International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 08 | Aug 2021 www.irjet.net p-ISSN: 2395-0072

Image Caption Generator using CNN-LSTM Model

Savitha S[1], Vishal V[2], Suhas CV[3], Sai Krishna Teja[4], Shivaprasad SB[5]

[1-5]*CMR Institute of Technology, Bangalore*

Abstract - *Caption for an image is generated by combining the keywords and punctuation, while there is no proper sentence to the caption which describes the image. Objects in the images are to be classified into human, animal, and plants by the model. To process these things, image processing and natural language processing are used to define the model. This can be done using CNN as an encoding vector layer by layer format, which will give a good idea about the image then they can be processed by the LSTM model, which will be used to predict or generate caption word by word. There are many datasets, which we can use to train our model. We are using the Flickr 8k dataset which is derived from the ImageNet dataset, which is very efficient for image processing.*

This model then generates a caption and we compare these results with the test dataset to generate the BLEU score. This BLEU score is used in determining whether the model is good or not.

Key Words: ***LSTM, CNN, Image Captioning, Computer vision, Natural Language processing, BLEU***

1. INTRODUCTION

Image is not just some dimensions with a colored object where each object tells what the image is about. Images speak something out and we just need to translate it into English words and join them to form a caption. Objects are the key to describe an image. To form a proper sentence out of it is another challenging work. We need to find the words that match the object in the image and combine the words to form a caption. Image captioning is representative of both Computer vision and Natural Language Processing. This helps our model to learn more than one sentence to identify the content of the image.

In this model, we are going to process this image using the deep learning model using a Convolution neural network class for visual identification on an image the output of it feeds into the LSTM model. We could have used the RNN instead of LSTM for the generating captions but LSTM is much better than RNN.

The first CNN model is used for the extraction of features of an image and then convert them into corresponding feature vectors. We will train the dataset using the CNN-LSTM model, then we will use the test dataset to evaluate the model and we can pass the image and get captions.

Figure 1

2. RELATED WORK

Deep learning development has changed the phase of Artificial Intelligence. In Deep learning, computer vision and [1] image processing has had rapid development in recent times. This rapid change has increased the graph of accuracy, speed, and performance in the field of image processing. It has changed the applications with which it is related like Self-Driving Cars, Visual Recognition, and Language Translation, etc.

Computer vision initially performs tasks like acquiring, processing, analyzing, and extracting symbolic information of the image to understand it. After the evolution of these methods, we got the convolution neural network which was started in 2014 which evolved from ILSVRC 2012 which was image classification of large data. Using these algorithms, CNN is now evolving a large dataset like the ImageNet which has around 14 million images.

Natural language processing (NLP) [3,4] is used to improve interactions between human language and computers. NLP gives us a better idea to program a particular task and gives us better ways of analyzing the data. As technology is advancing, NLP is coming with the best possible solutions using AI. This technology allows the computer to understand the human language, whether it can be in the form of voice or text or any other way. Using this technology a machine

will be able to understand human language and sentiment. Information retrieval is important in natural language processing. With the combination of both, they started a change in text processing. It started with language translation which is used to translate text from one language to another. Next, they have a trained system to give responses to their commands with specified actions. Also used for summarizing a huge set of data. NLP can be used in enterprise solutions, using which the productivity of employees can be increased and can prepare solutions for critical problems.

Images are processed using computer vision and natural language processing, which gives the caption for the image. [8] BLEU score algorithm is used for evaluating the machine predicted text. BLEU score compares word to word to provide better accuracy. This caption accuracy may vary from image to image due to the object identification in the image. This caption accuracy can be found using the BLEU score which lies between 0-1.

3. PROPOSED SYSTEM

3.1 System Architecture

The proposed system [8,9] generates a caption for the image given as input. This is possible by training the deep learning model with the dataset. To increase the quality of captions, we can do the cleaning of the dataset by removing unnecessary things and punctuations. We can use semantic feature extraction and making vocabulary. Using glove embeddings can increase the quality of captions generated.

By using an Encoder-Decoder model we can limit the length of the generated captions. So that our model will generate captions of that fixed length for a given input image.

This model is mainly divided into three modules:

1. Image recognition model - It is used for the extraction of features for a given image.
2. Semantic feature extraction model - It extracts the specific keyword in the dataset which improves the quality of captions.
3. Language model - It will generate captions based on the features of the given image.

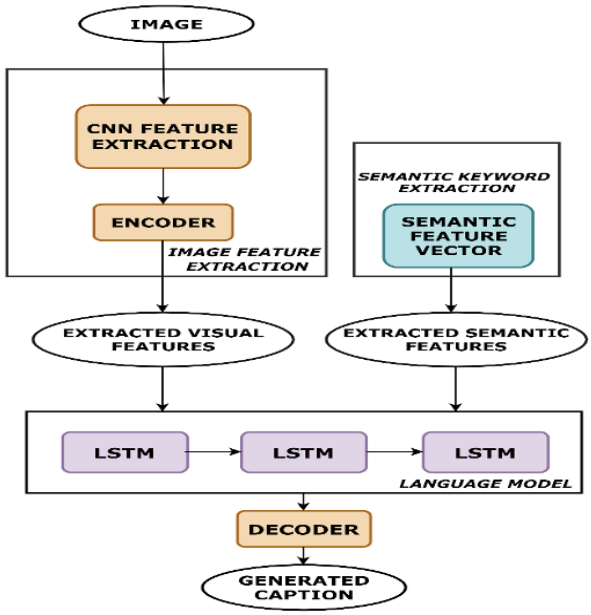


Figure 2

3.2 Image recognition model.

It involves extracting the features from the image like Color, Texture and patterns, Shape and objects, Position, etc... Using a Convolutional neural network (CNN).

Convolutional Neural Network (CNN) is used for identifying patterns and understanding them. It is a type of ANN. CNN's are a class of Deep Neural Networks that can detect and classify particular features from images and are widely used for analyzing visual images. Their main applications range from recognition of images, classification of images, analysis of medical images for disease identification, natural language processing, and computer vision.

CNN has five different layers which will perform different tasks.

The five different layers in CNN are:

1. Input layer – input layer contains the image data that is fed as input
2. Convolutional layer - This layer is used for feature extraction for images that are fed as input to the layer. This layer performs the mathematical operation of convolution between a filter of a particular size MxM and the input image. The dot product is performed between the image and the filter. The output gives information about the input image such as the edges, corners, and patterns is called the Feature map. This output is given to other layers as input to get any other features if required.

3. Activation function layer -This layer is used to learn and approximate any complex and continuous relationship in variables of a network. It is used to classify the data which can be forwarded in the network or data which can be left.

4. Pooling layer – The pooling layer is used for reducing computational costs by reducing the size of the feature map. This can be achieved by making only fewer connections between layers.

5. Fully connected layer - This layer consists of the neurons along with biases and weights. It is used to make a connection between the neurons of two different layers. It can be used for the classification of images after training.

Output layer - holds the extracted feature values at the end of the process

Inception v3 is a widely-used image recognition model that has been shown to attain greater than 78.1% accuracy on the ImageNet dataset.

3.3 Semantic Keywords Extraction.

Semantic Keyword extraction is the process of automatically identifying the terms that best describe the subject of a document, in this case, an image that is grammatically correct. The generated caption should consist of much more clear information on attributes of the image such as color. By using Semantic Keyword Extraction, the quality of captions will be increased and the captions generated will be more grammatically correct.

After getting the vocabulary or keywords, these are given as input to the perceptron network. This network is a multi-layered network with each layer designated with a unique function. Some layers like the dropout layer, softmax layer, and Dense layer are responsible for the identification of attributes while will help the model to generate meaningful captions.

The steps involved are:

Step 1: The dataset will be cleaned by removing punctuations.

Step 2: Data loading and preprocessing.

Step 3: converting words to vectors

Step 4: Creating a vocabulary with unique words

Step 5: From the dictionary created in the previous step, the top 400 semantics will be selected.

Step 6: The vector that is created in the previous step is given to multi layered network.

3.4 Language Model.

The language model we used is CNN-LSTM because [10] the experimental facts reveal that LSTM is better than RNN for text generation and image processing because it doesn't have a vanishing gradient problem.

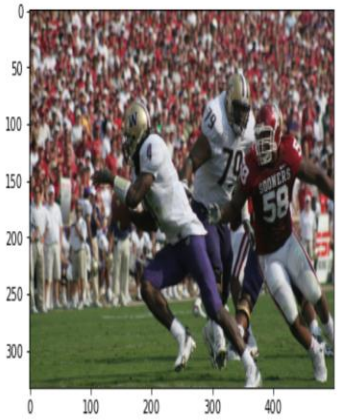
LSTM is like a better version of RNN which is developed to tackle the vanishing gradient problem of RNN. They have feedback connections that will let us process entire data sequences like video or speech. That's why they have a huge number of applications in image recognition, speech recognition, handwriting recognition, and video processing.

The features of images that are extracted from previous steps are given as input to this model. The sequence processor with the help of the tokenizer class will generate captions.

Since LSTM layers have a memory unit that recognizes which word may come after the first word and generate the meaningful words. This model will take the image vector as input and generates a caption. This input vector is obtained from the CNN model, in this case, the InceptionV3 model. We will use the Second layer of the CNN model to get a vector from an image.

4. EXPERIMENTAL RESULTS

Greedy Search: group of people racing on snowy hill

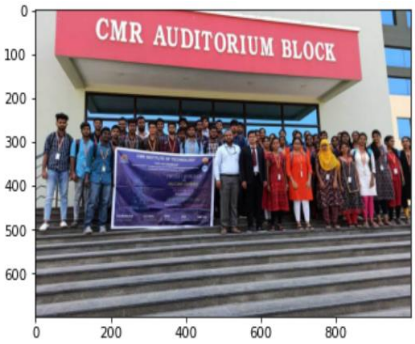


Greedy Search: football player in red and white uniform is being tackled by oklahoma player

Greedy Search: two children are playing hockey in the snow

Greedy Search: two boys playing soccer

Greedy Search: black dog is running across grassy yard



Greedy Search: group of people are standing around in front of building

EPOCHS	DROPOUT	BLEU SCORE
10	0.5	0.614561
20	0.5	0.628900
30	0.5	0.634662
40	0.5	0.649467
50	0.5	0.632956

Table 1

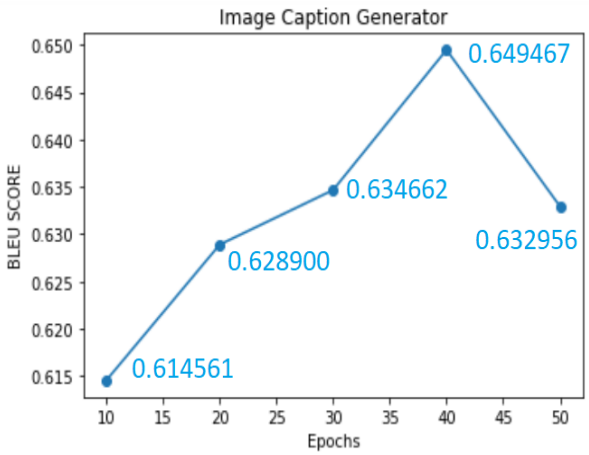


Figure 3

```
Bleu_Score -1 = 0.649467
Bleu_Score -2 = 0.446046
Bleu_Score -3 = 0.318250
Bleu_Score -4 = 0.219204
```

Figure 4

Our model has achieved a BLEU score of 0.64 which is greater compared to the reference model. In table 1, we can see how several epochs affect the accuracy of our model. We can see after some epochs it accuracy started decreasing is due to the over fitting of the model. The Dropout we used is 0.5, it is also used to avoid over fitting of the model. If we use dropout it will ignore some layers of input information to ensure the model will not be over fitted. Figure 3 is just a graphical representation of table 1, which will give us a better understanding of table 1.

Figure 4 depicts how BLEU scores can change by changing weights.

5. CONCLUSION

In this paper, we have the CNN-LSTM model to generate captions, using the Flickr 8k dataset. This is an image processing model using Natural language processing and computer vision to generate captions. We have added a feature that is the voice-over for captions. In this model, we have used layer two of InceptionV3 to extract features of images with Imagenet weights. We have used glove embeddings to increase the accuracy or correctness of captions. The complete model has made a uniqueness in image recognition models.

This model has got a BLEU score of 0.64 is a good score. As technology is improving rapidly, in the future we can expect some better models which can do wonders in image processing. This image processing concept will become vital seeing the development of self-driving cars. In recent years we can see a huge development in neural networks and computer vision. The development of next-level LSTM is going on, in the coming days we can expect some models which give better results.

The quality of captions not only depends upon the model, but it also depends on preprocessing of data and using the correct dataset. The accuracy of the model depends on the number of epochs that the model is trained.

6. REFERENCES

[1] Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan, Show and Tell: A Neural Image Caption Generator 17 Nov 2014

[2] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, Zbigniew Wojna, Rethinking the Inception Architecture for Computer Vision, 2 Dec 2015

[3] Prakash M Nadkarni, Lucila Ohno-Machado and Wendy W Chapman, Natural language processing: an introduction, 2011 Sep-Oct

[4] IBM, https://www.ibm.com/cloud/learn/natural-language-processing , [online], 2 July 2020

[5] Wikipedia, https://en.wikipedia.org/wiki/Natural_language_processing, [online]

[6] Upgrade,https://www.upgrad.com/blog/basic-cnn-architecture/ , [online], Dec 17 2020

[7]Medium,https://medium.com/techiepedia/binary-image-classifier-cnn-using-tensorflow-a3f5d6746697 , [online], 29 Aug 2020

[8] L Abisha Anto Ignatious, S Jeevitha, M Madhur Ambigai, M Hemalatha, A Semantic Driven CNN - LSTM Architecture for Personalised Image Caption Generation, 2019 11th International Conference on Advanced Computing (ICAC)

[9] Moses soh, Learning CNN-LSTM Architectures for Image Caption Generation

[10] Haoran Wang, Yue Zhang and Xiaosheng Yu, An Overview of Image Caption Generation Methods