e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:07/July-2021 Impact Factor- 5.354 www.irjmets.com

IMAGE CAPTION GENERATOR USING IMAGE FEATURES AND LSTM NETWORKS

Thejesh M[*1], Muni Pranay Polampalli[*2], Sunil Kumar Reddy D[*3], Yuvaraj G[*4], Dr. Anand Kumar R[*5]

[*1,2,3,4]B-Tech Student, Department Of Computer Science And Engineering,
Madanapalle Institute Of Technology & Science, Madanapalle, Andhra Pradesh, India.
[*5]Assistant Professor, Department Of Computer Science And Engineering,
Madanapalle Institute Of Technology & Science, Madanapalle, Andhra Pradesh, India.

ABSTRACT

At present, Image caption generator has raised an enormous interest in multimedia but understanding the background details of the image and automatically generating the text related to image with less latency and computationally efficient manner is a challenging job. As it contains two major categories that is image/video and the text, we should get into Deep Learning and Natural Language Processing tasks which are complex and requires high computational power which makes the job challenging. Inspired by many related works in this field that have shown a vast number of models and different encodings of the image with CNN architectures. Our proposed approach aims to build an encoder-decoder architecture where the encoder is a pre-trained model like VGG16, Resnet50, InceptionV3 and MobileNet. It may be a single model or the combination of vectors consisting of high-level features from two or more models works exceptional in most of the cases. Besides extracting high-level features from images, we also maintain the image color composition using OpenCV techniques which also helps the model to extract the features from small components in the image. Whereas the decoder part is a slightly modified LSTM network and contains the Time Distributed Layer which are helpful for time series data as we post the entire sentence or sequence as a time series problem and also in case of videos as well which works fabulously for storing sequence information and text generation.

Keywords: Deep Learning, Natural Language Processing, LSTM Networks, Time Distributed Layer, Opencv.

I. INTRODUCTION

Every day, we are exposed to a significant number of images from a variety of sources, including the internet, news articles, schematics in documents, and advertisements. These resources provide visuals that visitors must interpret for themselves. Majority of photos do not contain a description, yet humans can make sense of them without them. However, if people want automated image captions from the machine, the system must be able to understand and interpret them. The most accelerated technologies of this era are deep learning and machine learning. Artificial intelligence is now compared to the human mind and it does great work than people in some fields. New research in this area occurs every day and this field is growing very quickly because we now have enough computational power to do this. Deep learning is a machine learning branch that uses many-layered neural networks. Deep learning networks are often enhanced by increasing the amount of data used to train them. The technique of generating captions for an image is known as Image Captioning. We must first comprehend the significance of this challenge in real-world scenarios. Let us consider, few scenarios in which a solution to this problem could be extremely beneficial. Automatic captioning could help Image Search become as good as Web Search, because every image could be transformed into a caption first, and then searches could be conducted based on the caption. The model is a form of encoder-decoder architecture that was trained on Flickr 8k dataset provided by the University of Illinois at Urbana-Champaign. For this topic, there are numerous open-source datasets accessible, such as Flickr 8k (which contains 8k photos), Flickr 30k (which contains 30k photos), MS COCO (which has 180k photos), and so on. The images were chosen from six separate Flickr groups and do not feature any well-known persons or places, but they are manually selected to reflect a diversity of scenarios and circumstances.

II. METHODOLOGY

The Key functions that are to be noticed, processed, and implemented are:

1. Accessing and reading Image and Caption Dataset (Flickr 8k).

2. Preparing the Image Data and processing it.
3. Developing Encoder model (Resnet50) and extracting features from images.
4. Processing the Text data, cleaning and performing word embeddings.
5. Developing a deep learning-based LSTM Model
6. Training the model with Progressive Loading with Data generators.
7. Evaluate the trained model with test dataset.
8. Creating a web-app and good User Interface.

Image captioning is a deep learning-based model which contains complex architectures and corresponding weights that are to be trained for better performance and predictions. Integrating different functions is one of the crucial things in training the model because each layer should be compatible with the next corresponding layers. Choosing a right Optimizer and loss function makes the model performance better in short period of time. The computational power to train the deep learning models is high and we are balancing with simple model architectures. The respective pre-trained models should be downloaded, and all the packages should be installed.

Initially an image from any source is given as input to the encoder-decoder model where the given image is processed into a pre-trained Resnet50 model and the 2048 dimensional vector is extracted from the model which contains high-level image features of the image and then the vector is embedded into a dimensions of the sentence vector which is initially a <start> tag and whose two vectors are concatenated and passed over a Time Distributed Layer followed by LSTM (long short term memory) cells. The output of each LSTM cell is passed as a input of next LSTM cell until the <end> tag is reached. The outputs of all LSTM cells are concatenated to produce the final output. The final output sentence is a brief description of the given input Image.

III. MODELING AND ANALYSIS

Encoder :

The encoder is used to encode the given data to extract the insights from it. In our project the encoder is used to extract the high-level features from the images. The encoder we are using here is ResNet50 as shown in the figure 4.6. ResNet50, or Residual Networks, is a well-known neural network that is utilized as the backbone for many computer vision tasks. In 2015, this model was the winner of the ImageNet challenge. The fundamental breakthrough with ResNet was, it allowed us to successfully train extraordinarily deep neural networks with 150+ layers. Due to the problem of vanishing gradients, training very deep neural networks was difficult before ResNet. ResNet is a sophisticated backbone model that is utilized in a wide range of computer vision applications. To add the output from an earlier layer to a later layer, ResNet uses skip connections. This helps in resolving the vanishing gradient issue.

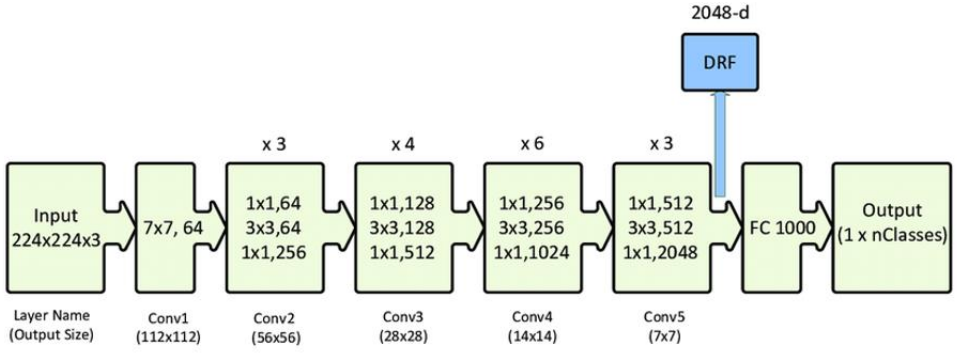

Figure 3.1 Encoder

The ResNet-50 model is divided into five stages, each with its own convolution and identity block. There are 3 convolution layers in each convolution block as shown in Fig 4.8, and 3 convolution layers in each identity block. There are around 23 million trainable parameters in the ResNet-50. There was a small change made for the ResNet50 and above that previously, shortcut connections skipped two layers, but now they skip three layers, and 1 * 1 convolution layers were added, which we will go over in detail with the ResNet50 Architecture.

Decoder:

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:07/July-2021 Impact Factor- 5.354 www.irjmets.com

Long Short-Term Memory networks – usually just called "LSTMs" – are a special kind of RNN, capable of learning long-term dependencies. They work tremendously well on a large variety of problems and are now widely used. LSTMs are specifically developed to prevent the problem of long-term dependency. They do not have to work hard to remember knowledge for lengthy periods of time, it's nearly second nature to them. All recurrent neural networks are made up of a series of repeated neural network modules. This repeating module in ordinary RNNs will have a relatively simple structure, such as a single tanh layer.

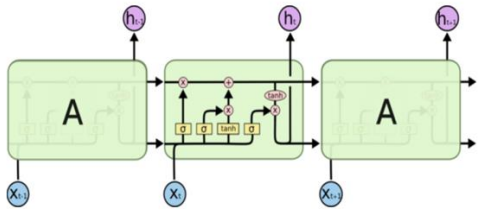

Figure 3.2 LSTM Architecture

The cell state, the horizontal line going through the top of the diagram, is the key to LSTMs. The state of the cell is similar to that of a conveyor belt. With only a few tiny linear interactions, it flows straight down the entire chain. It is incredibly easy for data to simply travel along it unaltered. The LSTM can delete or add information to the cell state, which is carefully controlled by the structure.

dense_22_input (InputLayer)	[(None, 2048)]	0	
embedding_8 (Embedding)	(None, 40, 128)	1056512	embedding_8_input[0][0]
dense_22 (Dense)	(None, 128)	262272	dense_22_input[0][0]
lstm_24 (LSTM)	(None, 40, 256)	394240	embedding_8[0][0]
repeat_vector_6 (RepeatVector)	(None, 40, 128)	0	dense_22[0][0]
time_distributed_8 (TimeDistrib	(None, 40, 128)	32896	lstm_24[0][0]
concatenate_8 (Concatenate)	(None, 40, 256)	0	repeat_vector_6[0][0] time_distributed_8[0][0]
lstm_25 (LSTM)	(None, 40, 128)	197120	concatenate_8[0][0]
lstm_26 (LSTM)	(None, 512)	1312768	lstm_25[0][0]
dense_24 (Dense)	(None, 8254)	4234302	lstm_26[0][0]
activation_8 (Activation)	(None, 8254)	0	dense_24[0][0]

Total params: 7,490,110
Trainable params: 7,490,110

Figure 3.3 Model

Training & Validation Analysis:

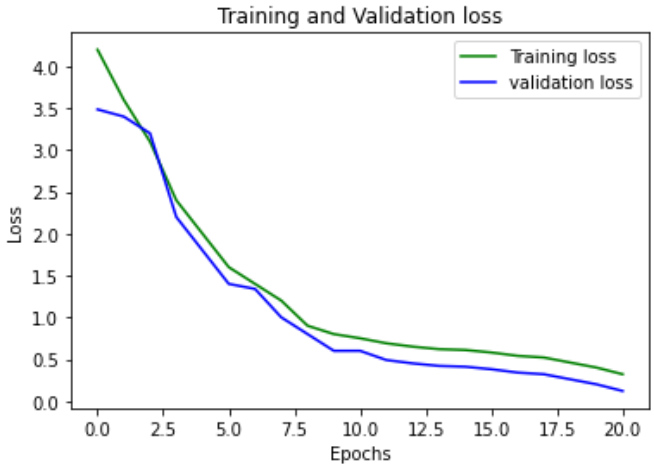

Figure 3.4 Training and Validation Loss

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:07/July-2021 Impact Factor- 5.354 www.irjmets.com

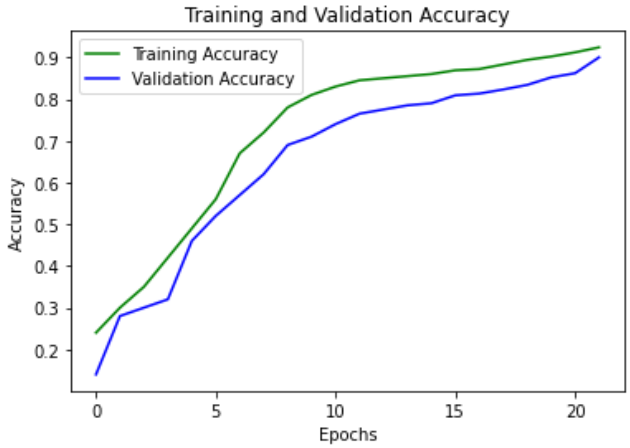

Figure 3.5 Training and Validation Accuracy

IV. RESULTS AND DISCUSSION

```
print(Argmax_Search)
```

```
Two dogs wrestle in the grass .
```

Figure 4.1 Output

Figure 4.2 Output

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:07/July-2021 Impact Factor- 5.354 www.irjmets.com

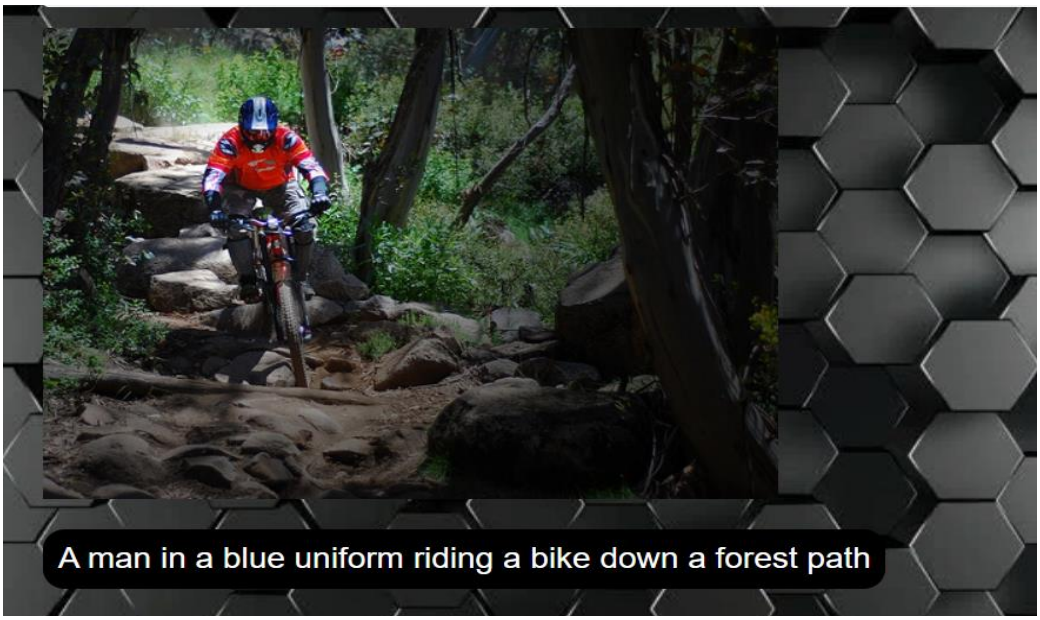

Figure 4.3 Output

Table

Testing of Captions generated on Images		
Number of Tests Carried	Number of Tests passed	Number of Tests Failed
45	42	3

The following test case scenarios were used in the integrated system testing to prove the working of the developed system.

- Encoder Output vectors and their corresponding Shapes.
- Word embeddings and the vector dimensions of the LSTM input.
- Model prediction on Validation and Test data.
- Semantic meaning in the generated captions of the Image.
- Validating the model with irrelevant images of the Train data.
- Checking all the compatibilities of the vectors and input shapes.
- Display of web-based Application and the GUI. All test cases were successfully tested. The system developed is user friendly and no special training is required for caption generation.

V. CONCLUSION

The proposed model is robust, with low-computational power and does not require any special training as the accuracy of the model on training data is over 92.4% with the loss of 0.32 and the validation accuracy of 90% with the loss of 0.12. This architecture uses of the existing developments of the pre-trained models and various types of deep learning-based and Natural language techniques. As the whole system is simplified, the scalability of major applications with this model is tremendous. As we used Time distributed layers followed by LSTM cells the sequence information of the sentence will get stored and when generating captions for videos as the frames of the video are time-based, these time-distributed layers will perform tremendously well of such data. The generated captions while validating the model with test data are accurate and well-defined semantic meaning.

VI. REFERENCES

[1] Ding, S., Qu, S., Xi, Y., Sangaiah, A. K., & Wan, S. (2019). Image caption generation with high-level image features. *Pattern Recognition Letters*, *123*, 89-95.

[2] 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.

[3] Aneja, J., Deshpande, A., & Schwing, A. G. (2018). Convolutional image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5561-5570).

[4] Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2625-2634).

[5] Gong, Y., Wang, L., Hodosh, M., Hockenmaier, J., & Lazebnik, S. (2014, September). Improving image-sentence embeddings using large weakly annotated photo collections. In *European conference on computer vision* (pp. 529-545). Springer, Cham.

[6] Farhadi, A., Hejrati, M., Sadeghi, M. A., Young, P., Rashtchian, C., Hockenmaier, J., & Forsyth, D. (2010, September). Every picture tells a story: Generating sentences from images. In *European conference on computer vision* (pp. 15-29). Springer, Berlin, Heidelberg.