Visual Captioning Using LSTM

Sudarshan Kumar Prajapati[1], Shubham Saini[2], Stuti Gupta[3],Sparsh Verma[4],Ms .Parul Sharma[5]

[1]Student, Computer Science Department, RKGIT, Ghaziabad, Uttar Pradesh, India
[2]Student, Computer Science Department, RKGIT, Ghaziabad, Uttar Pradesh, India
[3]Student, Computer Science Department, RKGIT, Ghaziabad, Uttar Pradesh, India
[4]Student, Computer Science Department, RKGIT, Ghaziabad, Uttar Pradesh, India
[5]Assistant Professor, Computer Science Department, RKGIT, Ghaziabad, Uttar Pradesh, India

---

Abstract –Computer Vision has become ubiquitous in every society, with applications in several fields. Inthis project, the main focus is on one of the visual recognition facets of computer vision, i.e image captioning. The problem of generating languagedescriptions for visual data has been studied for a long time but in the field ofvideos. In the recent few years emphasis has been lead still on image descriptions with natural text. Due to the recent advancements in the field of object detection,the task of scene description in an image has become easier.The aim of the project is to train convolutional neural networks with several hundreds of hyper-parameters and apply it on a huge data-set of images (Image-Net),and combine the results of this image classifier with a recurrent neural network togenerate a caption for the classified image. In this report, the detailed architectureof the model developed is being presented.

Key Words:Deep Learning, CNN, LSTM, VGG-16, Flickr30K,

1.INTRODUCTION

Artificial Intelligence(AI)[6] is now at the heart of the innovation economy and thus thebase for this project is also the same. In the recent past, a field of AI namely DeepLearning[3] has turned a lot of heads due to its impressive results in terms of accuracywhen compared to the already existing Machine learning algorithms[2]. The task of beingable to generate a meaningful sentence from an image is a difficult task but can havegreat impact, for instance helping the visually impaired to have a better understandingof images.

The task of image captioning is significantly harder than that of image classification,which has been the main focus in the computer vision[4] community. A description ofan image must capture the relationship between the objects in the image. In addition tothe visual understanding of the image, the above semantic knowledge has to be expressedin a natural language like English, which means that a language model is needed. Theattempts made in the past have all been to stitch the two models together.

In the model proposed in the project, the approach is to combine this into a singlemodel which consists of a Convolutional Neural Network (CNN) encoder[7] which helps increating image encodings. Here VGG16 architecture[7] is used with some modifications.

In this project some of the recent and advanced classification architectures could havebeen used but that would have increased the training time significantly. These encoded images are then passed to an LSTM network[9] which is a type of Recurrent NeuralNetwork[9]. The network architecture used for the LSTM network work in a similarfashion as the ones used in machine translators. The input to the network is an image
which is first converted into a 224*224 dimension. Flickr 8k dataset is used to train themodel. The model outputs a generated caption based on the dictionary it forms from thetokens of caption in the training set.

In the project,it's tried to train a simple Convolutional Neural Network and achievedecent results within a few hours of training. Thus, by the end of this section,it isconcluded that CNN is a good fit to be used as the image encoder for the captioningmodel. The BLEU score metric is used to compare the accuracy of the model proposedwith the ones already present. At the end,a few examples tested on the model arereported.

1.1 Purpose

Image captioning is very important to real world scenario. Let's see few applicationswhere this can be very useful.

- Self Driving cars - if it can properly caption the scene around the car, it can giveboost to self driving system.
- Aid to blind - People who can not see the scene around them can at least listen thevoice generated for captions.
- CCTV cameras are everywhere today, but if it can generate the relevant captions,then an alarm can be raised as soon as there is some malicious activity going onsomewhere.

1.2 Scope

The scope of this software is limited. It will work only on the system in which it is installed.

2. Methodology

For the task of image captioning, the first thing is to determine a fit model for the task of encoding the image. Deep learning concepts are being used in this project. VGG-16 model is used as CNN, and LSTM as RNN part.

2.1 Feature Extraction Using CNN

Convolutional Neural Networks (ConvNets or CNNs) are a category of Artificial Neural Networks which have proven to be very effective in the field of image recognition and

classification. They have been used extensively for the task of object detection, self- driving cars, image captioning, etc. Here VGG-16 architecture is being used to extract the image features and for training the model.
A basic convnet is shown in the fig. below:

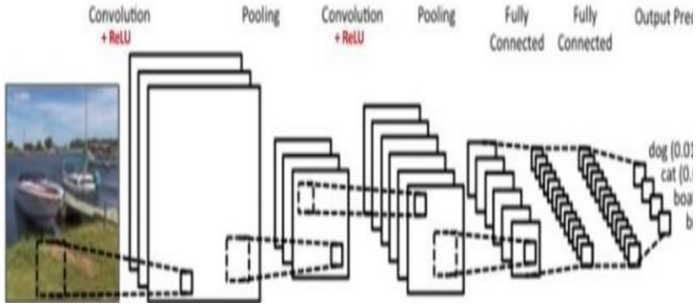

Fig-1: A simple convnetarchitecture

In this project some of the recent and advanced classification

2.2 Dataset

For the task of image captioning, Flickr 8k dataset is used. The dataset contains 8000 images with 5 captions per image. The dataset by default is split into image and text folders.Each image has a unique id and the caption for each of these images is stored corresponding to the respective id. Thedataset contains 6000 training images, 1000 developmentimages, and 1000 test images. A sample from the data is given in below figure. Other datasets like Flickr30k and MSCOCO for image captioning exist but both thesedatasets have more than 30,000 images thus processing them becomescomputationallyvery expensive. Captions generated usingthese datasets may prove to be better than theones generated after training on Flickr8k because the dictionary of words used by RNNdecoder would be larger in the case of Flickr30k and MS COCO.

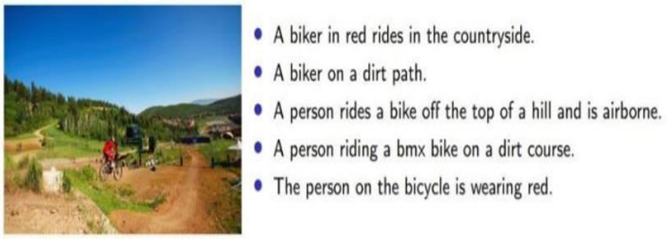

Fig-2:Sample image and corresponding captions from the Flickr8k dataset

2.3 VGG-16 architecture

Imagenet Large Scale Visual Recognition Competition (ILSVRC) have provided various open-source deep learning frameworks like ZFnet, Alexnet, Vgg16, Resnetetc have shown great potential in the field of image classification. For the task of image encoding in this model, Vgg16 is used which is a 16- layered network proposed in ILSVRC 2014[7]. VGG16 significantly decreased the top-5 error rate in the year 2014 to 7.3%.

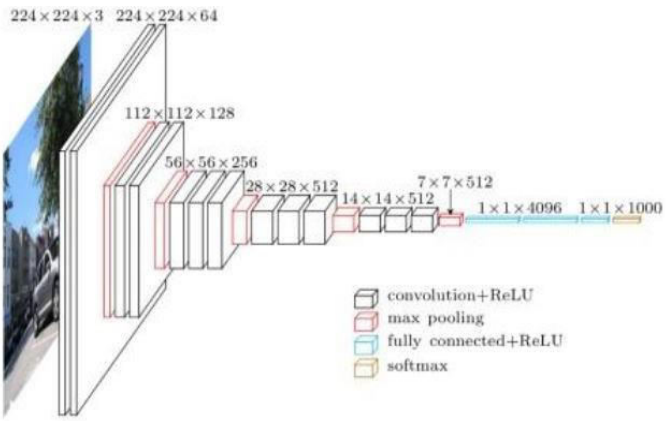

Fig-3: VGG16 architecture[7]

The convolution layer consists of 3*3 filters and the stride length is fixed at 1. Max pooling is done using a 2*2-pixel window with a stride length of 2. All the images need to be converted into a 224*224-dimensional image. A Rectified Linear Unit (ReLU)
activation function follows every convolution layer. The output of the ReLU function is given below:

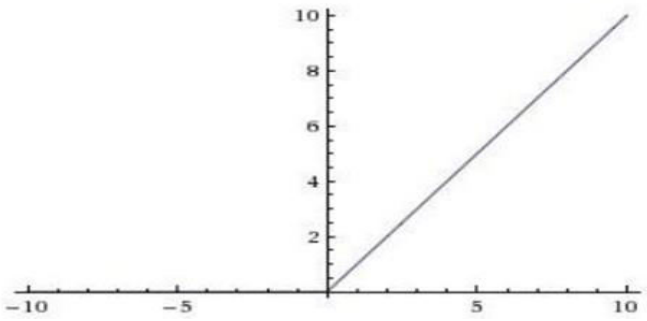

Fig-4:Rectified linear unit activation function [7]

3. CONCLUSIONS

This end-to-end system neural network system is capable of viewing an image and generating a reasonable description in English depending on the words in its dictionarygenerated on the basis of tokens in the captions of train images. The model has a convolutional neural network encoder and an LSTM decoder that helps in the generation ofsentences. The purpose of the model is to maximize the likelihood of the sentence giventhe image.

ACKNOWLEDGEMENT

We should like to acknowledge our mentor Ms.ParulSharma– Assistant Professor, Computer Science Department, RKGIT, Ghaziabad, Uttar Pradesh, India

REFERENCES

1 Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Doll´ar, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1473–1482, 2015.

2 SonuPratap Singh Gurjar, Shivam Gupta, and Rajeev Srivastava. Automatic image an- notation model using lstm approach.

3 Xin Jia. Image recognition method based on deep learning. In *2017 29th Chinese Control And Decision Conference (CCDC)*, pages 4730–4735. IEEE, 2017.

4 Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localiza- tion networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4565–4574, 2016.

5 Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recog- nition*, pages 3128–3137, 2015.

6 Stephen Lucci and Danny Kopec. *Artificial intelligence in the 21st century*. Stylus Pub- lishing, LLC, 2015.

7 Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

8 Oriol Vinyals, Alexander Toshev, SamyBengio, and DumitruErhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

9 Cheng Wang, Haojin Yang, Christian Bartz, and Christoph Meinel. Image captioning with deep bidirectional lstms. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 988–997. ACM, 2016.

10 Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, RuslanSalakhudi- nov, Rich Zemel, and YoshuaBengio. Show, attend and tell: Neural image caption gen- eration with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.