2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)
Amity University, Noida, India. Sep 3-4, 2021

2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO) | 978-1-6654-1703-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICRITO51393.2021.9596486

Generating Image Captions using Deep Learning and Natural Language Processing

Subhash Chand Gupta
Assistant Prof. Dept. of Computer Science and Engineering
Amity University
Noida, India
scgupta@gmail.com

Nidhi Raj Singh
Amity School of Engineering and Technology
Amity University
Noida, India
nidhirajsingh09@gmail.com

Tulsi Sharma
Amity School of Engineering and Technology
Amity University
Noida, India
sharmatulsi2015@gmail.com

Akshita Tyagi
Amity School of Engineering and Technology
Amity University
Noida, India
akshita.tyagi16@gmail.com

Rana Majumdar
MSIT, Techno India
rana.majumdarwb@gmail.com

***Abstract*: In today's world, there is rapid progress in the field of artificial intelligence and image captioning. It becomes a fascinating task that has saw widespread interest. The task of image captioning comprises image description engendered based on the hybrid combination of deep learning, natural language processing, and various approaches of machine learning and computer vision. In this work authors emphasize on how the model generates a short description as an output of the input image using the functionalities of Deep Learning and Natural Language Processing, for helping visually impaired people, and can also be cast-off in various web sites to automate the generation of captions reducing the task of recitation with great ease.**

Keywords: - Image Captioning; Natural Language Processing; Deep Learning; Convolutional Neural Network; Recurrent Neural Network; Long Short Term Memory (LSTM).

I. INTRODUCTION

Image caption generation is a famous subject amongst researchers and a lot of work has been done in this field in the recent past. Many techniques were produced in order to generate accurate captions but not all were as efficient as the methods which are available today. Today the efficiency and accuracy has enhanced by great deals. With the passage of time, the researchers have come up with lots of new methodologies for classifying images. Deep Learning is a widely used concept and is almost a recommendation for every classification. Neural Network, convolutional neural network, recurrent neural network, are some of the well-known deep learning techniques widely used for caption generation. The main purpose of this system is to generate image captions given as input by the users. This will help in organizing and labelling images in efficient manner without human involvement. By generating captions for multiple images of a same file one can organize or arrange those files easily and quickly. The people who are blind or the ones who have low vision can understand the images by their caption or description provided by the image captioning process. The images added in a website could be understood well when it has a valid description to it. Therefore, the process of generating a website could be done quickly by just adding the images and the caption for it could be found by applying the image caption generator. Hence, image captioning has become a critical tool for web development.

II. LITERATURE REVIEW

From the literature review, it is known that researchers are working on this issue and giving their opinions, interestingly a comprehensive Survey of Deep Learning for Image Captioning engraved by researchers [1] to deliberate the limitations and performance of image captioning for a given dataset. An evaluation matrix which is a deep learning modus operandi has also been discussed. Commonly used methods for generating captions are reinforcement learning and GAN based methods. Visual space based methods perform explicit mapping while multimodal space based methods use vision and language models. Methods which use CNN are based on encoder decoder architecture. On the other hand, some researchers [2] in their work presents a generative model on a deep recurrent architecture that merge the advances in computer vision and machine translation and used to generate phrases suitable for image. Specifically, NIC, an end-to-end neural network system that can involuntarily examine image and create a phrase that narrate the picture in English. Again some people [3] in their work boons a system that has two parts. The first part uses recognition algorithms and vision based detection for choosing the best words for the description of image. The second part known as surface realization chooses words and contracts natural language descriptions. They mentioned that the advantage of this system is automatic mining and parsing of huge texts collections. This model is fully automatic and effective in generating natural language. Excitingly numerous authors showed their interests in this area and keep working on various aspects, [5] in their work exhibits bi-directional mapping between the images and its descriptions is also conferred. This work practices RNN for visual representation and caption generation purposes. The first bi-directional model is capable enough to generate both novel image descriptions and visual features. This model is able to learn long term interactions. Demonstrations were made on the task of image retrieval, sentence retrieval, sentence generation and various datasets. Implementing CNN i.e. convolutional neural network, specifically (DCNN) dynamic convolutional neural network is used for modelling of

sentences in a semantic manner [6]. They emphasized on how DCNN (Dynamic Convolutional Neural Network) used k-max pooling which is known as a global pooling operation. The neural network handles the given input sentences of different lengths. SCST, Self-critical Sequence Training; is performed for captioning of images. The SCST is a popular form of reinforce algorithm. Researchers apply Reinforce algorithm throughout the training because this algorithm helps the data to be trained in a more effective way on non-differentiable metrics and provides a very remarkable improvement in captioning of images [7]. Equally some of them [8] in their work aimed to generate Image Captions through visual attention. Through visual attention model the content of images is described automatically. Standard back propagation techniques are also used with visual attention to train the image captioning model in a deterministic manner. Visual attention model helped in generating image captions to a greater extent and showed that through attention one can correspond remarkably to human intuitions. Again there is an encoder and decoder approach which can be combined with visual attention to produce tremendous results in different areas of interest. Keeping in the concept of Image Captioning with Semantic Attention few authors [9] tried to generate image captions through Semantic Attention. In their study they mentioned image captions can be generated through 2 approaches. These 2 approaches are top down approach and bottom up approach. In case of top down, the image captions are generated by examining the image from top to bottom and in case of bottom up the reverse process is followed. Both of these approaches are combined together to form semantic attention. Semantic Attention is a model which combines both the top down and bottom up approach to produce effective outcomes out of the image and further couples these 2 approaches with RNN i.e. Recurrent Neural Network that can attend the richer outcomes of the image captions selectively. Other researchers in their work mention the application areas of this concept like detecting human widget or even sentiment analysis using svm classifiers [12], [13]. Significantly image capturing and analysis is also demonstrated by few of them by using this concept [14].

III. PROBLEM STATEMENT

These days' social media have become a significant part of our life as one keep posting pictures/moments and also spend a significant amount of time in finding suitable captions for the same, and definitely a new software would save a lot of time, and would also be beneficial for disabled people. The main purpose here is to develop a model which can generate captions for images for a given input by the user. Here authors try to put emphasis on image captioning using a hybrid RNN & CNN approach after consulting various published research studies and tried to reduce the complexity to some extent. Knowing that RNN is suitable for problems related to speech/text analysis and generation; and CNN is suitable for image processing. So here in this work CNN will investigate the image for generating captions or related words and RNN will put words related to the identified parts or objects with the assistance of earlier knowledge, LSTM the modified version of RNN supports to remember the data that it trained from and connects that data with the current one. This will help to organize and label the images in an efficient manner without human intervention. As a result, the accessibility to the images will be easy.

IV. DESIGN AND METHODOLOGY

A. System Design

This framework is used to develop a model, labels for the images according to the picture present in the images. The basic flow illustration of the proposed system is constructed further all the components of this diagram are analyzed properly.

B. Block Diagram of Proposed System

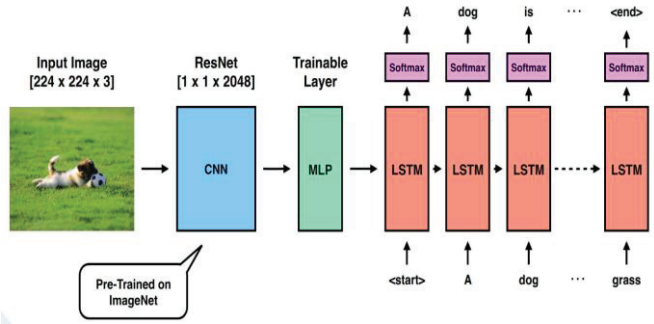

Fig. 1. Diagrammatic representation of the proposed model

C. Workflow of the Proposed System

Here the intended idea is to engender captions or descriptions for diverse images used as sample inputs. For accomplishing this, Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) are the two types of deep learning algorithms are implemented.

The image or picture gets passed through CNN to recognize objects and scenario present in a particular image. Through CNN, few more concepts were applied like pooling, padding and use of filters, etc. Basically, CNN is used to bring out all the features present in a sample image.

Additionally, in transfer learning where two main projects are considered: Glove and Inception V3. Glove defines the set of NLP vectors for common words and Inception V3 is used for bringing out important features from the image.

Further, through Natural Language Processing (NLP) the images are processed to simpler formats which makes easier for us to communicate with the computers. NLP techniques like tokenization, stemming is performed on the words before passing these words to RNN.

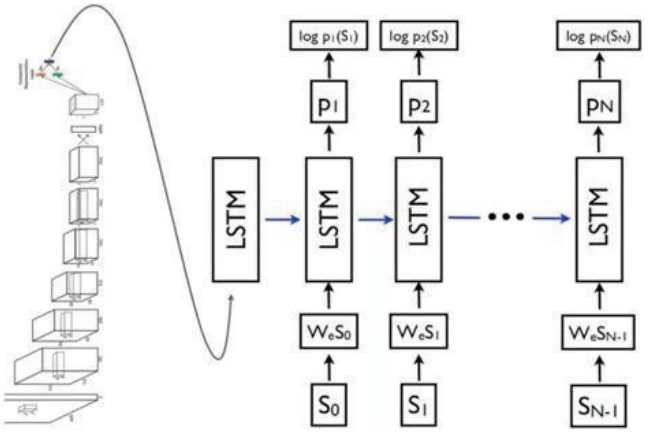

Fig. 2. Basic Workflow of the model

Finally, the set of words formed through CNN are send to the RNN. RNN model provides a meaningful description or sentence from the input provided by CNN.

D. Load Pre-Trained Network

Pre-trained networks are used to provide labels to the pictures. These pre-trained networks are present in huge amount. Around 1k classes of images are trained by these pre-trained networks. Few pre-trained networks are: Google Net, Alex net, VGG16.

E. Use Of Transfer Learning

Transfer learning is done for the training and classification of images into different classes. Through transfer learning, unknown classes could be trained and classified with higher accuracy.

As mentioned earlier, transfer learning consists of two important modules those are Glove and Inception V3. Here Inception V3 model with pre-trained weights is castoff for recognizing objects or figures in an input image.

F. Flickr8k Dataset

Flickr8k dataset comprises of images containing appropriate captions. Authors concentrated on Flickr8k dataset in their work because they found this dataset very suitable according to their problem statements. This dataset has suitable 1-line caption for every image which makes it suitable for the model.

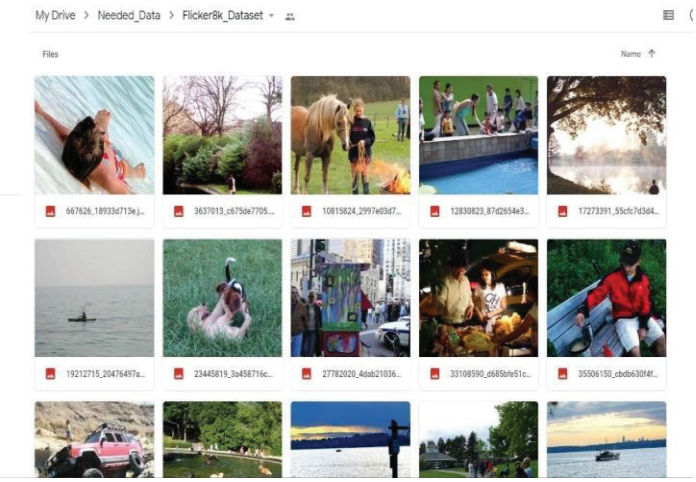

Fig. 3. Images available in our dataset

G. Platform Used

In this work authors used Google Colab for coding and executing purpose. Usually, it provides a platform for machine learning and data analysis. Colab notebooks can be loaded from GitHub and these notebooks are stored in Google Drive.

V. SIMULATION AND EXPERIMENTATION

A. System Implementation

From maintenance perspective database, consists of varied images of different categories are designated and each of them are provided with a meaningful caption. The first step will be object identification; this could be achieved by the image through the CNN. Further, the application of RNN and Long Short -Term Memory (LSTM) helps to assign a meaningful description for images.

B. Flickr Dataset

This Flickr8k dataset has 8000 images in PNG format and are of high quality in nature. The resolution for the images is 1024x1024.

C. Model Training

CNN reduces the image into essential features by using operations like convolution and pooling and then bringing those features into use for classifying the images. One of the main features of CNN is to detect patterns. RNN has several loops; these spirals transfer/communicate information regarding the object/image/data that they have gather knowledge about from past experience/training. These spirals/loops keep the information preserve and pass it on to next layer and take us one a step forward in our model. CNN and RNN are combined together to create a trained model.

D. Epoch

Epoch is simply the number of counts the model will be trailing across the data, that is, an epoch refers to one Passover through full dataset. More the number of epochs, more will be the improvement of the model, but only until a certain point after which model will not improve plus it will start taking longer to run the model.

E. Iterations

The term iteration indicates the number of times a batch of data is being passed through the algorithm. It is said that, when one iteration is completed when a batch of data is passed through neural network. For example, a dataset of 10 sample images, batch size 2 and specified number of epochs is 3, then number of iterations will be (10/2=5, 5*3=15) 15 iterations.

Below diagram shows an example of caption generation.

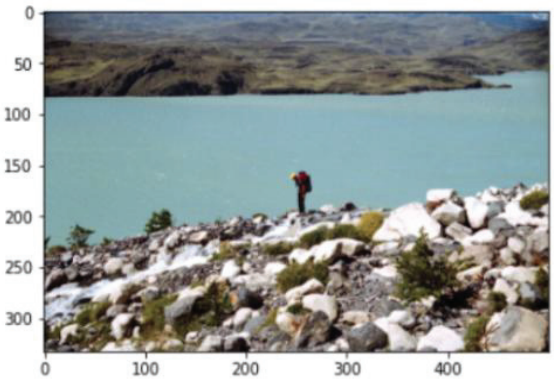

Fig. 4. Man standing at edge of Water

F. Predictions

The model will be ready to generate captions after the training and testing is completed. For this sample pictures will be given as an input and as output captions for the input picture will be generated.

VI. CONCLUSION

Generation of image captions is found to be an essential tool as it can be used for dissimilar meadows for their different purposes. By generating captions for multiple images of a same file one can organize or arrange those files easily and quickly. The people who are blind or the ones who have low vision can understand the images by their caption or description provided by the image captioning process. The images added in a website could be understood well when it has a valid description to it. Therefore, the process of generating a website could be done quickly by just adding the images and the caption for it could be found by applying the image caption generator. Hence, image captioning gaining its popularity and importance in the area of image processing knowhow.

REFERENCES

[1] Hossain, M. Z., Sohel, F., Shiratuddin, M. F., & Laga, H. (2019), "A comprehensive survey of deep learning for image captioning", ACM Computing Surveys (CSUR), 51(6), 1-36.

[2] Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2016). "Show and tell: Lessons learned from the 2015 mscoco image captioning challenge", IEEE transactions on pattern analysis and machine intelligence, 39(4), 652-663.

[3] Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Li, S., Choi, Y., & Berg, T. L. (2013). "Baby talk: Understanding and generating simple image descriptions", IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(12), 2891-2903.

[4] Chen, X., & Zitnick, C. L. (2014). "Learning a recurrent visual representation for image caption generation", . arXiv preprint arXiv:1411.5654.

[5] Kalchbrenner, N., Grefenstette, E., Blunsom, P. (2014)." A convolutional neural network for modelling sentences.", arXiv preprint arXiv:1404.2188.

[6] Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J., & Goel, V. (2017). "Self-critical sequence training for image captioning", IEEE Conference on Computer Vision and Pattern Recognition (pp. 7008-7024).

[7] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., & Bengio, Y. (2015, June). "Show, attend and tell: Neural image caption generation with visual attention", In International conference on machine learning (pp. 2048-2057).

[8] You, Q., Jin, H., Wang, Z., Fang, C., Luo, J. (2016). "Image captioning with semantic attention", In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4651-4659).

[9] Nguyen, D. K., & Okatani, T. (2019). "Multi-task learning of hierarchical vision-language representation", IEEE Conference on Computer Vision and Pattern Recognition (pp. 10492-10501).

[10] Sharma, S., Suhubdy, D., Michalski, V., Kahou, S. E., Bengio, Y. (2018). "Chat painter: Improving text to image generation using dialogue", . arXiv preprint arXiv:1802.08216.

[11] Sehgal, S., Sharma, J.,Chaudhary, N. (2020, June). Generating Image Captions based on Deep Learning and Natural Language Processing. In 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO) (pp. 165-169). IEEE.

[12] N. S. Ghosh, R. Majumdar, B. Giri and A. Ghosh, "Detection of Human Activity by Widget," *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, 2020, pp. 1330-1334, doi: 10.1109/ICRITO48877.2020.9197982.

[13] P. Das, A. Ghosh and R. Majumdar, "Determining Attention Mechanism for Visual Sentiment Analysis of an Image using SVM Classifier in Deep learning based Architecture," *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, 2020, pp. 339-343, doi: 10.1109/ICRITO48877.2020.9197899.

[14] S. S. Khan, R. Majumdar, P. P. Maut, A. Ghosh and V. P. Mishra, "Analyzing and Applying Captured Object with Machine Learning Techniques," *2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*, 2019, pp. 287-290, doi: 10.1109/ICCIKE47802.2019.9004382.

[15] Navaney, P., Dubey, G., Rana, A., "SMS Spam Filtering Using Supervised Machine Learning Algorithms", Proceedings of the 8th International Conference Confluence 2018 on Cloud Computing, Data Science and Engineering, Confluence 2018, 2018, pp. 43–48, 8442564

[16] Tyagi, N., Rana, A., Kansal, V., "Creating Elasticity with Enhanced Weighted Optimization Load Balancing Algorithm in Cloud Computing", Proceedings - 2019 Amity International Conference on Artificial Intelligence, AICAI 2019, 2019, pp. 600–604, 8701375