the weight matrices and b are the biases. In the Forward Propagation, the above process is used to update in the input gate, forget gate and the output gate.

Experimental Setup

Data Set:
We have based our model experiments on Flickr8k data set. The Model can be effectively trained using this data set. Each Image has 5 captions provided with it. This data set is suitable for Desktop and laptops. The Data set has to files shown as below:

1. Flickr8k_Dataset: It has total 8092 images with different sizes , shapes and colors. Out of total 8092, 6000 images are used for training the model and rest for the testing the model.
2. Flickr8k_text: It has text file describing the Training and Test data. Flickr8k.token.txt has 5 captions per image for training the model stored in the key- value pair where key is the Image id and the value is the List of Captions.

We have considered an input of images with size 200X200X3 pixels. The images are input to the Convolutional Neural Network .

Results and Discussion

A Pre-processed Model is built on the Data set. The Image is taken as input as shown in Fig. 4. Every Image goes through the Pre-processing Model which includes with Grayscale, Threshold and Edge Detection as shown in Fig. 5. The input image is converted into Grayscale, which is further gone under the Threshold and Edge Detection Process in order to remove the noise from the Image and the machine model is trained with variation of images so that the model is able to extract features even if unseen Image is feed.

Fig. 5. Shows the process of how every image goes under various transitions so that the machine is able to adapt the process of identifying an image better even if there is slight changes in any image or its object. It helps machine to adapt the skill of recognizing the objects or scene which reduces the error rate in caption generation.

The Fig. 6 shows the Image objects which are detected along with the accuracy of the image. The Pre-tained CNN Model identifies the objects from the images which are the feature vector for the next Model which is LSTM.

The Objects in the Fig. 6 are correctly identified using CNN with Convolutional layer and Pooling Layer.

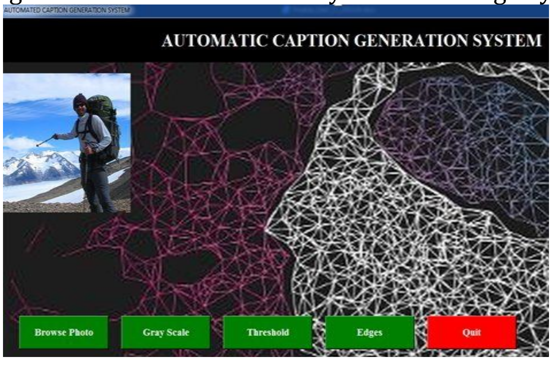
AUTOMATED CAPTION GENERATION SYSTEM
AUTOMATIC CAPTION GENERATION SYSTEM
Browse Photo
Gray Scale
Threshold
Edges
Quit

Fig. 4. Input Image

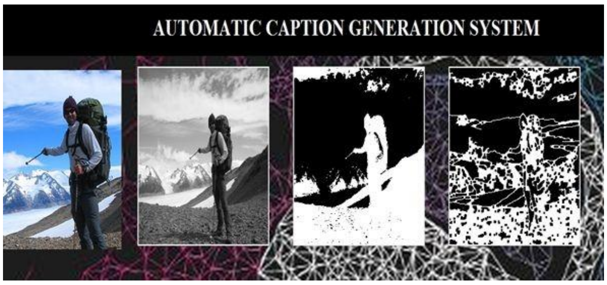

Fig. 5. Pre-Processed Image

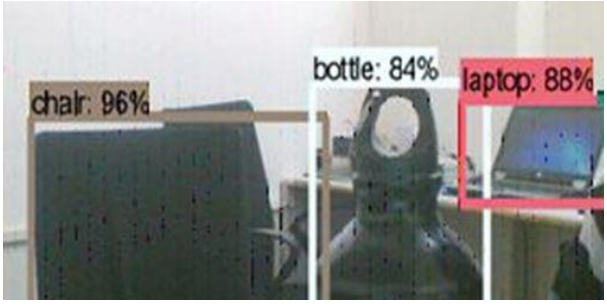

Fig. 6. Object Identification in Feature Generation

Conclusion

In this paper, we have introduced CNN Algorithm and LSTM Algorithm to generate automatic Image Captioning. The Model uses CNN algorithm to extract the Features from the Image in the form of Feature Vector. The features are stored in the form of vector which are further passed to the LSTM algorithm which is advanced RNN to form sequential sentences from the vectors by maintaining the relationship between the objects. The approach helps to reduce the error rate and the RNN vanishing gradient problem is solved using LSTM. The Captioning model can efficiently learn the Image-Caption Pairs and produce significant quality result which a precise caption for the Image. In Future, this model can be extended to a architecture which uses long descriptions along with the Images as inputs to the model to generate captions .

Acknowledgement

The authors would like to thank researchers as well as the publishers for making their resource available and also teachers for their guidance. We are thankful to the authorities of Savitribai Phule University, Pune and concern members of cPGCON2020 conference, organized by the Department of Computer Engineering, Sangamner for their constant guidelines and support. We are also thankful to the reviewers for their valuable suggestions.

References

[1] Kun Fu , Jin Li, Junqi Jin, and Changshui Zhang, Fellow , "Image-Text Surgery: Efficient Concept Learning in Image Captioning by Generating Pseudopairs", IEEE Trans, 2162-237X, 2018 .

[2] Parth Shah, Vishvajit Bakrola, Supriya Pati,"Image Captioning using Deep Neural Architectures",IEEE International Conference on Innovations in information Embedded and Communication Systems (ICIIECS), 2017 .

[3] Mingxing Zhang, Yang Yang, Hanwang Zhang, Yanli Ji, Heng Tao Shen, Tat-Seng Chua , "More is Better: Precise and Detailed Image Captioning using Online Positive Recall and Missing Concepts Mining",IEEE Transactions on image processing,2018 .

[4] Chetan Amritkar, Vaishali Jabade," Image Caption Generation using Deep Learning Technique", IEEE 978-1-53865257-2/18/$31.00 , 2018 .

[5] Aghasi Poghosyan , Hakob Sarukhanyan, "Long ShortTerm Memory with Read-only Unit in Neural Image Caption Generator ",IEEE, 978-1-5386-2830-0/17/$31.00,2017 .

[6] Vishwash Batra, Yulan He, George Vogiatzis,"Neural Caption Generation for News Images", School of Engineering and Applied Science, Aston University,2018.

[7] Jie Wu, Tianshui Chen, Hefeng Wu, Zhi Yang1, Qing Wang, and Liang Lin"Concrete Image Captioning By Integrating Content Sensitive And Global Discrimination Objective", in IEEE International Conference on Multimedia and Expo (ICME) 2019.

[8] Multitask Learning for Cross-domain Image Captioning Min Yang, Wei Zhao, Wei Xu, Yabing Feng, Zhou Zhao, Xiaojun Chen, Kai Lei , IEEE transactions on multimedia,2018 .