JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

DEVELOPING AN VOICE BASED ADVERTISEMENT GENERATOR FROM IMAGES USING CNN& LSTM

[1]BIKKI SANDYA, [2]Dr.I. KULLAYAMMA

[1] M. Tech Student, Department of Electronics and Communication Engineering, SV University College of Engineering, SV University-Tirupati, Andhra Pradesh, India.

[2] Professor, Department of Electronics and Communication Engineering, SV University College of Engineering, SV University – Tirupati, Andhra Pradesh, India.

Abstract: The century is extensively dependent on image-based information processing. Image processing can be described as the process of manipulating an image to either improve its quality or extract valuable information from it. In lines with that, image captioning is an increasingly emerging technique used in a variety of applications such as usage in virtual assistants, image indexing, visually impaired persons, social media and several other natural processing applications.

The project aims to develop a targeted voice-based advertisements from images using CNN and LSTM, with using MATLAB tools. For an efficient image caption generator, the system needs to identify the information and should further generate the most relevant and brief description for a syntactically and semantically correct image. An image caption generation is a task that involves the NLP (natural language processing) concept for understanding the description of an image. The main target ofthe proposed project is to obtain ads for specific images by combining CNN and LSTM. First, the description of the image is obtained. Then based on the information in the captions, specific ads are obtained. After obtaining the description, it will be converted into text and the text into a voice. Image description is a best solution used for a visually impaired people who are unable to comprehend visuals. The project could have great impact in social media platforms where images are the primary source of data.

IndexTerms **- NLP (natural language processing), CNN (Convolutional neural network), LSTM (Long short-term memory), RNN (recurrent neural network).**

I. INTRODUCTION

In recent years Deep learning is one of the most used trends in Machine Learning and artificial intelligence, it is a machine learning Technique inspired by the Human brain, it uses the algorithm like convolutional neural network, recurrent neural network, long short term memory etc., where there are many developments had already made for visually impaired people, voice based Image caption generator is used to identify the objects and information present in the image, which could improve the lives of Visually impaired people, Using CNN and LSTM together can be best fit for this project because LSTM is similar to RNN, and the RNN algorithm is depending on the LSTM because its having the feedback connectivity and also LSTM process the entire sequence of data.

The main challenge of deep learning is when we deal with large data we need to go deeper that is analyzing the huge data need to done thoroughly, The structure of text descriptions should be relevant to the objects present in the image,

and the relationship between the objects and it's descriptions need to be clarified, Our ultimate aim of the project is to train the dataset with the good result and with the high accuracy. Val dataset is utilized with the huge collection of photographs used for computer vision and image processing algorithms. So this voice based caption generator act as a eyes for the people don't have the ability to conceptualize the scene happen around themselves, they can roam anywhere without the support of anyone else.

II. LITERATURE REVIEW

[1]. **Neha Tuniya, Shariva Dhekane, Vaishnavi Agrawal, Vibha Vyas(2021),"Image Based Caption Generator using Attention Mechanism"**-12th International Conference on computing communication and Networking Technologies,2018. This introduced a system that uses an attention mechanism alongside an encoder and a decoder to generate the captions. It used a pre-trained Convolutional Neural Network (CNN) viz. Using Inception V3 for image feature extraction followed by a Recurrent Neural Network (RNN). GRU to generate a relevant caption. To generate captions, the model used an attention mechanism that is Bahdanau attention. MS COCO dataset is used to train the model. The results validated the model's reasonable ability to understand images and generate text.

[2]. **Mirza Muhammad Ali Baig, Mian Ihtisham Shah, Muhammad Abdullah Wajahat, Nauman Zafar and Omar Arif (2018)," Image Caption Generator with Novel Object Injection"**-The data set usually used is MSCOCO (Microsoft: Common Objects in Context) data set. This covers about 80 object classes, which is an insufficient amount for creating robust solutions that aren't limited to constraints of the data at hand. To overcome the problem, they proposed the solution of identifying unknown objects and classes by using semantic word embeddings and existing state-of-the-art object identificationon algorithms. This model used a pre-trained caption generator and works on output ofthe generator to inject objects that are not present in the data-set into the caption.

[3]. **Sumathi, T., and Hemalatha, M., presented a combined hierarchical model for automatic image annotation and retrieval at the International Conference on Advanced Computing (ICAC) in 2011**.- In image process [3] used support vector machine and JEC to extracting the depth feature for an image by applying the Gaussian effect to get a better idea of a user given image. In [3] they used JEC for image feature extraction method. It will createa feature vector for annotation an image in various dimensions. It's simply the process of utilizing various models to process the image. After extracting the features from the JEC it applied to the SVM model to performing various operations like, rotating image in flat wise, axis wise and position wise manner to map the important features it contains keyword while annotating the image. It helpful for us to identify or recognize the object.

[4]. **Dong-Jin Kim, Donggeun Yoo, Bonggeun Sim,In So kweon,''Senetence Learning Deep Convolutional neural Network for Image Caption Generation'', In : 13th International Conference on Ubiquitous Robots and Ambient Intelligence (URAl)-2016.** And in [4] CNN and RNN algorithm is used to performing the caption generation process by including the attention model to predict the proper words LSTM is more effective when compared to RNN that's why we have used long short-term memory for our model. For recognizing the unusual objects, they have used CNN after it will pass on the obtained information of the image is fed to the first step of long short-term. To address the challenge of predicting only the initial word of a sentence, researchers have employed a technique called Guided Long Short-Term Memory (LSTM). In this method the guide carried out through the entire process, with the previously obtained word and it does not change during the process.

[5]. Varsha Kesavan, Vaidehi Muley ,Megha kolhekar, "Deep Learning based Image Caption Generation" Global Conference for Advancement in Technology (GCAT)-2019 - In [5] Based on the transfer learning approach to develop automated image captioning for user given image. Here they have using the VCG16 for encoding process. And then recurrent neural network to encode the input to produce constant dimensionalvector for getting the proper description. They used various algorithm like VCG16, RESNET and inception model to compare the accuracy that are obtained between them to use more effective one. Next appropriate caption is created for a user given image.

[6]. Ren C Luo, Yu-Ting, Hsu, Yu-Cheng, Wen, Huan-Jun, and Ye delve into "Visual Image Caption Generation for Service Robotics and Industrial Applications" in their IEEE 2019 publication.- In[6] based on the image detection, like how the face detection,object detection in the self driving machines, it detects each and every object that where present in front of the cameras and predict the proper word about the object that is box,pen,bottles,it combined with the previous pretrained model also and added the new created unknown words to the previous dataset values. By using this methodology it will become some time consuming process and results in the irrelevant description creation.

[7]. **N. Komal Kumar, K. Laxman, A. Mohan, D. Vigneswari, J. Yuvaraj (2019)," Detection and Recognition of Objects in Image Caption Generator System: A Deep Learning Approach"- 5th International Conference on Advanced Computing & Communication Systems (ICACCS),2019**-The paper addressed the problem of tedious task involved in capturing mechanism that collaborates both image processing and computer vision. The model proposed to detect, recognize and generate worthwhile captions for a given image using deep learning. The Regional Object Detector (RODe) serves for detection, recognition and caption generation. The proposed method emphasizes deep learning to enhance the current image caption generation system. Experiments had been conducted on the Flickr 8k dataset using python language to demonstrate the proposed method.

[8]. **Yu, M.T., Sein, M.M.: "Automatic image captioning system using integration of N cut and colour-based segmentation method". In: Society of Instrument and Control Engineers Annual Conference (SCCEAC)-(2011) -** In [8] entire world is focusing on the image caption generation, but no one is interested in predicting the emotions and sentiments that present in the image, for that they created the model that defines both the description and emotion that included in the image .by using CNN to extract the feature and CAST to predict emotions.

III. METHODOLOGY:

In Existing method [1] The attention mechanism introduced after the convolutional neural network encoder makes the model pay attention to the most relevant information in the input scene image so that the decoder only uses specific parts of the image to generate the caption. This improves the caption as compared to a conventional encoder-decoder-based model. The obtained captions generated by the model is intelligible.

In proposed method for image captioning, encoder-decoder is used. The visual data is encoded by deep convolutional neural network and the coded visual content is fed into Recurrent Neural Network which generates captions. The captions obtained by this method are close to the natural language and improvement in this method is capable of generating more accurate results.The Proposed methodology for voice based captions which is not only deals with internal images but also give a descriptions for external input images .once a description is created that text description will be read out as voice outputs then the audio is saved in the separate folder that contains all the audio files for the future references. For developing this model we have used convolutional neural network and long short-term memory. Convolutional neural network for indentifythe various features or objects that are present in the image. It will be helpful for the entire system predict the proper result then it will feed into the long short-term memory to produce the sequence of words that properly describe about the image.

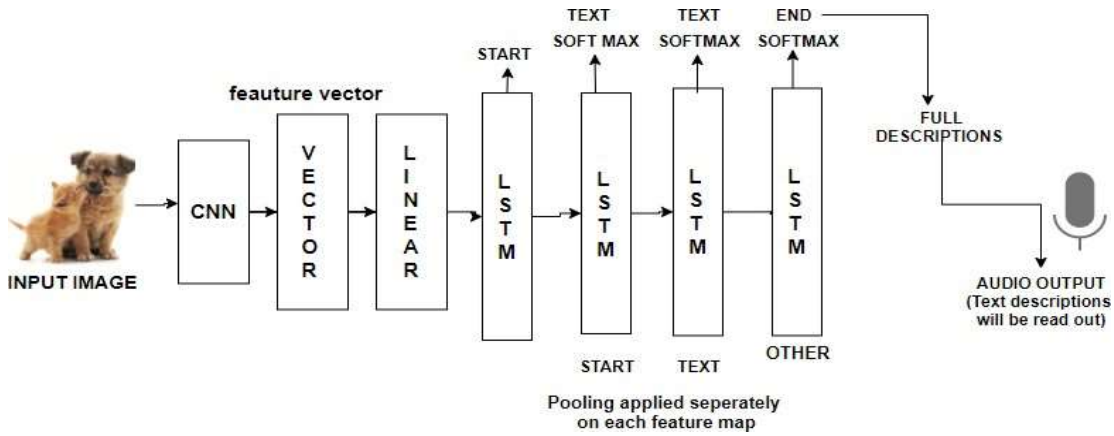


Fig1; Architecture of Proposed Model

Input image taken from the user side convolutional neural network identify the objects that present in the image it extracts the important features of an image and store those feature vector values, using pooling functions it will predict the features. Once the process completed it will move on to the long short term memory layer for the sequence sentence prediction based on the previous one, here softmax function is used to predict the output accurately and for overcoming the

over fitting problem ,when we are working with the neural network most of the nodes having the output that are related to the previous one its results in overfitting, to avoid those problem softmax layer is used .If the output of this layer is between the range of zero to one ,if the range is greater are lesser it results in the error .and system will not predict the correct description for an image.

This system integrates both CNN and RNN, where CNN is used to extract input image features and to classify these image features and these extracted feature signals are utilized by neural networks, where RNN is used to generate captions by predicting the next word consequently. Where there are many developments had already made for visually impaired people, voice-based Image caption generator is used to identify the objects and information present in the image, which could improve the lives of Visually impaired people.

Convolutional Neural Network:

CNN's employ image recognition and classification for tasks like object detection and facial recognition. They consist of neurons capable of learning weights and biases to optimize performance. Each individual neuron gets a large number ofinputs, weights them, then sends the result via an activation function to produce an output. A convolutional or CNN is a class of deep neural network it is mostly used for analyzing visual images and classification, and also it is used in various field like image recognition, NLP and speech recognition.

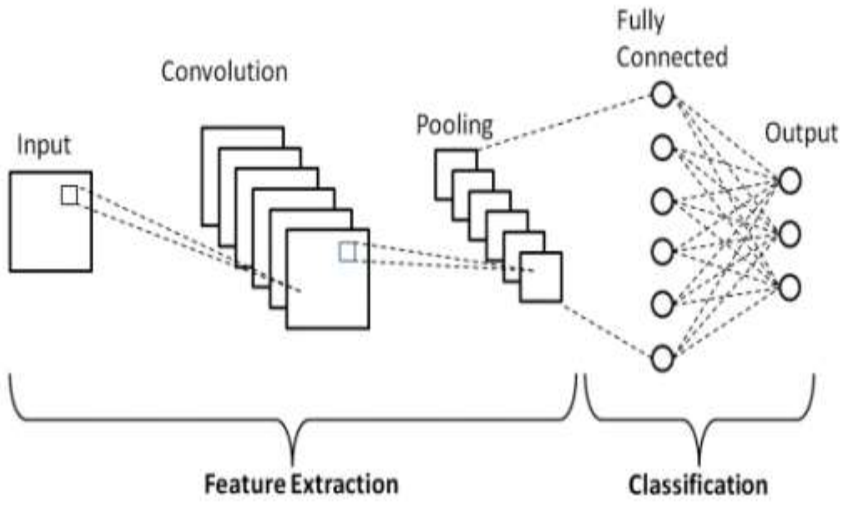


Fig 2: **Architecture of Convolutional Neural Network**

It has three layers namely, convolutional layer, pooling layer, and fully connected layer. The main advantage of using convolutional neural network is, it can identify the objects and faces present in the image. A convolutional tool dissects and recognizes diverse features within an image, a procedure known as Feature Extraction. Following this, a fully connected layer harnesses the outcomes of the convolutional process to forecast the image's category, leveraging the features extracted in earlier stages.

A SoftMax layer applies the softmax function to its input. Create a softmax layer using softmax layer. A classification layer computes the cross-entropy loss for multi-class classification problems with mutually exclusive classes. Create a classification layer using classification Layer. In classification tasks, it's typical to include a softmax layer followed by a classification layer after the final fully connected layer. The SoftMax function, often termed the normalized exponential, serves as the multi-class extension of the logistic sigmoid function. In typical classification networks, the classification layer typically follows the softmax layer. In the classification layer, train Network takes the values from the softmax function.

Recurrent Neural Network:

Recurrent Neural Network (RNN) are a type of neural there the output from previous step is fed as input to the current step. In traditional neural networks, inputs and outputs are independent, but for tasks like predicting the next word in a sentence, context matters. Recurrent Neural Networks (RNNs) address. This by incorporating a Hidden Layer that retains information about previous inputs. The key feature of RNNs is the hidden state, which stores sequence information.

LONG SHORT-TERM MEMORY:

Long short-term memory is a type of RNN, it is used for sequence prediction problems. The non-relevant information will be removed by using LSTM, and long short term memory have the efficient performance when compared to the RNN, it can be sustainable get the information with the long duration of time. It can be able to predict the information from the next data

or previous data. The main challenge in LSTM is it will take more time to drain the data depending on the size of the dataset. A CNN will be employed to extract information from the image, while an LSTM will generate captions for the input image.

DATASET COLLECTION AND DATA CLEANING:

The dataset used is flicker dataset, that contains images and descriptions that descriptions are in the form of dictionary with keys and values, it's a easy way to map the description with input images. Every text dataset needs to be done with the data cleaning process. That involves clearing the symbols like special characters like asterisk, semicolon, colon, double quotes. Then the keywords start with digits or ends with digits will be cleaned in this module. Compressing the long sentence which contains the inappropriate words.

EXTRACTING FEATURE VECTOR:

Using a pre-trained Inception model, we extract features from the image, aiding in the implementation process. This can't do anything the trained model will do everything because its already trained by the large image net dataset it will classify the various difference in the image. It will take 299*299*3 as an input image and removing the end classification layers for getting the 2048 feature vector. It can accept any image format including PNG, JPG, and others. The neural network reduces large set of features extracts from the original input into smaller recurrent neural network-compatible feature vector. The primary reason for referring to CNN as an 'Encoder' is its capability to extract and encode meaningful features from input data. In this context, within the module, CNN operates within both supervised and unsupervised learning paradigms. Specifically, when CNN is trained with labeled image data and their corresponding descriptions, it falls under supervised learning.

Provided with an external data source, the system generates outputs through pattern recognition learned from the trained data. This process is typically categorized as unsupervised learning. When we using the data generator it first going through the CNN layer and performing some process like pooling, next passes through the LSTM model it taken the output of CNN model and fit the first input with the second generated word with the help of dense. Comparing each pixel of an image long term short memory will forecast the suited description.

BLEU (bilingual evaluation understudy):

BLEU (Bilingual Evaluation Understudy) is an algorithm designed to assess the quality of text that has undergone machine translation from one natural language to another. Recognized as one of the initial metrics to demonstrate a strong correlation with human assessments of quality, BLEU remains widely used as an automated and cost-effective evaluation metric.

BLEU(N) = Brevity Penalty * Geometric Average Precision Score (N)

$$\text{Brevity Penalty} = \begin{cases} e^{1-\frac{r}{c}}, & C \leq T \\ 1, & C > T \end{cases}$$

$$\text{Geometric Average Precision(N)} = \exp\left(\sum_{n=1}^{N} W_n \log p_n\right)$$

Where,

C = number of words in the predicted sentence

r = number of words in the target sentence

W_n = uniform weights

P_n = N-gram precision

The model is trained and tested for various BLEU scores. BLEU scores tell us the quality of text that was machine-translated from a natural language. From the model has the highest BLEU score when the dropout value is 0.5.

Table 1: BLEU (bilingual evaluation understudy)

Images	BLEU Score (bilingual evaluation understudy)
Image-1	0.3868
Image-2	0.4562
Image-3	0.3868
Image-4	0.4431

Image Accuracy:

$$\text{ACCURACY} = \frac{TN+TP}{TN+TP+FP+FN} * 100$$

Where in:

True positive (TP) is the number of cases correctly identified as image features.

True negative (TN) is the number of cases incorrectly identified as image features.

False positive (FP) is the number of cases correctly identified as image description.

False negative (FN) is the number of cases incorrectly identified as image description.

Table 2: Voice based images Accuracy

IMAGES	ACCURACY
Image-1	87.1345
Image-2	86.8738
Image-3	85.1355
Image-4	86.4074

IV. RESULTS:

First process is to uploading the image it may be from the dataset which we have gathered or else it may be user own image. After that step it enter into various module then it will print the related description for an user given input, once the captions is created then it will play the audio of an caption generated.

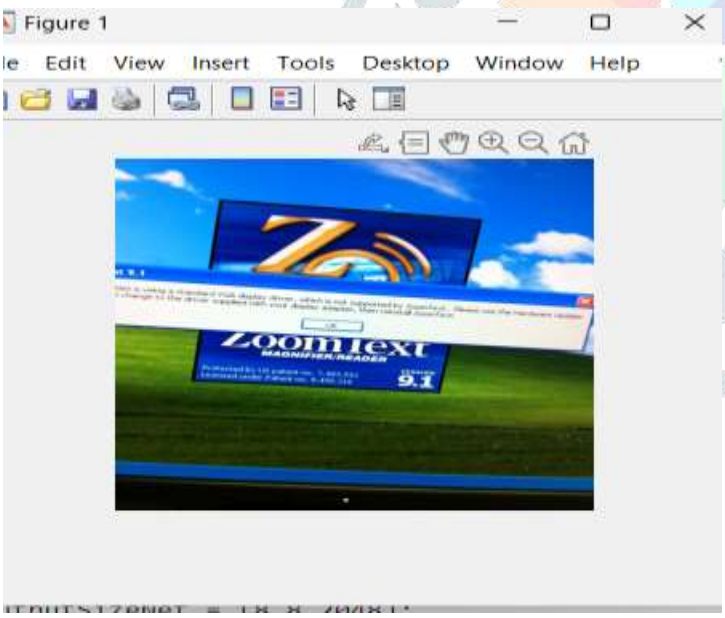
INPUT IMAGE 1:

```
caption =

    "a person is holding a bottle of medicine in their hand"
```

VOICE OUTPUT: **https://s19.aconvert.com/convert/p3r68-cdx67/fgxh9-va55u.mp3**

INPUT IMAGE 2:

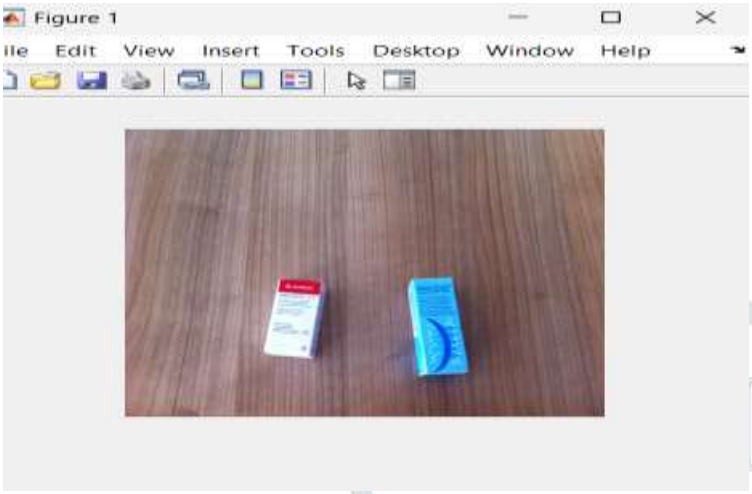



```
caption =

    "a computer screen is showing a window with a message"
```

VOICE OUTPUT: **https://s17.aconvert.com/convert/p3r68-cdx67/ihlkm-d2rpf.mp3**

INPUT IMAGE 3:

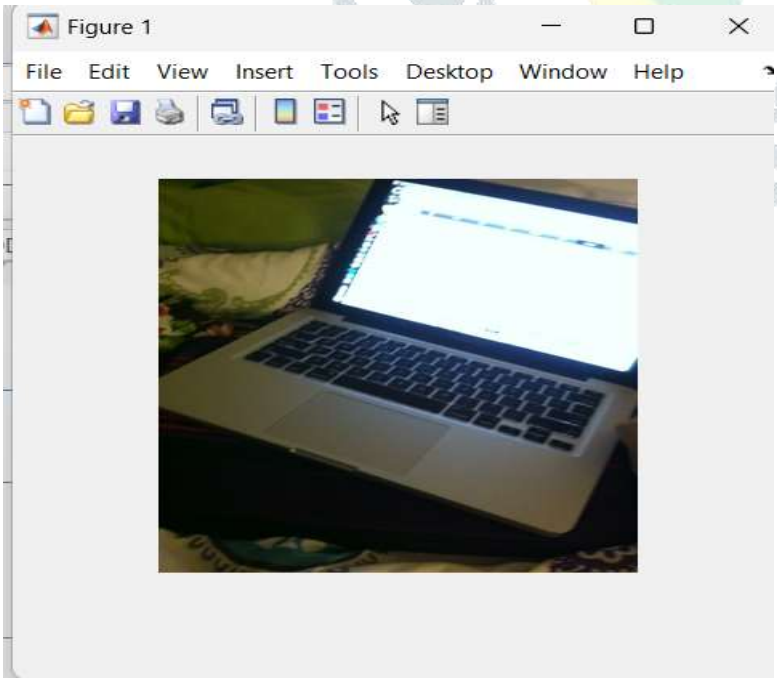



```
caption =

    "a box of medicine is on top of a wooden table the other is on the wooden table the product is a blue and"
```

VOICE OUTPUT: https://s21.aconvert.com/convert/p3r68-cdx67/59010-jk24g.mp3

INPUT IMAGE 4:




```
caption =

    "a laptop computer is on top of a bed"
```

VOICE OUTPUT: **https://s31.aconvert.com/convert/p3r68-cdx67/rjewa-yly8f.mp3**

V. CONCLUSION:

A voice-based image caption generator has been developed employing a CNN-LSTM model.Main key aspects of our project depend on the dataset, the proposed model is trained for testing the user input, so that it can predict the descriptions from the image. Out dataset consists of 7000 images. The proposed model is required to be trained on huge dataset that contains more than 7000+ images and generates caption based on image input.

VI. FUTURE SCOPE:

The project can be best applied for social media-based marketing strategies. The cost effectiveness and compatibility make it a boon for small companies with conventional systems to go for affordable and effective marketing strategies. This system now works on a model that was trained with a smaller set of datasets, by increasing the base dataset with which the model was trained, there is much potential for our project to cover domains other than marketing, like research purposes-to analyses the attitudes of today's society and thus helping in anthropological studies. The system currently works on English language. Altering and adding more datasets in different languages can expand the operation.

ACKNOWLEDGMENT

The satisfaction that accompanies the successful completion of this model would be incomplete without acknowledging the individuals whose constant guidance and encouragement made it possible. Their unwavering support has crowned our efforts with success, and for that, we are sincerely grateful.

VII. REFERENCES

[1]. Neha Tuniya, Shariva Dhekane, Vaishnavi Agrawal, Vibha Vyas (2021),"Image Based Caption Generator using Attention Mechanism"-12th International Conference on computing communication and Networking Technologies,2018.

[2]. Mirza Muhammad Ali Baig, Mian Ihtisham Shah, Muhammad Abdullah Wajahat, Nauman Zafar and Omar Arif (2018)," Image Caption Generator with Novel Object Injection"-School of Electrical Engineering and Computer Science, National University of Sciences and Technology, Islamabad, Pakistan.

[3]. S. Bengio, A. Toshev, O. Vinyals (2017),"Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge"- IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 4, 2017.

[4]. Ho-Jin Choi and Seung-Ho Han (2020), "Domain-Specific Image Caption Generator with Semantic Ontology" presented at the IEEE International Conference on Big Data and Smart Computing (Big Comp), 2020.Aayush Yadav, Aman Gill, Anurag Mishra, Nand Kumar Bansode, Pranay Mathur (2017)," Camera2Caption: A Real-Time Image Caption Generator"-International Conference on Computational Intelligence in Data Science (ICCIDS),2017.

[5]. Aayush Yadav, Aman Gill, Anurag Mishra, Nand Kumar Bansode, Pranay Mathur (2017)," Camera2Caption: A Real-Time Image Caption Generator" - International Conference on Computational Intelligence in Data Science (ICCIDS),2017

[6]. N. Komal Kumar, K. Laxman, A. Mohan, D. Vigneswaran, J. Yuvaraj (2019)," Detection and Recognition of Objects in Image Caption Generator System: A Deep Learning Approach"- 5th International Conference on Advanced Computing & Communication Systems (ICACCS),2019.

[7]. Sumathi, T., and Hemalatha, M. proposed a unified hierarchical model for automated image annotation and retrieval, as presented in their paper "A combined hierarchical model for automatic image annotation and retrieval," which was featured in the International Conference on Advanced Computing (ICAC) in 2011.

[8]. Dong-Jin Kim, Donggeun Yoo, Bonggeun Sim, In So Kweon, "Senetence Learning Deep convolutional neural Network for Image Caption Generation ", In: 13th International Conference on Ubiquitous Robots and Ambient Intelligence (URAl)-2016

[9]. Varsha Kesavan, Vaidehi Muley, Megha kolhekar, “Deep Learning based Image Caption Generation” Global Conference for Advancement in Technology (GCAT)-2019 .

[10]. Ren C. Luo, Yu-Ting, Hsu, Yu-Cheng, Wen, Huan-Jun, Ye (2019), "Visual Image Caption Generation for Service Robotics and Industrial Application" presented at the IEEE conference.

[11]. Yu, M.T., Sein, M.M.: “Automatic image captioning system using integration of N cut and color-based segmentation method”. In: Society of Instrument and Control Engineers Annual Conference (SCCEAC)- (2011).