e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:03/Issue:09/September-2021 Impact Factor- 6.752 www.irjmets.com

IMAGE CAPTION GENERATOR USING CNN AND RNN (LSTM)

Ankit Kumar Yadav[*1], Eby Joyal Nadar[*2], Kishan Chaudhary[*3], Monika Pal[*4]

[*1,2,3]Department of I.T., St. Francis Institute Of Technology), Mumbai, Maharashtra, India.

[*4]Asst. Professor, Department of I.T., St. Francis Institute Of Technology),
Mumbai, Maharashtra, India.

ABSTRACT

Automatically generating captions from an image is one of the primary goals of computer vision. It can play a significant role in robotics area, such as reading picture books for babies, helping visually impaired and much more. Image caption generation area have received attentions since the development of Deep Learning especially, a model called Show and Tell which uses Long-short term memory (LSTM).LSTM is one of the most remarkable models in neural caption generation. We used transfer learning to detect features of objects in the images using InceptionV3 and then generate natural captions using LSTM on the image datasets. We have proposed a technique combining the best features of two state-of-the-art models which can efficiently and accurately provide very natural captions. The Convolution Neural Network part extracts the features of image, the vocabulary of our model is embedded with the Global Vector for word representations. The features and the vocabulary are both inputted to the caption model generation model which produces the output.

Keywords: Caption, CNN, LSTM, Caption Generator, Deep Learning, Image Captions.

I. INTRODUCTION

Being able to automatically describe the content of an image using properly formed English sentences is a very challenging task. This task is significantly harder, for example, than the well-studied image classification or object recognition tasks, which have been a main focus in a computer vision community. Indeed, a description must capture not only the objects contained in an image, but also, it must express how these objects relate to each other as well as their attributes and the activities they are involved in. Moreover, the above semantic knowledge has to be expressed in a natural language like English, which means that a language is needed in addition to visual understanding.

In this project, we will be implementing a caption generator system using Convolution Neural Networks (CNN) and Long short term memory (LSTM).The image features will be extracted using InceptionV3 which is a CNN model trained on the ImageNet dataset and then we feed the features into the LSTM model which will be responsible for generating the image captions.

Caption Generator has great potential impacts. As an example, it could help visually impaired people better understand the content of images on the web. Also, it could provide more accurate and compact information of images/videos in scenarios such as image sharing in social networks or video surveillance systems. All these, as well as the thought of learning something new and applying it excited us to create something useful in real life such as this project.

Our aim is to develop a system that can accurately identify objects in an image and can generate captions describing the scene of the image in a matter of seconds and then also compare its results with other state-of-art models using BLEU-SCORES.

II. RELATED WORKS

High level image understanding has drawn much attention in the recent years with the increasing success of deep learning and ever increasing data, the encoder-decoder architecture based on CNN and LSTM has been the main approach to solve Image Caption Generation problem. A number of researches on methods have been performed.

[1] proposed a deep residual recurrent neural network with two contributions. First, an easy -to-train deep stacked Long Short Term Memory(LSTM) language model. Second, to overcome the overfitting problem by larger-scale parameters in deeper LSTM, a novel dropout method was added. In [2], they described image using three methods : CNN- RNN based, CNN- CNN based, and Reinforcement based framework and compared them.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:03/Issue:09/September-2021 Impact Factor- 6.752 www.irjmets.com

However, they found that most of their models are not sensitive to the number of objects contained in the image such as "two" or "group". [3] proposed double LSTM image caption generation model with scene factors by using Resnet, Places365- CNN to extract deep scene features and training a multilayer perceptron to predict scene information in image. Double LSTM enabled models to describe images more accurately and improved semantic understanding of image effectively. In [4], they are using CNN and RNN for caption generation. Pre-trained Convolutional Neural Network (CNN) is used for the image identification task. This network acts as an image encoder. The last hidden layer of this network is used as an input to Recurrent Neural Network (RNN).This RNN network is a decoder which generates sentences. [5] have evaluated encoder-decoder caption generation frame-works with different choices of CNN encoders and observed that there is a wide variation in terms of both the scores, as evaluated with commonly used metrics (BLEU, METEOR, CIDER, SPICE, ROUGE-L), and also the generated captions while using different CNN encoders. This article indicates that we should use CNN model either Resnet, Densenet or Inception which has shown great results in their comparisons.

In [6], they present a model that can automatically generate an image description and is based on a recurrent neural network with modified LSTM cells with an additional gate responsible for image features. Their algorithm uses joint probability in LSTM to generate the next word and the caption. However, using just probability to generate the next word using lSTM can sometimes produce words out of context. This can be improved by using Global Vector for Word Representations as an embedding layer which handles this easily. [7] employed a regional object detector, RNN-based attribute prediction, and an encoder–decoder language generator embedded with two RNNs to produce refined and detailed descriptions of a given image. Their proposed system focused on a local based approach to improve upon existing holistic methods, which relates specifically to image regions of people and objects in an image. Their system showed impressive performance and outperforms state -of-the-art methods using various evaluation metrics. When dealing with cross-domain indoor scene images showed superiority over existing methods. However, detailed attributes, such as those related to garments , were not considered much. Saliency detection was also missing. Their model was not equipped with transfer learning to deal with image description generation for images such as cartoons and oil paintings. [8] consists of a phrase-based image captioning model using a hierarchical Long Short-Term Memory (phi-LSTM) architecture to generate image description. The phi-LSTM decodes image captions from phrase to sentence. It consists of a phrase decoder to decode the noun phrases of variable length, and an abbreviated sentence decoder to decode the abbreviated form of the image description. A complete image caption is formed by combining the generated phrases with sentences during the inference stage. There was only a small improvement in terms of relations and cardinality, because the CNN encoder they used does not hold any information regarding the relative position of the objects. Therefore, object relations are mostly inferred from the local statistics of training data. These literatures helped us identify a few areas to work upon such as a more detailed description of the objects that can be formed for the image, improving the Grammatical forms in the sentences and also working on creating fast and efficient models to generate captions in at most a second or two.

III. METHODOLOGY

To develop a system that can accurately describe an image and generate caption describing the scene and contents of the image in less than a second. Also provide a comparison of accuracy with other models in literature, in a graphical way. With the help of advanced CNN models and RNN models, InceptionV3 and LSTM respectively, develop a system that can accurately and efficiently generate captions for an input image. The model should predict the caption within 1 second. The bleu score of the model should be higher or equivalent to any of the state of art models found in the research phase. The generation of captions will be limited to 1 line of description. The captioning can be done within or less than a second. The comparison of different models will be done using their BLEU-SCORES using common test data with the help of a bar graph.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:03/Issue:09/September-2021 Impact Factor- 6.752 www.irjmets.com

A. Proposed Methodology

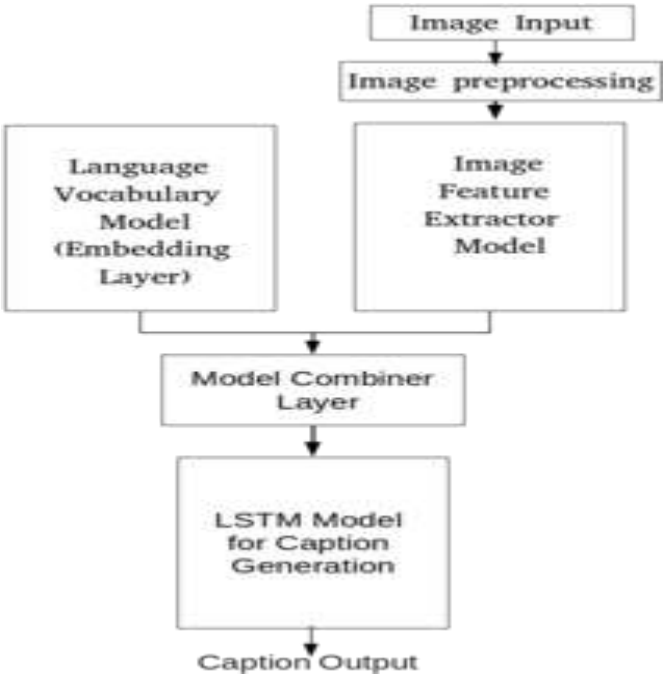

Figure 1: Architecture Diagram.

First of all, image preprocessing happens to convert the to required size. Once this is done, the features of the image are extracted using the InceptionV3 model of CNN. A layer then combines the feature layer output with an embedding layer (Vocabulary layer) which is embedded with Global Vector representations of words. Image feature along with vocabulary and an output sequence which initially have <startseq> token is passed to the LSTM model. The LSTM model with the help of vocabulary and feature vector generates captions. The captioning process utilises the features as long as it doesn't exhaust and the <endseq> token is encountered. Finally the full sentence is returned as output.

As shown in the above block diagram we'll be using a pretrained CNN model that can accurately identify objects. The algorithm used for the extraction of image features will be InceptionV3 based on [8]. The CNN will extract all the features of the image and then finally convert the feature vector into a linear vector. This linear vector can be used by the LSTM model for sentence generation.

The (Long Short Term Memory)LSTM model in RNN is very useful for generating sentences from features of an image. Each feature will be taken one by one and the output will be formed. This sentence will be transferred to next layer of LSTM and similar process will be repeated for each feature. Finally we'll get a caption generated for the input image.

Very user-friendly applications that a user can easily operate without any prior knowledge of the system, availability on the internet so users need not install anything to use except a browser and an internet connection and being, easily upgradable are some of the features of our system.

IV. IMPLEMENTATION

Implementation involves the following steps:

Step 1: Model Designing

The units and layouts are implemented here. The model's architecture is designed. The model is implemented using Keras version 2.4.3 .We used Inception-V3 which were pre-trained on ImageNet dataset and acts as encoder. The input images are first resized to 299 x 299, in pre-processing step before being fed to the CNN. The language model is built, trained iteratively on small datasets and tested for accuracy. Until the accuracy is improved the model is reiterated using different hyper-parameters. Once the model is formed with good accuracy among all the developed versions, it is taken to the next phase.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:03/Issue:09/September-2021 Impact Factor- 6.752 www.irjmets.com

Step 2: Gathering Training Datasets

To train the model dataset is the most important requirement. Hence, we took data from various sources such as Flickr8K, Flickr30K to include variety. We also created our own dataset apart from the online data sources.

Step 3: Training

The model developed so far is trained using the extensive datasets gathered from the various sources. During training we used ADAM for optimization with a batch size of 64. We trained the model for 25 epochs.

Step 4: Testing and Analysis

In this phase, the model is tested on various images in the testing datasets and their BLEU-SCORES are evaluated against the testing dataset captions. The BLEU-SCORES are also compared with other models in the reference papers. The model is also tested on some real-life images downloaded randomly from the internet.

Step 5: Web Application Development

In this phase the exported model is being used to generate using the input images. A website is being developed with a very user-friendly UI using Flask, HTML5, CSS, JS.

Step 6: Deployment

In this phase the whole system is containerized inside a docker container and then deployed on a real server that can handle requests from users from anywhere and anytime and generate captions.

V. RESULTS AND DISCUSSIONS

The dataset used for testing the model with other models was the Flicker8k testing dataset. The BLEU-SCORE criteria was used for evaluating the accuracy of our model with others. The BLEU-SCORE for a common image is shown in table V below with the bar chart also illustrated for the same.

Table 1. BLEU-SCORE Comparison

Paper	BLEU-1	BLEU-2	BLEU-3	BLEU-4
[1]	56.7	43.8	32.5	22.9
[2]	74.6	59.9	42.3	30.5
[3]	71.5	54.4	40.6	29.6
[5]	65.32	46.92	32.81	22.58
Our Model	**89.85**	**78.94**	**67.71**	**58.20**

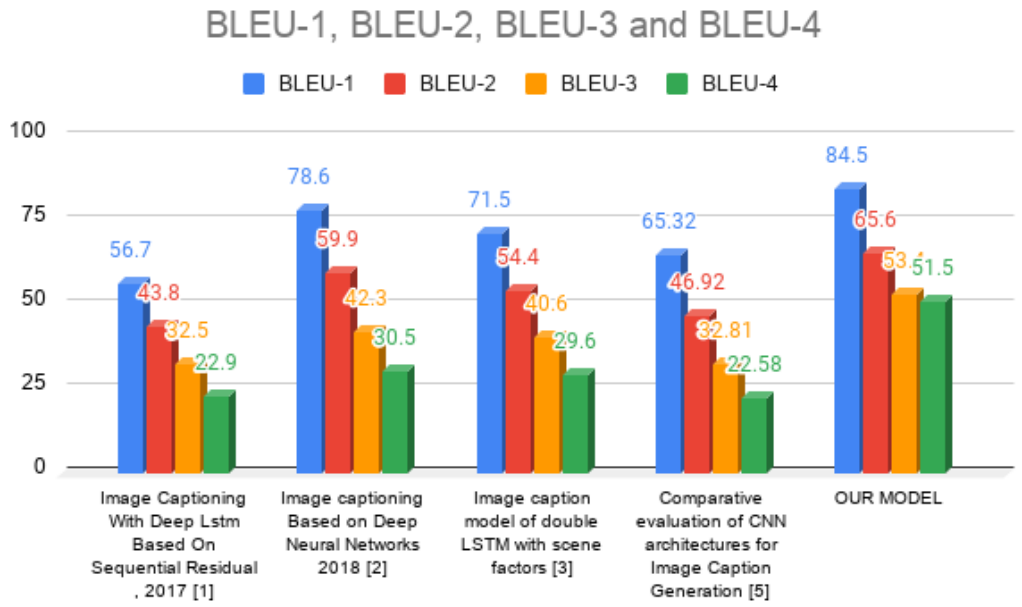

Figure 2: Bar Graph for BLEU-SCORES.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:03/Issue:09/September-2021 Impact Factor- 6.752 www.irjmets.com

The result achieved was mainly due to the GloVe embedding layer we used in our model. This helped us generate captions that are in context with the image. Attention can help improve the model much better when used in conjunction with this layer. Few samples of captions produced by our model are illustrated in Figure 3. Few of the captions are not very accurate because of not much of the same type of data in the training. A large and varied dataset can be used for training to improve upon such errors. However, considering the models mentioned the captions are better.

Figure 3: Examples of captions generated from our model.

VI. CONCLUSION

We have developed the system to generate captions using CNN and LSTM algorithms. Our model was able to generate captions which have a BLEU score of at least 0.7 which is very good in comparison to other state-of-art models. However, some errors occur while generating the captions at times when more than one objects are present or an unknown object is present. With larger dataset this can be improved upon with training. A much

larger and varied dataset will help to increase the accuracy of the model. Our model was able to generate captions very fast in less than a second most of the time with few exceptions, this helped us achieve our one of the objectives we had set in the beginning of our project. With this project we can conclude that, GloVe representations can help generate language models that are much more accurate and error free to be used in real life applications as it was visible in our project.

ACKNOWLEDGEMENTS

We are extremely grateful to our college St. Francis Institute of Technology for the confidence bestowed in us. We express our sincere gratitude to our respected director Bro. Jose Thuruthiyil, our principal Dr. Sincy George and our HOD Dr. Joanne Gomes for encouragement and facilities provided to us. We owe our profound gratitude to our project Mrs. Monika Pal Ma'am for guiding us throughout the development of our project. We would also like to thank Mumbai University for including this project in our curriculum as it not only helped us learn new things and apply them in practice but also taught us skills like teamwork, time management and much more.

VII. REFERENCES

[1] K. Xu, H. Wang and P. Tang, "Image captioning with deep LSTM based on sequential residual," 2017 IEEE International Conference on Multimedia and Expo (ICME), 2017, pp. 361-366, doi: 10.1109/ICME.2017.8019408.

[2] Liu, Shuang & Bai, Liang & Hu, Yanli & Wang, Haoran. (2018). Image Captioning Based on Deep Neural Networks. MATEC Web of Conferences. 232. 01052. 10.1051/matecconf/201823201052.

[3] Yuqing Peng, Xuan Liu, Weihua Wang, Xiaosong Zhao, Ming Wei, "Image caption model of double LSTM with scene factors" , Image and Vision Computing,Volume 86,2019,Pages 38-44,ISSN 0262 8856, https://doi.org/10.1016/j.imavis.2019.03.003.

[4] C. Amritkar and V. Jabade, "Image Caption Generation Using Deep Learning Technique," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), 2018, pp. 1-4, doi: 10.1109/ICCUBEA.2018.8697360.

[5] Sulabh Katiyar and Samir Kumar Borgohain, "Comparative Evaluation of CNN Architectures for Image Caption Generation" International Journal of Advanced Computer Science and Applications(IJACSA), 11(12), 2020. http://dx.doi.org/10.14569/IJACSA.2020.0111291

[6] A. Poghosyan and H. Sarukhanyan, "Short-term memory with read-only unit in neural image caption generator," 2017 Computer Science and Information Technologies (CSIT), 2017, pp. 162-167, doi: 10.1109/CSITechnol.2017.8312163.

[7] Kinghorn, Philip, Zhang, Li and Shao, Ling (2018) , "A region -based image caption generator with refined descriptions". Neurocomputing, 272. pp. 416-424. ISSN 0925-2312

[8] Ying Hua Tan, Chee Seng Chan, "Phrase-based image caption generator with hierarchical LSTM network", Neurocomputing, Volume 333, 2019, Pages 86-100, ISSN 0925-2312, https://doi.org/10.1016/j.neucom.2018.12.026.