IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)
An International Open Access, Peer-reviewed, Refereed Journal

Image Captioning Using Deep Learning Techniques

[1]Saraswathi P, [2]Dr. K. Suresh Babu
[1]MTech Student, [2]Associate Professor
School of Information Technology,
Jawaharlal Nehru Technological University, Hyderabad, India

Abstract: With the use of the internet, a lot of data is being collected in the form of images. This type of data can be used for further analysis or processing only if we are able to draw out what the image represents or depicts. Image captioning in simple words can be defined as the process of producing or generating a description in the form of text for an image. The results have many applications in real life. This paper presents an effort to leverage the benefits of the techniques of Deep Learning in producing the captions for images based on the objects present, the properties of the objects, the actions that are being performed by them and the interaction between the objects and with their surroundings using a CNN and RNN model.

***Index Terms* – Image Captioning, CNN, LSTM.**

I. Introduction

Generating captions for images is a vital task involving the dual techniques of computer vision and natural language processing. Computer vision to understand the content of the image and to turn the understanding into words using natural language processing's language models. Imitating the human capability of giving descriptions for images by a machine is itself a remarkable step, the main challenge is to capture how objects relate, interact to each other in the image and to express them in a natural language that has many applications. Few of them are: Aiding the visually impaired -Self Driving car Image indexing, surveillance camera. It has been seen that in recent times deep learning models have achieved state- of- the- art optimal results in this particular field. Single end-to-end model can be defined to generate the captions instead of a pipeline of models. The dataset used is Flickr8K dataset and to measure the model's performance, BLEU standard metric is used. Google Colaboratory, a product from Google research is used to run the model.

II. Literature Survey

Many methodologies have been proposed for the solution of image captioning as images have been a major sharing media on the internet since its inception. Krizhevsky et al. [1] classified 1.2million images into 1000 different classes by training a large, deep convolutional network. The deep network has 60 million parameters and 650,000 neurons, consisting of five convolutional layers some of which are followed by max-pooling layers and three fully connected layers with 1000-way SoftMax as the last layer. Zhao et al. [2] provided a review on different deep learning frameworks for object detection starting from pipeline models that form the base for other architectures. Several future directions have been proposed to gain an understanding on object detection. Karpathy and Fei Fei [3] presented a model that describes the images and their regions with the use of datasets of images and their descriptions by understanding the inner correlation of visual data and language. The model developed was a convolutional network over the image areas, a bidirectional recurrent network over sentences and aligning the two modalities through multi modal embedding. Aneja et al. [4] discussed a convolutional approach for image captioning and showed that it performs on par with existing LSTM techniques also compared with RNN based learning. Vinyals et al. [5] presented a model based on a deep recurrent architecture combining convolutional network and recurrent network to generate plain english descriptions for an image. Marc Tanti et al. [6] proposed a merge model where the hidden layer vector size shrunk up to four times in the RNN layer. Papineni et al. [7] proposed a method for evaluation of automatic machine translation - BLEU, Bilingual Evaluation which is quick and language independent.

III. Methodology, Implementation and Metrics:

i. Methodology:

The dataset adapted for this caption generation model is Flickr8K dataset among various datasets that are available- Flickr8K, Flickr30K, MS COCO, Pascal VOC datasets. The dataset description is stated in the paper “Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics” from 2013[2]. The images were selected from six different Flickr groups and mostly avoid any well-known people or locations to depict a variety of scenes and situations. The dataset contains 8092 images in jpeg format with text descriptions of the photographs. The dataset is divided into- 6000 images of training data, development and test dataset consist of 1000 images each.

The model has three main parts: a feature extractor where in the features of the image are extracted using a CNN model, sequence processor where the recurrent network is present to handle the input text and lastly a merger where the input from the previous layers is fed to get a prediction.

Feature Extractor:

The images of the Flickr8k dataset are fed to the convolution network to extract the features. The convolution network used in this model is VGG16, a 16-layer network-13 convolution layers and 3 fully connected layers. The dropout layers are used to reduce the overfitting and a dense layer at the end to reduce the 4096-vector representation of the image to 256 feature vector length. The output is fed to the Long Short-Term layer.

Sequence processor:

A 34-word length input sequence is fed to this model. The model has embedding layers where the padded values are ignored using a mask followed by a Long Short-Term Layer.

Merger:

Both the previous layers give a 256-vector length as output. These are fed to the merger layer where the inputs are combined using an additional operation which is then fed to the neuron layer and then a dense layer that produces the next word in the caption using SoftMax prediction from the vocabulary that was processed in the sequence processor part.

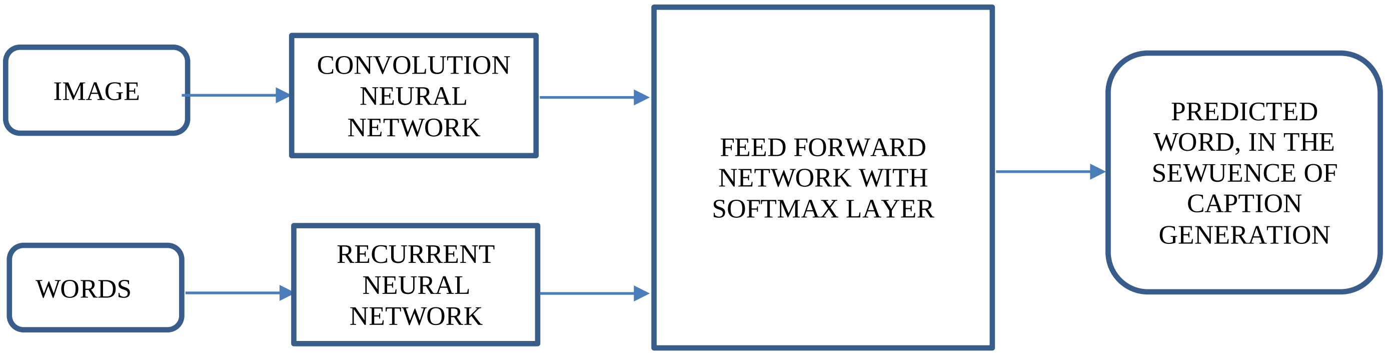


IMAGE
CONVOLUTION NEURAL NETWORK
WORDS
RECURRENT NEURAL NETWORK
FEED FORWARD NETWORK WITH SOFTMAX LAYER
PREDICTED WORD, IN THE SEWUENCE OF CAPTION GENERATION

Fig. 1 Architecture

ii. Implementation and Metrics

The model implementation was done using the Python environment and Keras and TensorFlow was used for the deep learning model implementation. Keras is a framework used for creating deep neural networks and TensorFlow is a library by Google works as background for Keras framework. The model was trained on Google Colaboratory Pro with GPU: 1xTesla K80 having 2496 CUDA cores, 12GB GDDR5 VRAM. The steps involved in the implementation are:

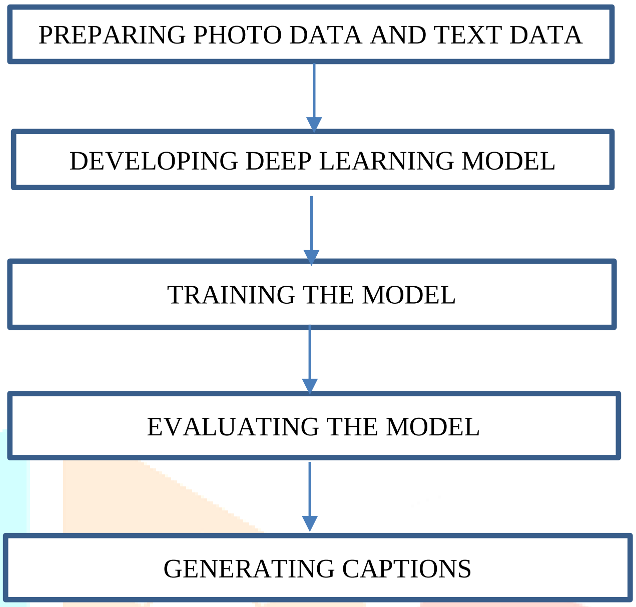

PREPARING PHOTO DATA AND TEXT DATA
DEVELOPING DEEP LEARNING MODEL
TRAINING THE MODEL
EVALUATING THE MODEL
GENERATING CAPTIONS

Fig. 2 Steps used in proposed model

In the first step, the image features are stored by extracting them using a pretrained VGG16 by removing the last layer provided by Keras framework and it is loaded, fed to the model to avoid the redundancy of running the whole dataset every time a new language model needs to be tested. Also, storing helps in real time implementation of the model. The captions are pre-processed so that the only data enrich content is present and are suffixed and prefixed with end and start tokens respectively so that the model can understand when it is started and generation to be ended. In second step, a dep learning model is developed for caption generation using CNN and RNN. The architecture of the proposed model can be seen in Figure 3.

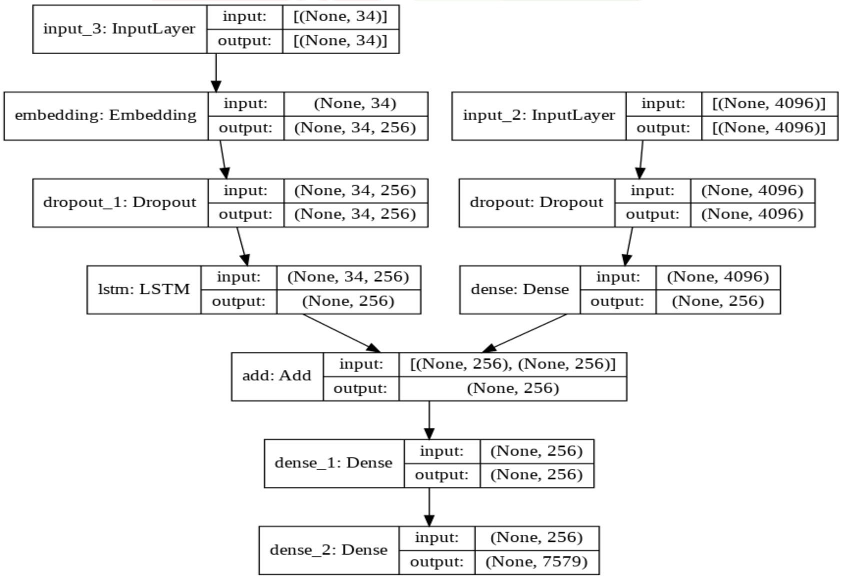
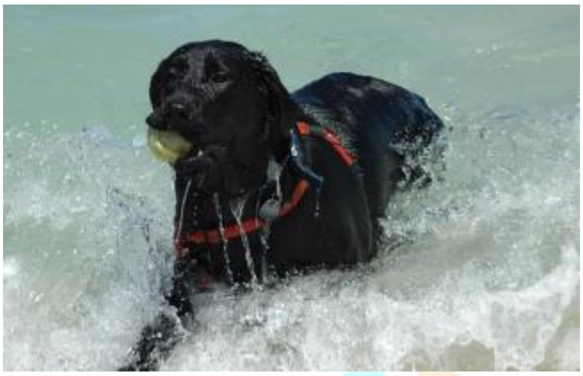


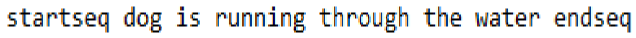
Fig. 3 Proposed Image Captioning Model Architecture

During the training phase the image from the training set of 6000 images is fed to the model. The Vgg16 part is trained to identify the objects in the image and the LSTM part to predict the words in the sequence. In the pre-processing text part of the model the captions. Later on, the model is tested on a testing dataset and the evaluation of the model is done by BLEU score. BLEU score is a Bilingual Evaluation Understudy used to evaluate the performance of machine translation. At the last phase, an image is given as input to the model where the output is predicted caption.

IV. RESULTS

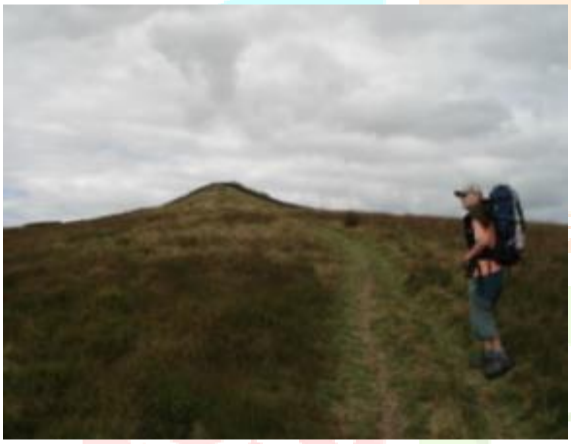
The model with best validation results is saved in a Hierarchical Data Format 5 file and later the model is trained, evaluated. And used for the caption generation. The average BLEU score of 47.8 is obtained for the validation set. The following are the results obtained corresponding to the input image:

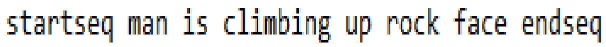




```
startseq dog is running through the water endseq
```

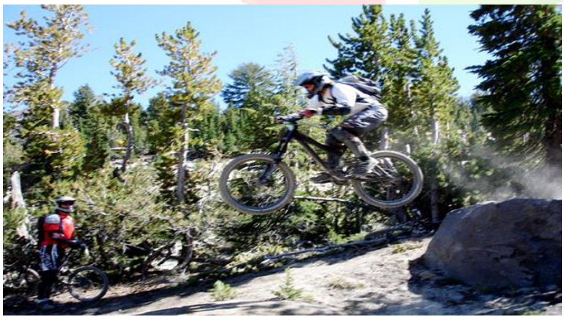
Fig. 4 Input image-1 and corresponding output from the model

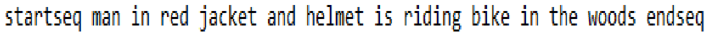




```
startseq man is climbing up rock face endseq
```

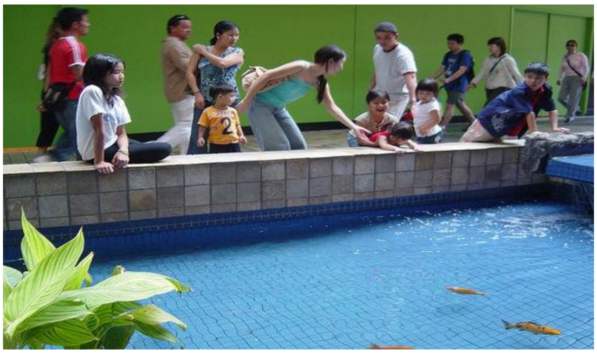
Fig. 5 Input image-2 and corresponding output caption generated by the model





```
startseq man in red jacket and helmet is riding bike in the woods endseq
```

Fig. 6 Input image -3 and corresponding output caption generated by the model



```
startseq man is sitting on the edge of the water endseq
```

Fig. 7 Input image -4 and corresponding output caption generated by the model

V. CONCLUSIONS AND FUTURE WORK

The deep learning techniques have been implemented to develop a model that can generate captions similar to human sentences for a given image. The proposed model able to predict the captions but not so accurately for few. The results can be improved by using larger GPUs, larger datasets and training the model on that. Further these techniques can be used for various applications as stated earlier like aiding visually impaired People, automated cars and so.

VI. REFERENCES

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, Networks, [Online] Available: https://papers.nips.cc/paper/4824-imagenetclassificationwith-deep-convolutional-Neural-networks.pdf

[2] Z. Zhao, P. Zheng, S. Xu and X. Wu, "Object Detection With Deep Learning: A Review," in *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 11, pp. 3212-3232, Nov. 2019, doi: 10.1109/TNNLS.2018.2876865.

[3] Andrej Karpathy, Li Fei-Fei, Deep Visual Semantic Alignments for Generating Image Descriptions, [Online] Available: https://cs.stanford.edu/people/karpathy/cvpr2015.pdf

[4] Jyoti Aneja, Aditya Deshpande, Alexander Schwing, Convolutional Image Captioning, [Online] Available: https://arxiv.org/pdf/1711.09151.pdf

[5] Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan, Show and Tell: A Neural Image Caption Generator, [Online] Available: https://arxiv.org/pdf/1411.4555.pdf

[6] TANTI, M., GATT, A., & CAMILLERI, K. (2018). Where to put the image in an image caption generator. *Natural Language Engineering,24*(3),467-489. doi:10.1017/S1351324918000098

[7] BLEU: A Method for Automatic Evaluation of Machine Translation Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu IBM T. J. Watson Research Center Yorktown Heights, NY 10598, USA