e-ISSN:2582-7219

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

Volume 5, Issue 6, June 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.54

International Journal Of Multidisciplinary Research In Science, Engineering and Technology (IJMRSET)

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.54|

| Volume 5, Issue 6, June 2022 |

| DOI:10.15680/IJMRSET.2022.0506018 |

Image Captioning - A Deep Learning Approach

B. Mythri[1], G. Lakshmi Priya[2], A. Akanksha[3], D. Harshitha[4]

Dept. of Electronics and Communications Engineering, Vasireddy Venkatadri Institute of Technology, Gundur, Andhrapradesh, India

ABSTRACT: We are very interested in how machines can automatically describe the content of images using human language. In order to gain a deeper insight of this computer vision topic, we decided to implement current state-of-the-art image caption generator Show, attend and tell: Neural image caption generator with visual attention [12]. Our neural network based image caption generator is implemented in Python powered by Pytorch machine learning library. We have identified five major components in our pipeline: (R1) data prepossessing; (R2) Convolutional Neural Network (CNN) as an encoder; (R3) attention mechanism; (R4) Recurrent Neural Network (RNN) as a decoder; (R5) Beam Search to find most optimal caption; (R6) Sentence Generation and evaluation. BLEU-4 score is picked for evaluating the quality and accuracy of the generated caption. We evenly distributed the five components described above among our group and each member has made equal contributions to push the project forward. We have successfully finished the implementation of the all five components and are able to train our network on Google Colab (which provides free GPU resources). Our implementation of this image caption generator has achieved a very decent accuracy quantified by BLEU-4 score (15.5) which is very close to the result reported in the original paper (18.5). As we finished training the network and obtained a satisfying performance, we continue to visualize the attention mechanism.

I. INTRODUCTION

Caption generation is an interesting artificial intelligence problem where a descriptive sentence is generated for a given image. It involves the dual techniques from computer vision to understand the content of the image and a language model from the field of natural language processing to turn the understanding of the image into words in the right order. Image captioning has various applications such as recommendations in editing applications, usage in virtual assistants, for image indexing, for visually impaired persons, for social media, and several other natural language processing applications. Recently, deep learning methods have achieved state-of-the-art results on examples of this problem. It has been demonstrated that deep learning models are able to achieve optimum results in the field of caption generation problems. Instead of requiring complex data preparation or a pipeline of specifically designed models, a single end-to-end model can be defined to predict a caption, given a photo. In order to evaluate our model, we measure its performance on the Flickr8K dataset using the BLEU standard metric. These results show that our proposed model performs better than standard models regarding image captioning in performance evaluation.

II. RELATED WORK

The image captioning problem and its proposed solutions have existed since the advent of the Internet and its widespread adoption as a medium to share images. Numerous algorithms and techniques have been put forward by researchers from different perspectives. Krizhevsky et al. [1] implemented a neural network using non-saturating neurons and a very efficient a unique method GPU implementation of the convolution function. By employing a regularization method called dropout, they succeeded in reducing overfitting. Their neural network consisted of maxpooling layers and a final 1000-way softmax. Deng et al. [2] introduced a new database which they called ImageNet; an extensive collection of images built using the core of the WordNet structure. ImageNet organized the different classes of images in a densely populated semantic hierarchy. Karpathy and FeiFei [3] made use of datasets of images and their sentence descriptions to learn about the inner correspondences visual data and language. Their work described a Multimodal Recurrent Neural Network architecture that utilises the inferred co-linear arrangement of features in order to learn how to generate novel descriptions of images. Yang et al. [4] proposed a system for the automatic generation of a natural language description of an image, which will help immensely in furthering image understanding. The proposed multimodel neural network method, consisting of object detection and localization modules, is very similar to the human visual system which is able to learns how to describe the content of images

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.54|

| Volume 5, Issue 6, June 2022 |

| DOI:10.15680/IJMRSET.2022.0506018 |

automatically. In order to address the problem of LSTM units being complex and inherently sequential across time, Aneja et al. [5] proposed a convolutional network model for machine translation and conditional image generation. Pan et. al [6] experimented extensively with multiple network architectures on large datasets consisting of varying content styles, and proposed a unique model showing noteworthy improvement on captioning accuracy over the previously proposed models. Vinyals et al. [7] presented a generative model consisting of a deep recurrent architecture that leverages machine translation and computer vision, used to generate natural descriptions of an image by ensuring highest probability of the generated sentence to accurately describe the target image. Xu et al. [8] introduced an attention based model that learned to describe the image regions automatically. The model was trained using standard backpropagation techniques by maximizing a variable lower bound. The model was able to automatically learn identify object boundaries while at the same time generate an accurate descriptive sentence.

III. DATASET AND EVALUATION METRICS

In deep learning ms coco dataset fli8k 30k are most common used. Most common of them are Pascal VOC dataset, Flickr 8K and MS-COCO Dataset. Flickr 8K Image captioning dataset [9] is used in the proposed model. Flickr 8K is a dataset consisting of 8,092 images from the Flickr.com website. This dataset contains collection of day-to-day activity with their related captions. First each object in image is labeled and after that description is added based on objects in an image.

We devide 8,000 images into three disjoint sets. The training data (DTrain) has 6000 images whereas the development and test dataset consist of 1000 images each.

In order to evaluate the image-caption pairs, we need to evaluate their ability to associate previously unseen images and captions with each other. The evaluation of model that generates natural language sentence can be done by the BLEU (Bilingual Evaluation Understudy) Score. It describes how natural sentence is compared to human generated sentence.

Our model to caption images are built on multimodal recurrent and convolutional neural networks. A Convolutional Neural Network is used to extract the features from an image which is then along with the captions is fed into an Recurrent Neural Network. The architecture of the image captioning model is shown in figure 1.

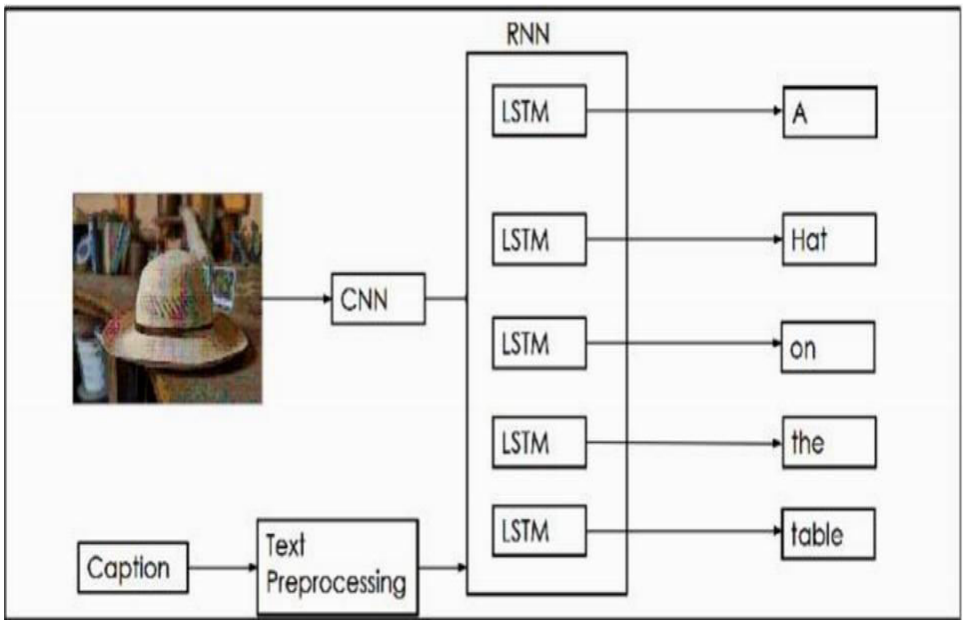


Figure 1. Architecture

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.54|

| Volume 5, Issue 6, June 2022 |

| DOI:10.15680/IJMRSET.2022.0506018 |

The model consists of 5 phases:

A . Data Set Collection

There are many data sets which can be used for training the deep learning model for generating captions for the images like ImageNet, COCO, FLICKR 8K, FLICKR 30K. We are using FLICKR 8K data set for training the model. FLICKR 8K dataset works efficiently for training the Image Caption Generating Deep Learning Model.The FLICKR 8K data set consists of 8000 images in which 6000 images can be used for training the deep learning model and 1000 images for development and 1000 images for testing the model. Flickr Text data set consists of five captions for each given image which describes about the actions performed in the given images.

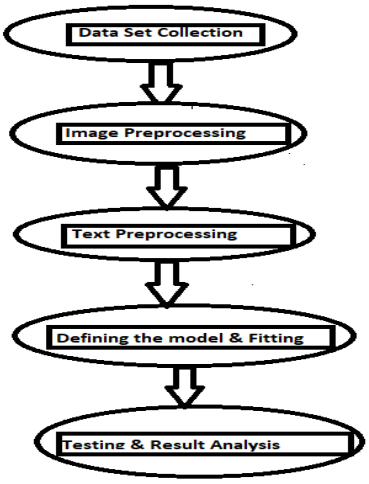


Figure 2. Model Implementation

B . Image Preprocessing

After loading the data sets we need to preprocess the images in order to give these images as input. As we cannot pass different sized images through the Convolution layer we need to resize every image so that they are in same size i.e; 224X224X3 .We are also converting the images to RGB by using inbuilt functions of cv2 library.

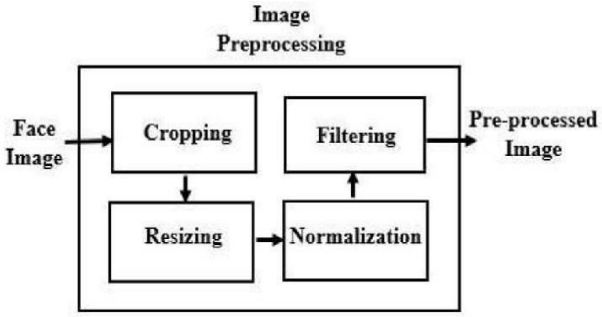


Figure 3. Image Preprocessing

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.54|

| Volume 5, Issue 6, June 2022 |

| DOI:10.15680/IJMRSET.2022.0506018 |

C . Text Preprocessing

After loading the captions for the images using FLICKR text data set we need to preprocess those captions so that there is no ambiguity or difficulty while generating vocabulary from the captions and also while training the deep learning model. We need to change all the uppercase letters in the captions to the lower case in order to eliminate ambiguity during vocabulary building and training of the model. As this model will generate captions one word at a time and previously generated words are used as inputs along with the image features as input , <start seq> and <end seq> are attached at the starting and end of each of the caption to signal the neural network about the starting of the caption and ending of the captions during the training and testing of the model.

D . Defining and Fitting the Model

After collecting the data set and preprocessing the images and captions and building vocabulary. We have to define the model for generation of captions. These are the steps followed for fitting the model: Model Training

Predictive Model

Model Implementation

IV. TRAINING PHASE

During training phase we provide pair of input image and its appropriate captions to the image captioning model. The VGG model is trained to identify all possible objects in an image. While LSTM part of model is trained to predict every word in the sentence after it has seen image as well as all previous words. For each caption we add two additional symbols to denote the starting and ending of the sequence. Whenever stop word is encountered it stops generating sentence and it marks end of string.

Loss function for model is calculated as, where I represents input image and S represents the generated caption. N is length of generated sentence. pt and St represent probability and predicted word at the time t respectively. During the process of training we have tried to minimize this loss function.

$$L(I, S) = -\sum_{t=1}^{N} \log p_t(S_t)$$

V. IMPLEMENTATION

The implementation of the model was done using the Python SciPy environment. Keras 2.0 was used to implement the deep learning model because of the presence of the VGG net which was used for the object identification. Tensorflow library is installed as a backend for the Keras framework for creating and training deep neural networks. TensorFlow is a deep learning library developed by Google. It provides heterogeneous platform for execution of algorithms i.e. it can be run on low power devices like mobile as well as large scale distributed system containing thousands of GPUs. The neural network was trained on the NvidiaGeforce 1050 graphics processing unit which has 640 Cuda cores. In order to define structure of our network TensorFlow uses graph definition. Once graph is defined it can be executed on any supported devices. The photo features are pre-computed using the pretrained model and saved. These features are then loaded and them into our model as the interpretation of a given photo in the dataset to reduce the redundancy of running each photo through the network every time we want to test a new language model configuration. The preloading of the image features is also done for real time implementation of the image captioning model. The architecture of the model is shown in Figure 2.

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.54|

| Volume 5, Issue 6, June 2022 |

| DOI:10.15680/IJMRSET.2022.0506018 |

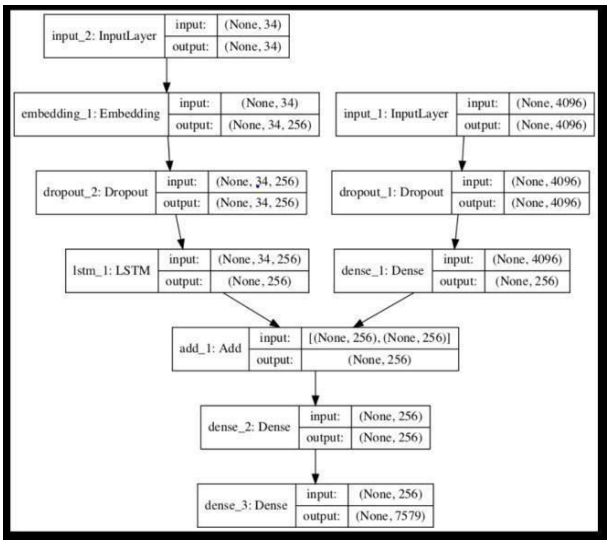


Figure 4. Image Captioning Model

VI. RESULTS AND COMPARISION

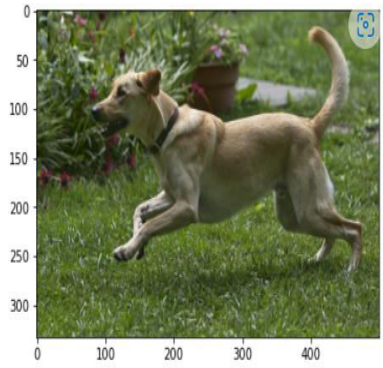
The image captioning model was implemented and we were able to generate moderately comparable captions with compared to human generated captions. The VGG net model first assigns probabilities to all the objects that are possibly present in the image, as shown in Figure 3. The model converts the image into word vector. This word vector is provided as input to LSTM cells which will then form sentence from this word vector. The generated sentences are shown in Fig 4. Generated sentence are black dog runs into the ocean next to a rock, while actual human generated sentences are black dog runs into the ocean next to a pile of seaweed., black dog runs into the ocean, a black dog runs into the ball, a black dog runs to a ball. This results in a BLEU score of 57 for this image. Similarly in Fig 5. Generated Sentence is 'A man wearing black shirt is standing in ice'. While calculating BLEU score of all image in validation dataset we get average score of 60.1 , Which shows that our generated sentence are very similar compared to human generated sentence.


```
Referance Captions:
A man be wear a Sooner red football shirt and helmet .
A Oklahoma Sooner football player wear his jersey number 28 .
A Sooner football player weas the number 28 and black armband .
Guy in red and white football uniform
The American footballer be wear a red and white strip .
Predicted Caption:
A football player in a red jersey .
bleu score:  1.0905741632475476e-154
```

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.54|

| Volume 5, Issue 6, June 2022 |

| DOI:10.15680/IJMRSET.2022.0506018 |




```
Referance Captions:
A brown dog run
A brown dog run over grass .
A brown dog with its front paw off the ground on a grassy surface near red and purple flower .
A dog run across a grassy lawn near some flower .
A yellow dog be play in a grassy area near flower .
Predicted Caption:
A brown dog run through a field .
bleu score:  0.38260294162784475
```

Figure 5. Input and Output

VII. CONCLUSION

In this paper, we have reviewed deep learning-based image captioning methods. We have given a taxonomy of image captioning techniques, shown generic block diagram of the major groups and highlighted their pros and cons. We discussed different evaluation metrics and datasets with their strengths and weaknesses. A brief summary of experimental results is also given. We briefly outlined potential research directions in this area. Although deep learning-based image captioning methods have achieved a remarkable progress in recent years, a robust image captioning method that is able to generate high quality captions for nearly all images is yet to be achieved. With the advent of novel deep learning network architectures, automatic image captioning will remain an active research area for some time. We have used Flickr_8k dataset which includes nearly 8000 images, and the corresponding captions are also stored in the text file. Although deep learning - based image captioning methods have achieved a remarkable progress in recent years, a robust image captioning method that is able to generate high quality captions for nearly all images is yet to be achieved. With the advent of novel deep learning network architectures, automatic image captioning will remain an active research area for sometime. The scope of image-captioning is very vast in the future as the users are increasing day by day on social media and most of them would post photos. So this project will help them to a greater extent.

REFERENCES

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

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li and Li Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database

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

[4] Zhongliang Yang, Yu-Jin Zhang, SadaqaturRehman, Yongfeng Huang, Image Captioning with Object Detection and Localization, [Online] Available: https://arxiv.org/ftp/arxiv/papers/1706/1706.02430.p df

[5] JyotiAneja, AdityaDeshpande, Alexander Schwing, Convolutional Image Captioning, [Online] Available: https://arxiv.org/pdf/1711.09151.pdf

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.54|

| Volume 5, Issue 6, June 2022 |

| DOI:10.15680/IJMRSET.2022.0506018 |

[6] Jia-Yu Pan, Hyung-Jeong Yang, Pinar Duygulu, Automatic Image Captioning, Conference: Conference: 0XOWLPHGLDDQG([SR,&0(¶ 2004 IEEE International Conference on, Volume: 3

[7] OriolVinyals, Alexander Toshev, SamyBengio, DumitruErhan, Show and Tell: A Neural Image Caption Generator, [Online] Available: https://arxiv.org/pdf/1411.4555.pdf

[8] Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, RuslanSalakhutdinov, Richard S. Zemel, YoshuaBengio, Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, [Online] Available: https://arxiv.org/pdf/1502.03044.pdf [9] M. Hodosh, 3<RXQJDQG-+RFNHQPDLHU´)UDPLQJ,PDJH Description as a Ranking Task: Data, Models and (YDOXDWLRQ0HWULFV´-RXUQDORI$UWLILFial Intelligence Research, Volume 47, pages 853-899

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

l

Impact Factor
7.54

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com