International Journal of Future Generation Communication and Networking
Vol. 13, No. 3, (2020), pp. 3175–3183

Automatic Image Captioning Using Neural Network

Ms. Priyanka S Raut[1], Mrs. Rushali A Deshmukh[2]

[12] Dept. of Computer Engineering JSPM's Rajarshi Shahu College of Engineering
Tathawade.Pune – 411033., India

[1] priyankaraut940@gmail.com, [2] radesh19@gmail.com

Abstract

The Automatic Image Captioning aims towards generating natural and simple language sentences to describe the content of an Image. Image Captioning has gained a lot of interest in recent times. Generating accurate captions is a very challenging task. Although various solutions have been generated using the traditional CNN and RNN neural network algorithm for Image Captioning, still a problem arises is that the model can only generate Captions for the Image Concepts, which are seen in the Training Data set. In this paper, we propose a 2 Stage Model Architecture methodology with Deep Convolutional Network, which is known as the CNN algorithm for extracting the features and Long Short Term Memory (LSTM) algorithm to generate accurate Image Caption and to overcome the problems arisen in the traditional neural algorithms methodology. We evaluate our approach on the Flicker8k data set.

Keywords: *Image Based Model, CNN, Feature Vector, Language Based Model, RNN, LSTM, Caption Generation.*

1. Introduction

The Automatic Generation of Image Captions is such a task in which the Machine learns and understand the visual world or images at a human level of knowledge and understanding of describing the content of Image. The machine Model is able to generate accurate Natural language Captions or sentences for the given Input Image. The model is an example of Supervised Learning where a Machine Model is trained with Images along with its Descriptions. However, in such tasks like Image Classification is used to classify the correct Objects in the Image. This is a very challenging task as we expect the machine to understand a very complex scene. Image Captioning aims to generate simple explanatory sentences describing the objects and their relationship with the given Image or Visual Description Generation. The automatic Image Captioning Model is a two-stage Model which is an Image Based Model as stage-I and Language Based Model as Stage-II. The First stage uses the Deep Convolutional Network for Extraction of Features using as many layers until accurate Image Features are extracted. In this Model, Image Features are the Image Objects, which is an important part of the Image caption. The Second Stage uses the Long Short Term Memory known as LSTM, which is a better version of the Recurrent NN algorithm for generating sequential sentences from the Features extracted from the Stage-I. This stage focuses on generating the Relationship between the objects. The Stage -II generates Precise captions. The model uses LSTM to overcome the problems arisen in the traditional Image Captioning Model which uses Recurrent Neural Network (RNN). The RNN algorithm suffers from vanishing and exploding gradient problem. The LSTM algorithm solves the gradient problem by using new gates, which maintains the information in the memory for long dependencies.

ISSN: 2233-7857 IJFGCN

2. Literature Survey

Kun Fu et al. [1] proposed a novel method, Image-Text Surgery for Image Captioning, which used and synthesized the pseudo image and sentence pairs. The paper depicts the efficient learning concepts in Image Caption Generation by using pseudo pairs. This pseudo image and sentence pairs were generated with the knowledge base, which used the seed data set syntax. The model system learns the concepts without any human image-sentence pairs, which are labeled pairs. They also introduced the adaptive replacement of visuals that filtered the pseudo data's unnecessary features with the above mechanism. They evaluated their experiments against the MSCOCO Data set. Their output had shown significant improvements in the caption quality. Their approach is robust to pseudo-data, which shows that stable performances in a broader range are different from the traditional Image Captioning. The paper also shows the construction of the knowledge base and the process to measure the Concept Similarity. Once this is done, the paper shows the generation of the Pseudo sentences.

Parth Shah et al. [2] presented about how advancement in the task of 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 as CNN. The CNN algorithm is explained and is used to generate captions. The presented methodology and its performance were 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 problem, such as the variance between the number of occurrences of positive and the negative concept samples. Another reason they highlighted is Incomplete Labeling in Training Captions. They proposed a Model called Online Positive Recall and the Missing of Concept Mining, which helped to overcome the problem. The re-weighting of the loss of different test samples based on their generated predictions and 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. The Captioning of Images is a very challenging and difficult task that requires both Image and Text Processing. The disadvantage of a Recurrent Neural Network is that it has a vanishing gradient issue.

Aghasi Poghosyan et al. [5] presented the most fundamental problem of the existing Image Captioning Models. The next word to be predicted in the captioning process depends on the last predicted word than Image Content. The proposed model has generated Image Description and used RNN with modified additional LSTM cell gate for Image Features, which generated more accurate captions. In this paper, how the next word is identified from the previous passed input word is depicted.

VishwashBatra et al. [6] proposed a methodology that focused on News images and generating captions automatically for newspaper articles, which is different from the traditional methods due to the input given to the system contained not only 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. This paper uses additional information, which are long descriptions of the image, which increase the word vocabulary. The word vocabulary helps to predict the objects in the image very easily. The description can be used as a hint to find out correct Image content. If a building is given in the image, it gets

ISSN: 2233-7857 IJFGCN

difficult to predict if it a government office or an old school. The additional description plays a vital role in predicting the exact Image Content. The paper shows how the description has played a vital role in Image Captioning. They also explained how the model could be useful for Online new Sites such as BBC, etc. They have conducted their experiment on the BBC News Corpus.

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 that better described the image content. They further used the Global Discriminative methods, which pulled the generated sentence better-related image than other methods. The model described in this paper uses two additional concepts for better caption generation. The first concept is to use the less frequently used words such An old Man is a less frequent word, and the Man is a frequently used word. Such less frequently used words can help to describe the image in a better way and accurate captions. The next concept is to pull the image closer, which best describes the sentence and pushes away all other images from the provided set. This paper shows better results than the traditional Image Captioning Model.

Min Yang et al. [8] presented the "MLADIC" algorithm for the cross-domain Captioning of Image. They further explained the steps to reduce the gap between different domains such as the source and the target. They then trained the system Model to learn and understand the alignment of the images and their text data. Then the system model was fed with the selected image and its text and information not paired in the target.

3. Proposed System

The Automatic Image Caption Generation Model aims to identify objects, their actions, their relationship, the background location, and generate meaningful Captions for a given input image. The model formulates the given problem as follows: Given input image I, generate a caption C that best describes the Image content. The Training Data set D consists of the Image-Caption Data.

3.1. Architecture

The Proposed Methodology shown in Figure 1, first takes Input Images and converts it to the Image vectors of a fixed size, which are then given to the network i.e., Encoder Module (CNN). The convolutional layer and Pooling layer generates the feature Vector. The last layer of CNN is a Fully Connected Layer. It is removed from our Proposed methodology as we are only generating the Feature Vector of the Image and not classifying it. After the ConvNet layer, a Pooling layer will be used. Our Encoder Module takes an input Image of size 299X299X3, and the output Feature Vector is of size 8x8x2048. Once the Image Feature Vector is generated, it is given to the Next Module, Decoder Module (LSTM) a special version of Recurrent Neural Network (RNN) which sequentially processes input feature vector and also has a "memory cell" which can hold information in the memory for a long period of time which generates Sequential Sentences from the Feature Vector. The last output is the caption generated to describe the image. The Encoder/ Decoder Model, which first identifies the objects, features into feature vector such as color, action, objects, size, etc, and then the decoder module forms a sentence based on the vector provided by the encoder module.

ISSN: 2233-7857 IJFGCN

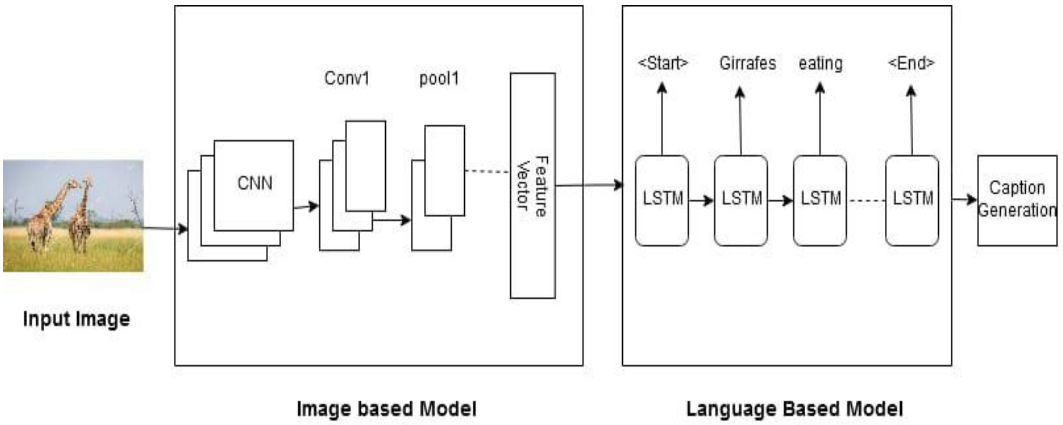

Figure 1. System Architecture

3.2. Modules

3.2.1. The Image Pre-processing

As the machine model does not understand the Images. The image needs to be converted into the matrix of the pixels and the color code of each image pixel at their respective locations. Pre-processing is the first task performed in the proposed Model where Input image with the various size is provided as input, which is transformed into a matrix of pixels around 299X299x3 size. In this module, the noise from the image is removed.

3.2.2. The Image based Model (CNN)

A Pre-trained CNN takes a matrix of Image Vector as input and extracts features from the input image in the form of a fixed size feature vector, which is 8x8x2048 by using CNN layers i.e., ConvNet layer and pooling layer along with ReLU. The Convolutional layer and pooling layer extracts feature such as Color of the Objects, Objects, which are stored where each word is given a unique index number. Dimensionality Reduction is also performed. The vector of image features is transformed in a linear way to have similar dimensions as the LSTM network input dimension. The model is also called the Encoder Model.

3.2.3. The Language Based Module (LSTM)

The Decoder module translates the features provided by the Image based Module to generate natural sentence i.e., caption using Long Short Term Memory (LSTM). The LSTM algorithm uses tokens for sequencing. The Decoder Module generates sequential Sentences based on the features. To train this model, we first have to define our label and its target tokens i.e. Text.

Example: Consider the image caption = ”A and B are eating cake”.

Label: [start,A,and,B,are,eating,cake,.]

Target: [A,and,B,are,eating,cake,end].

3.2.4. Caption Generation and comparative Result analysis

ISSN: 2233-7857 IJFGCN

The Caption Generation Module generates Image Caption which is the output of LSTM model. The AI System will be given a test images which should generate accurate captions.

4. Algorithm

4.1. The Convolutional Neural Network (CNN)

The Convolutional Layer and Pooling Layer are used as the Features Extractors from the Input Image.

4.1.1. Convolutional Layer

The first layer in CNN is a Convolutional Layer. The input to this layer is an array of pixel values of size 299x299x3. Each picture is considered as a matrix of pixel values. Pixel values range from 0 to 255 for a grayscale image. In CNN, the nbyn matrix known as a detector of feature is generated by using the filter slide and sliding it over the input picture, and it computes the dot product of it, which is known as the Activation Map. These filters behave as the feature detectors or Highlight features from the original input 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 features from the images are extracted, which helps our neural network improve and learn better in finding features from the given image. The network keeps becoming good at finding patterns in the new input pictures.

The Map size can be controlled with the below parameters that we need to pre-define before performing this step of convolution:

Depth: Depth is the number of filters that we use for the above defined convolution step/operations on the given image.

Stride: Stride is the quantity or number of pixels by which we slide our matrix called a matrix of filter over the given matrix. If the value of stride = 1, then shift the filter by one pixel.

Zero-padding: In Some cases, we need to pad the given matrix border with zeros, so that it becomes easier to apply the selected filter to the input image matrix borders.

Non Linearity (ReLU): After every Convolution operation, a ReLU operation is added. The Rectified Linear Unit is a nonlinear step. The output of this operation is given by Output Max (zero input). This operation is applied on a single image pixel, which replaces the negative image pixel in the map by 0 i.e. zero.

4.1.2. Pooling Layer

The Pooling layer is used for dimensionality reduction of features in the map, which does not hamper the important features and its data. This layer can be described in the following: Max, Min, Sum, and Average. The Pooling Layer reduces the size of the given matrix, which helps for processing faster in the neural network and also reduces the number of parameters that are not required. It avoids Overfitting of the data in the model. The CNN algorithm reshapes the loaded image into the fixed size Image pixel values i.e. 8x8x2048. The function extract_features() loads each photo from the given directory, processes it and then extracts the features. The features of the image are One-D with 4,096 element vector. The function gives a collection of the picture identifier to the features of the picture.

ISSN: 2233-7857 IJFGCN

The CNN Algorithm is stated as below:

Input: Pixel Matrix(299x299x3) of Image I to the System S.

Output: Feature Vector Generation, Fv of fixed size.

BEGIN

Initialize the required filters and weights required with random values and conv_2d layer with activation layer.

for each Image I in System S:

Step 1: Input the Image pixel matrix to the stage-1 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, perform Image processing.

Step 3: After every convNet layer, apply ReLU operation.

Step 4: Apply pooling layer for each feature in Feature Map for Dimensionality reduction.

Step 5: Fixed sized Feature vector Fv of Image I.

end for

END

When the network model is feed with unseen images to the ConvNet, Forward Propagation is used to extract features from the image. Furthermore, later are stored into feature vector generated from the Convolutional Neural Network CNN (Encoder) Module, which is later being given as input to the Long Short Term Memory LSTM (Decoder) Module of the proposed Model. The output of this model is not actual English words or text; rather it is the sequence of the indices which is unique for each word in the Dictionary.

4.2. Long Short Term Memory (LSTM)

The Traditional Recurrent Neural Network (RNN) can track 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 versions of RNN, which solves the problem of gradient vanishing. Long short-term memory an artificial and modified recurrent neural system 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, an input, output, and a forget gate. This cell has the functionality to remember data over the long time intervals and these gates, which controls the data flow inside or outside the system model. Partial caption is passed to the cell, which predicts the next word in the sequence. It uses the tokenizing method as it adds the start and end token to the input features.

We have an additional piece of information which is called MEMORY in LSTM for each time step. The LSTM cell consists of:

1. Forget Gate “f” (neural system with sigmoid)

2. Candidate layer “C”(NN with Tanh)

ISSN: 2233-7857 IJFGCN

3. Input Gate “I” (NN with sigmoid)

4. Output Gate “O” (NN with sigmoid)

5. Hidden state “H” (vector)

6. Memory state (vector)

The Model reads these partial caption to process the sequence of the words in the dictionary of words.

5. Experimental Setup

5.1. Data Set

We have based our model experiments on the Flickr8k data set. The model can be effectively trained using the Flickr8k data set. The data set has provided two files, one with Images and the other with Image reference and five captions associated with the image. This data set is suitable for small workstations such as Desktop and laptops. It contains the two files described below:

5.1.1. Flickr8k_Dataset

It has a total of 8092 images with different sizes, shapes, and colors. Out of total 8092, 6000 pictures are used to train the model, and 1000 pictures are used for Development, and the remaining 1k images are used for testing the model.

5.1.2. Flickr8k_text

This text file describes which images are used as Training and Test data. It contains a Flickr8k.token.txt which has five captions per image for training the model stored in the form of key-value pair where the key is the Image id, and the value is the List of Captions.

We have considered an input of images with size 299X299X3 pixels. The images are input to the Convolutional Neural Network.

6. Results and Discussion

A Model is built on the defined Data set, which is the Flickr8k data set. The image is given as input, as shown in Figure. 2. Every Image is the Pre-processed Model, which includes Grayscale, Threshold, and Edge Detection as shown in Figure 2. The input image is converted into Grayscale using conversion_to_grayscale function, which is further gone under the Threshold where the foreground and background images are separated by white and black color and Edge Detection Process in order to remove the noise from the image, the machine model is trained with a variation of images so that the model is able to extract features even if unseen image is feed. At every iteration, the Model learns from its training Images. The Figure. 2 shows that each image needs to processed before it is been given to the Model for generating captions. The Figure 3 Shows the objects in the form of extracted features from a given input picture, which is the outcome of the Convolutional Network algorithm in the Image Based Model. The output shows the percentage of the correctness of the identified object from the provided Image. These Objects are known as the Features, which are stored in the vector Form. Each word/object identified is added to the vocabulary, which are then replaced with a unique index in the model.

ISSN: 2233-7857 IJFGCN

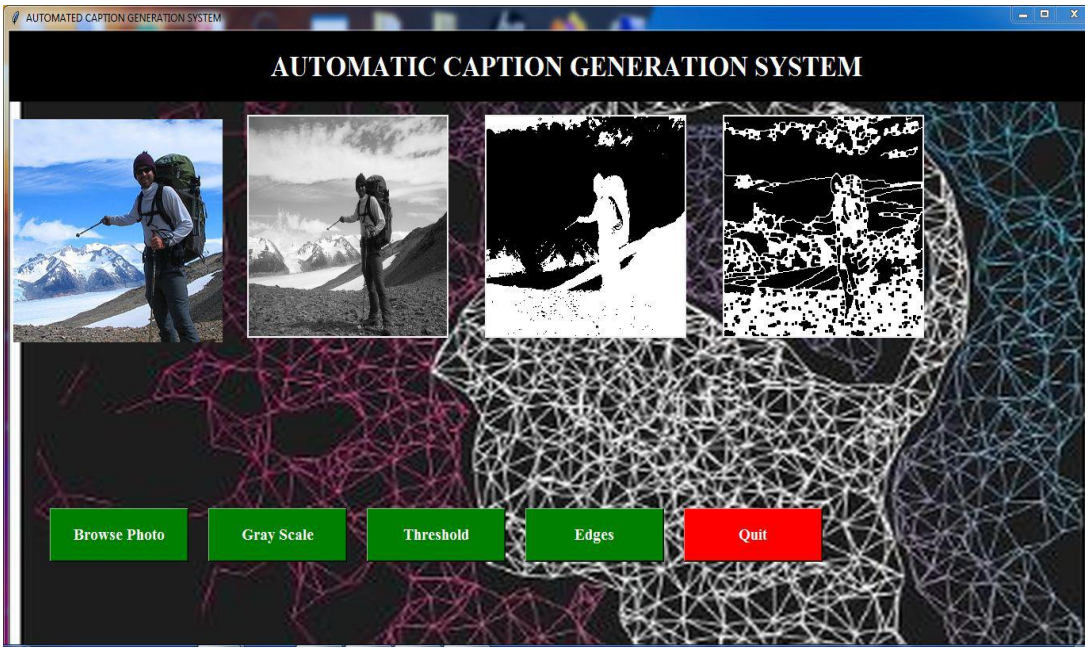

Figure 2. Image Pre-processing

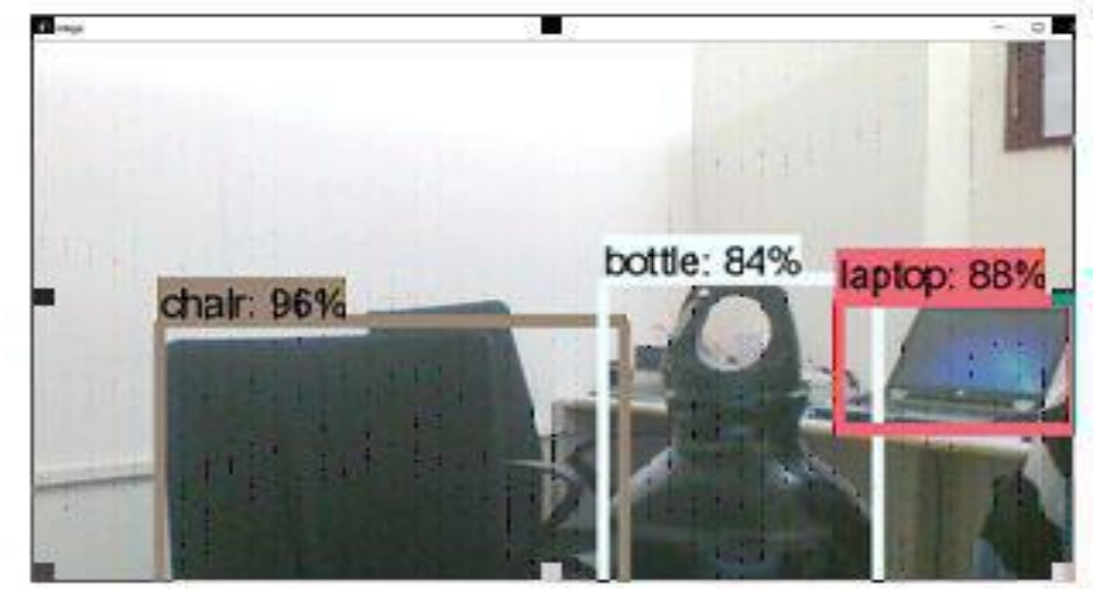

Figure 3. Object Identification in Feature Extraction

7. Conclusion

In this paper, we proposed a methodology with the Convolutional Network and an advanced Long short term memory Algorithm to Generate the best image describing Sentences and Caption. The Supervised Model was trained with 6000 Training Image with it's respective Captions using Convolutional layers to extract the Image Features, which were then stored in the form of Feature Vector. The Image features are stored in the form of a vector which are then further given to the next model, which uses the LSTM

ISSN: 2233-7857 IJFGCN

algorithm, an advanced Recurrent Neural network to process the sequential sentences from the Image vectors. The Language based Model generates the natural language sentences which justifies the Image Content by maintaining the Object's Relationship with each other correctly. The proposed approach reduces the error rate in the generated caption and also solves the RNN problem of gradient disappearing using an advanced version of Recurrent Network. The Artificial Intelligence Captioning model can efficiently learn the Image-Caption Pairs provided in the Training data set and produce significant quality captions that are accurate. In the future, this model can be extended to an architecture fed with long descriptions and Images as inputs to the AI Model to generate short and more accurate Captions.

Acknowledgments

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 ICATCSIT-2020 conference, organized by the JSPM's Rajarshi Shahu College of Engineering, Department of Computer Engineering,Tathawade ,Pune 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 978-1-5386-5257-2/18/$31.00 , **2018**.
[5] Aghasi Poghosyan,Hakob Sarukhanyan,"Long Short Term 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", IEEE International Conference on Multimedia and Expo (ICME) **2019**.
[8] Min Yang, Wei Zhao, Wei Xu, Yabing Feng, Zhou Zhao, Xiaojun Chen, Kai Lei"Multitask Learning for Cross-domain Image Captioning", IEEE Transactions on Multi Media, **2018**.
[9] K. Simonyan, A. Zisserman,"Very Deep Convolutional Networks for Large-Scale Image Recognition", International Conference on Learning Representations, **2015**.

ISSN: 2233-7857 IJFGCN