Grenze International Journal of Engineering and Technology, Jan Issue

Image Caption Generation using Deep Learning Techniques

Mr. Siddharaj D. Pujari[1] and Mr. Atul S. Patane[2]
[1]Walchand College of Engineering, Department of CSE, Sangli, India
Email: Siddharaj.pujari@walchandsangli.ac.in
[2]Walchand College of Engineering, Department of Applied Mechanics, Sangli, India
Email: atul.patane@ walchandsangli.ac.in

***Abstract*— With use of natural language processing(NLP) and computer vision(CV) the problem of identification of contents of an image is solvable using artificial intelligence. In this paper, a generative model is presented which uses a deep recurrent architecture that combines recent advances in machine translation and CV which is used to generate natural sentences describing an image. An image caption generative model is created using Convolutional Neural Networks (CNN) and Long short term memory (LSTM) which uses different techniques of CV to understand an image's context and describe it in a manner native to English. For training purpose, the Flicker8K [17] dataset is used. It consists of 8000 unique images. Caption generation is one of the interesting areas of Artificial Intelligence that has many challenges. After generating the caption, it is converted into audio format. The result shows that the designed system is providing promising results w.r.t caption generation.**

***Index Terms*— Computer vision, Convolutional Neural Networks, Long short-term Memory.**

I. INTRODUCTION

It is difficult to automatically describe the content of images using natural languages. Image caption generation uses CV and NLP ideas for identifying an image's environment and describing it in English. It is useful for generation of textual description of an image. The approach may produce an English sentence explaining the content of an image when given an input of an image. Image caption, generates descriptions of an image based on the content present in an image and displays using natural language, it is an essential part of scene understanding. Today, we encounter a large number of information available in form of images from various sources such as advertisements, internet, document diagrams and news articles. Most images do not have a description, viewers would have to interpret the meaning of such images on their own, although without their in-depth captions, a person may grasp them in considerable part. However, if people require automatic image captions from the machine, it must understand some kind of image captions.

There are various applications of this system such as it is useful for visually impaired persons by adding voice based description of generated caption, it might aid persons who are blind in better grasping the meaning contained in images. It can be done by first converting the image into text and then to voice. It might deliver more precise and condensed information of photos and movies in scenarios such as video surveillance systems or image sharing in social networks . Social media platforms like Instagram, Facebook can conclude directly from the image

Grenze ID: 01.GIJET.10.1.141

uploaded by the user. The answers for questions like what you are doing (in a way), what you wear (color), where you are (beach, cafe, etc.) can be easily answered by providing images or photos. Automated captioning could make Google Image Search as excellent as Google Search by automatically turning every image into a caption before the search is conducted using the caption.
The proposed system uses an image as an input and generates output in the form of English text and audio. For this purpose, it uses deep learning and NLP. In terms of accuracy this system works well to identify the correct image captions. The text to audio conversion system is useful for blind persons to identify scenarios going on based on the images. This system will also be useful in creating summarized data from several images and pass this data for other applications for processing purpose such as creating videos, short story making, streaming analysis etc.

II. Literature Survey

This research [1] focuses on detection, recognition and generation of captions for a given image using deep learning techniques. For this purpose, Regional Object Detector (ROD) is used. The experiments are conducted on the Flickr 8k dataset and it provides promising results in terms of accuracy.
This paper [2] introduces caption generation system using a combination of stochastic gradient Markov chain Monte Carlo and recurrent auto encoding variation which is based on recurrent neural network (RNN). This model improves a conventional RNN-based language model by considering within-sentence word dependencies and temporal transitions between sentences. Result shows that the developed model outperforms well and creates different sentences and paragraphs that are syntactically and semantically correct and similar. [2] [22]
The proposed method [3] uses deep recurrent architecture which combines advances of CV and machine translation. Experimentation is carried out on several datasets which shows that the model generates caption with higher accuracy and the fluency of the language it learns solely from image description.
In this research [4] three different image captioning models based on deep neural network are considered such as CNN-CNN based, CNN-RNN [20][21] based and Reinforcement-based. All these models are tested for common dataset to compare its accuracy. The results show that all the model outperforms well with respect to accuracy and can be applied for automatic image description. [4]

III. Proposed System

The proposed system starts with uploading of an image. To generate the captions, the Convolutional Neural Networks (CNN) and Long Short Term Memory (LSTM) algorithms are used. The Xception model based on CNN is used to extract features of an image which is trained using ImageNet [18] dataset and then the output passed into the LSTM model which generates the captions of an image. After generation of an image caption the audio is generated using python library. The proposed system is shown in figure 1.

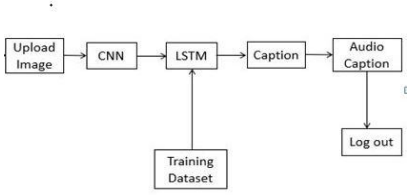

Figure 1. Architecture of Image caption generation system

The sub tasks involved in the proposed system are given below.

- *Uploading of an Image:* This step contains selection of appropriate image whose caption need to be generated.
- *Data Pre-processing:* The data set which contains images and its respective descriptions which will be used further at the time of training, in this step the available descriptions are cleaned i.e. removal of punctuation's, conversions of a caption into lower case letters, removing numbers from text these tasks are performed. This step also contains removal of all unique words from available descriptions and vocabulary creation. At the end it creates list of vocabulary and stores in a file.
- *Extracting the features from all images:* With help of Xception model, image features are extracted from all the images and every image is mapped with its name and corresponding feature array.
- *Loading data set for Training the model:* In this step the data set of images along with its features generated in the previous step and its descriptions will be loaded which will be used further in model training.

- *Vocabulary tokenizing:* Computers cannot understand English words directly, so for every vocabulary word a unique index number is assigned in this step.
- *Creating a Data generator:* A generator method is used to create different batches, to store huge amount of data in form of image features, vocabulary etc.
- *Defining the CNN-RNN model***:** With a dense layer, the extracted features from an image has a size of 2048 and reduced the dimensions to 256 nodes. To handle the textual input in embedding layer a sequence processor is used, followed by the LSTM layer. The Decoder is used to make the final prediction processed by the dense layer by combining the output from the two layers above it. The total number of nodes in the final layer are equal to our vocabulary size.
- *Training and Testing the model:* As part of model training, the 6000 training images are used by generating the input and output sequences in batches and model is trained. After training, the testing images are used to test the model accuracy, which loads the model and generates predictions. The predictions contain the index values of max length these predicted values are compared with the previous assigned index value stored in a file during pre-processing step.
- *Generation of audio:* After Generating text caption, the audio is generated audio for the same with python inbuilt library pyttsx3.

IV. Result Analysis

To perform the experimentation and to train the model, an ImageNet [18] dataset is used.
This dataset consists of 8000 images and corresponding image captions. Every image contains 5 captions which are numbered from 0 to 4. These images and captions are used to train our system. After completion of training, some images are used to test the accuracy of our proposed system. The figure 2 shows image uploading and corresponding captions generated by the system.
After performing testing of our model the accuracy recorded is 66 percent. The accuracy is totally dependent on the quality of an image. The figure 3 shows variation of result based on quality of image. For blurry images the corresponding generated captions are totally wrong, while for good quality images the captions generated by our model are more accurate.

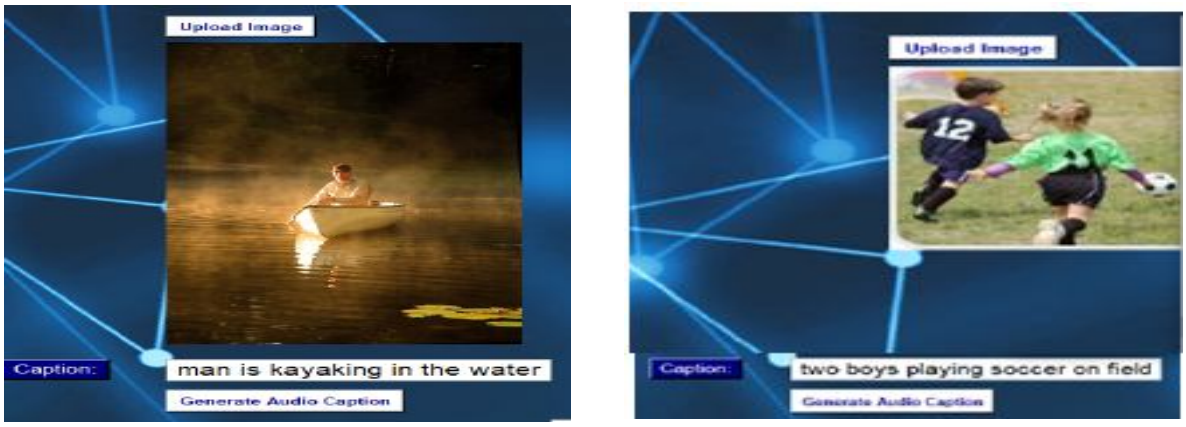

Fig 2: Image caption generation system

The table 1 and figure 4 shows the amount of time required to generate the captions after uploading of an image, it also includes the amount of time required for pre-processing of data. The statistics shows that time is dependent on size of caption and hardware type used.
The table 2. Shows accuracy of caption generation w.r.t. size of an image. The table shows image size (in pixels) and corresponding accuracy achieved by algorithm after training on images with given sizes. The result shows that based on size of an image it can be concluded that as size of image increases, the accuracy achieved in detecting captions also increases. To detect more accurate captions image size should be more than 224*224.
Table 3. shows that relationship of model accuracy for generating captions w.r.t. no of objects in an image. The objects in an image is nothing but persons, stationary objects like pole, big obstacles, lake, pond, flowers etc. If an image contains less number of objects, then the testing accuracy of caption generation is more. If objects in an image are more then caption generation accuracy decreases.

Image	Caption Generated
	Little boy in red shirt is swinging on swing.
	The little boy is playing with ball in the grass
	White crane is flying over the water
	Dog is running through the grass

Fig 3. Quality of Image: Accuracy decreases for blurry images

TABLE.I. TIME REQUIRED FOR GENERATION OF IMAGE CAPTIONS

Image caption word length	**Time(s)**
5	0.45
7	0.69
8	0.75
9	0.85
10	0.96
12	1.20
15	1.43
20	2.20

TABLE.II. IMAGE SIZE VS ACCURACY OF CAPTION GENERATION.

Image size (in pixels)	**Accuracy (in%)**
72*72	74.24
144*144	78.42
144*288	75.29
224*224	85.41
288*576	85.60
300*300	88.42

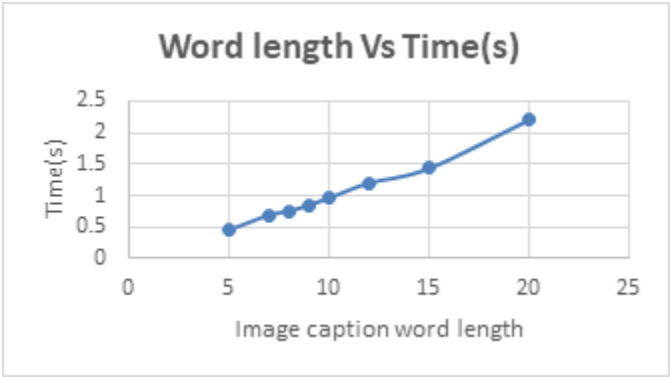

Fig 4. Word length and required Time(s)

TABLE.III. NO OF OBJECTS IN AN IMAGE VS ACCURACY OF CAPTION GENERATION

No of objects in an image	**Accuracy (in%)**
2	88.45
3	87.56
4	82.36
5	80.14
6	74.20
7	72.15

V. CONCLUSION

The proposed image caption generation system is based on deep learning and CNN-LSTM model. The system automatically detects the objects in the images and generates descriptions for the images in text as well as in audio format. The result shows that the system generates descriptions in more accurate way and it is totally dependent on quality of an image. The time required to generate the descriptions is also very less and it is also dependent on description of an image. As part of the future work, the system can be extended and used for saving the browsing history. This system can be used by visually impaired people independently, for this purpose it need to be deployed on braille system hardware.

REFERENCES

[1] N. Komal Kumar , D. Vigneswari et.al, “Detection and Recognition of Objects in Image Caption Generator System: A Deep Learning Approach”, 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), 2019.

[2] Dandan Guo et.al, "Recurrent Hierarchical Topic-Guided RNN for Language Generation", Proceedings of the 37 th International Conference on Machine Learning, Online, PMLR 119, 2020.

[3] Oriol Vinyals et.al, “Show and Tell: A Neural Image Caption Generator”, IEEE Explore, 2015.

[4] Shuang Liu et.al,"Image Captioning Based on Deep Neural Networks", MATEC Web of Conferences 232, 01052 (2018) https://doi.org/10.1051/matecconf/201823201052 EITCE 2018.

[5] Guiguang Ding, Minghai Chen et.al, "Neural Image Caption Generation with Weighted Training and Reference", Cognitive Computation (2019) 11:763–777

[6] Manjunath Jogin et.al, "Feature Extraction using Convolution Neural Networks (CNN) and Deep Learning", 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT-2018)

[7] S. ALBAWI and T. A. MOHAMMED, “Understanding of a Convolutional Neural Network”, in ICET, Antalya, 2017.

[8] O. Vinyals, A. Toshev, S. Bengio and D. Erhan, “A Neural Image Caption Generator”, CVPR 2015 Open Access Repository, vol. Xiv, 17 November 2014.

[9] D. S. Whitehead, L. Huang, H. and S.-F. Chang, “Entityaware Image CaptionGeneration”, in Empirical Methods in Natural Language Processing, Brussels, 2018.

[10] C. Elamri and T. Planque, “Automated Neural Image Caption Generator for Visually Impaired People”, California, 2016.

[11] G. Ding, M. Chen, S. Zhao, H. Chen, J. Han and Q. Liu, “Neural Image Caption Generation with Weighted Training and Reference”, Cognitive Computation, 08 August 2018.

[12] J. Chen, W. Dong and M. Li, “Image Caption Generator Based on Deep Neural Networks”, March 2018.

[13] R. Staniute and D. Sesok, “A Systematic Literature Review on Image Captioning”, Applied Sciences, vol. 9, no. 10, 16 March 2019.

[14] A. Farhadi, M. Hejrati, M. A. Sadeghi and P. Young, “Every Picture Tells a Story: Generating Sentences from Images”, in ACM Digital Library, 2010.

[15] J. Donahue, L. A. Hendricks and M. Rohrbach, ”Longterm Recurrent Convolutional Networks for Visual Recognition and Description,” CVPR 2015, vol. 14, 31 May 2016.

[16] Young P, Lai A, Hodosh M, Hockenmaier J. From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions. Trans Assoc Comput Linguist. 2014

[17] S. Hochreiter, “LONG SHORT-TERM MEMORY”, Neural Computation, December 1997.

[18] Fliker8Kdataset: https://www.kaggle.com/datasets/adityajn105/flickr8k

[19] ImageNet dataset: https://www.image-net.org/

[20] Siddharaj D Pujari, Anusha K. “Effective Prediction of Autism Using Ensemble Method”, Artificial Intelligence for Innovative Healthcare Informatics, Springer Nature Switzerland AG 2022.

[21] Prachi Chobhare1, Gayatri Bagul2, Harshad Soyam3, Rahulkumar Pardeshi4, Siddharaj Pujari5 and Atul Patane6,” Development of IoT-Based Smart Bin System for Effective Garbage Management”, Grenze International Journal of Engineering and Technology, Grenze Scientific Society, 2023

[22] Mr. Siddharaj D Pujari, Dr. K Anusha, “A Review on Prediction of Autism using Machine Learning Algorithm”, International Journal of Advanced Science and Technology Vol. 29, No. 6, (2020), pp. 4669 – 4678.