International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue III, March 2018- Available at www.ijraset.com

Automated Image Captioning Using ConvNets and Recurrent Neural Network

Prof. P. Singam[1], Ashvini Bhandarkar[2], Bhavana Mahalle[3], Chetna Tule[4], Kiran Kewate[5]
[1, 2, 3, 4, 5,] *Department of Computer Technology, KDK College of Engineering, Nagpur(India)*

Abstract: ***We present a model that generates free-form natural language descriptions of image regions. Our model leverages datasets of images and their sentence descriptions to learn about the inter-modal correspondences between text and visual data. Our approach is based on a novel combination of Convolutional Neural Networks over image regions, Recurrent Neural Networks over sentences, and a structured objective that aligns the two modalities through a multimodal embedding. We then describe a Recurrent Neural Network architecture that uses the inferred alignments to learn to generate novel descriptions of image regions. We introduce a system to automatically generate natural language descriptions from images that takes an input image and generates its description in text. It also generates descriptions that are notably more true to the specific image content than previous work.***
Keywords: Recurrent Neural Network, Convolution Neural Network, datasets.

I. INTRODUCTION

An artificial neural network is an interconnected group of nodes, as like vast network of neurons in a brain. Neural networks have been used on a variety of tasks, including computer vision, speech recognition, machine translation, social network filtering, playing board and video games, medical diagnosis and in many other domains. Internet being source of images, it is very important to understand these images. Over flowing visual data on Internet motivate us to understand image. Image being 3D collection of numbers, it makes easy to understand it. There is lot of work going on in computer vision field on detecting, recognizing the things in a image. We humans can easily recognize an image what there in that but giving same ability to machine is what we are trying to do here. Recognizing what is there in image is 2 step process. Semantics help us to make a sentence from an image. In this project we try to make system understand the semantics in an image. To achieve this, we make use of several deep learning models. [1] By combining deep Convolution neural network for image classification with recurrent neural network for sequence modelling, create a single network that generates descriptions of images. Our proposed model helps use to identify the content in an image.[2] Convolution neural network(CNN) and Recurrent neural network(RNN) helps us to do so. Image of size 32*32*3 is given as input to CNN which in turn produce a feature vector which is given as input to RNN to generate sequence of words called sentences. [6]We merged RNN and CNN. Some 'n' frames of video are taken at a time in batch. All these 'n' frames are feed to independent CNN simultaneously. [8] Input of each layer of each CNN is fused with input of same layer of next CNN in sequence. Since we are using neural network, [6] accuracy of our model is better than image classification algorithms. Here We are using IAPR12 dataset which contains 20000 images from all over the world.

II. RELATED WORK

A. Literature Review

We discuss below some of the approaches used for image captioning:
Szegedy et al [2] proposed a 22 layer CNN, which was state of art until 2014. Figure 2.1 shows basic building block of Google Net. Here a single layer is composed of fusion of 3 convolution layers each with kernel size as 1x1Ren et al [4] currently modelled a state of art model for object detection, called Faster RCNN. It uses Region of Interest (ROI) pooling layer in order to get a set of regions in images where probability of object being present is high. Later on each of these regions data augmentation is applied forming a set of cropped portion of images over which CNN model would be run. CNN model would have to bottleneck layers. One for object classification using a softmax classifier and other for predicting bounding box co-ordinates of the object using a regression model.
Ballas et al. [6], merged RNN and CNN. Some 'n' frames of video are taken at a time in batch. All these 'n' frames are feed to independent CNN simultaneously. Input of each layer of each CNN is fused with input of same layer of next CNN in sequence.

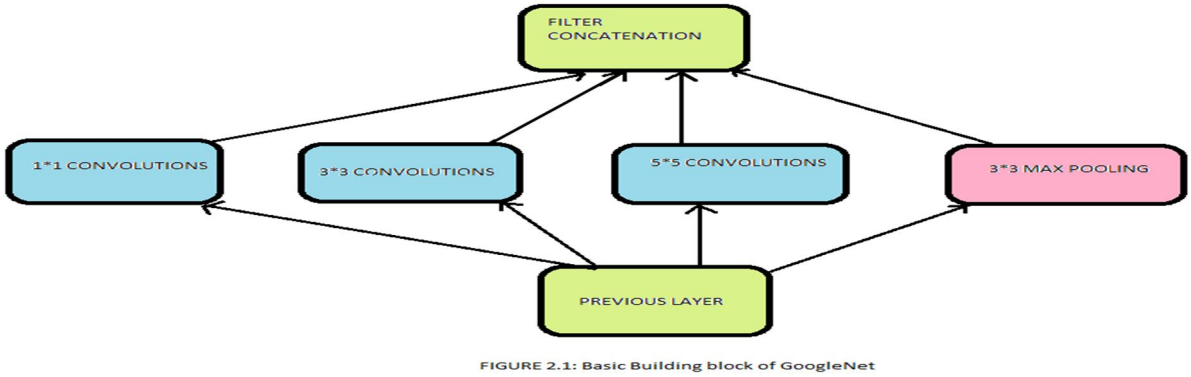

Figure 2.1: Basic building block of GoogleNet

arpathy et al [7] proposed a Slow Fusion technique wherein multiple frames are converged gradually till the final layer. However, their wasn't much improvement in accuracy. For single frame it was 59.3% and for slow fusion it was 60.9%.
Simonyan and Zisserman [8] feed video frames to a CNN and optical flow images of those images to another CNN. Finally, the output of both the CNN are fused together to produce single output. This method demonstrated improvement in accuracy Baccouche et al [9] used a 3D-CNN and stacked a LSTM on top of it.
Simply recognizing scene is insufficient as contextual information. In order to get some insight into context in Yatskar et al [1] tried to provide structured summary of image by defining attributes like agent, place, source, tool and item. They used Frame Net2.2 Problem Statement By combining deep Convolution neural network for image classification with recurrent neural network for sequence modelling, create a single network that generates descriptions of images.

B. Research Objectives

1) Image classification using inception V3 to generate feature vector of standard dataset images.
2) Generating sequences of words i.e. generation of sentences using RNN.
3) Combining both models to get description of an image in sentence form.

III. METHODOLOGY

A. Overview of steps involved in this approach

1) Input to the system would be an image which is system interpret as 3D matrix containing pixel values from 0 to 255
2) Image is then passed through many layers like convol, ReLu and max pooling
3) In Fully connected layer, feature vector is obtained and accordingly there values are generated.
4) nstead of using soft max classification to find out the loss, this feature vector us then pass on to RNN
5) Using this feature vector, sequence of sentence is generated through different hidden layers
6) Finally, loss is calculated and back propagation is done till desire output is generated.

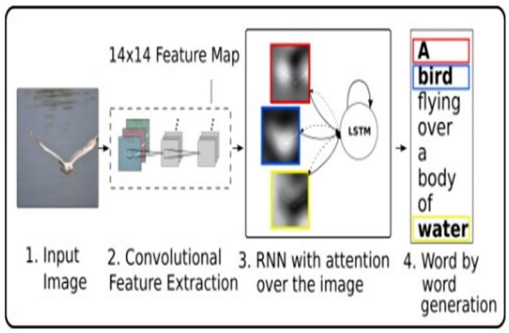

Figure 3.1: Overview of architecture of the system

B. Image Classification and CNN

Image is collection of pixels whose values are in between 0-255.System treat this image as 3D matrix. There is no specific algorithm to decide how to convolute this image and what should be the parameters. Filter of some size is taken to convolute over the image. This feature matrix has to pass through different layers of convol, Relu, Pooling is as shown in fig 3.2:

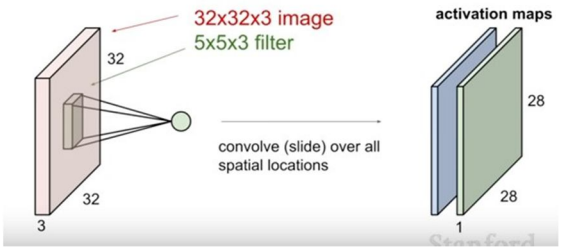

Figure 3.2: CNN Initial Step

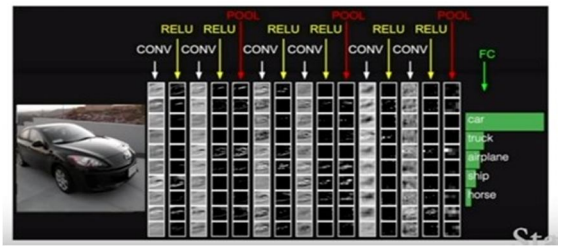

Figure 3.3: Fully connected layer contains neurons that connect to the entire input volume as in ordinary neural network.

C. Recurrent Neural Network and LSTM

We use RNN (Recurrent Neural Network) for character level language modelling. In character level language modelling we train RNN to predict next character in sequence on a training data. We use one-hot representation to represent a character in vector form. RNN maintains a state internally. It may receive an input vector x and may update weights and/or change its state and may output a vector at certain time steps. Figure 3.6 shows architecture of RNN for character level language modelling.

$$h_t = f_w(h_t{-}1, x_t) \quad (3.1)$$

$$h_t = tanh(W_{hh} * h_t{-}1 + W_{xh} * x_t) \quad (3.2)$$

$$yt = Why * ht \quad (3.3)$$

t: time step**,** **hr**: vector representing hidden state at time 't'**,** Xt: Input vector received at time step t**,** W_{hh}: Weight between hidden layer at two consequent time intervals**,** W_{hy}: Weight between hidden layer and output layer, W_{hx}: Weight between hidden layer and input layer.

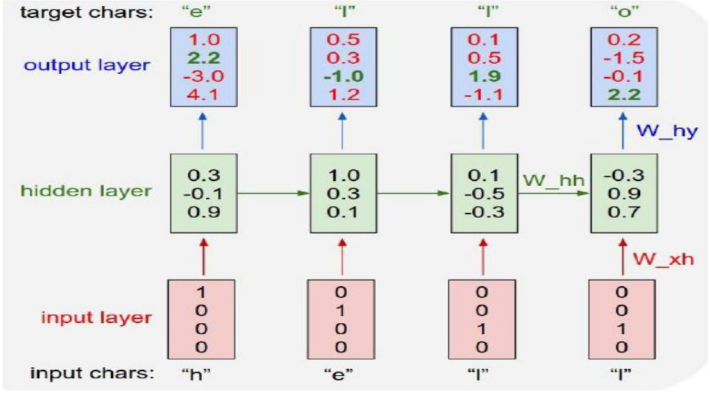

Figure 3.4 RNN for character level modelling.

RNN suffers from problem of vanishing gradients. This happens when epochs are h $\begin{pmatrix} i \\ f \\ o \\ g \end{pmatrix} = \begin{pmatrix} \text{sigm} \\ \text{sigm} \\ \text{sigm} \\ \tanh \end{pmatrix} W^l \begin{pmatrix} h_t^{l-1} \\ h_{t-1}^l \end{pmatrix}$ propagation is

high. To mitigate with this problem a variation of RNN, LSTM is use in our model. (3.4)

$ct = f * ct\text{-}1 + i * g$ (3.5) $ht = O * tanh(ct)$ (3.6)

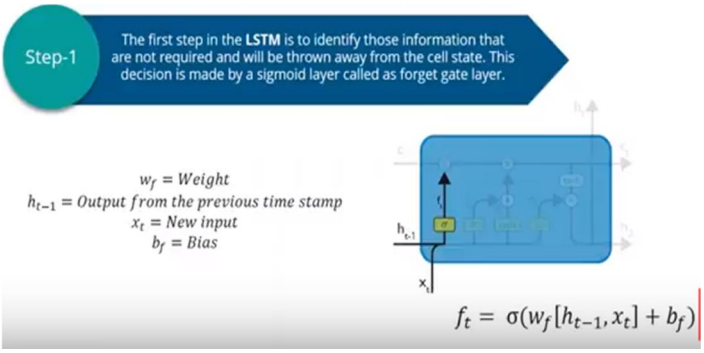

Figure 3.5: Step 1 of LSTM

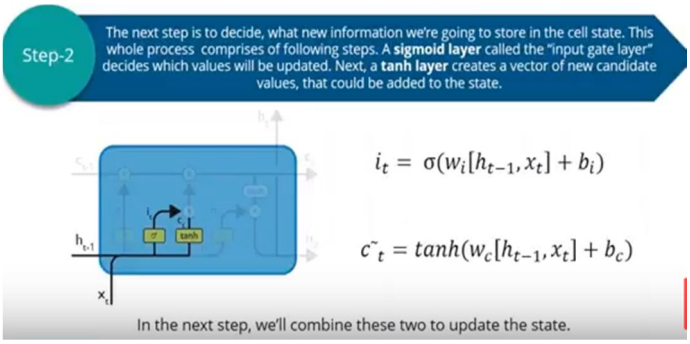

Figure 3.6: Step 2 of LSTM

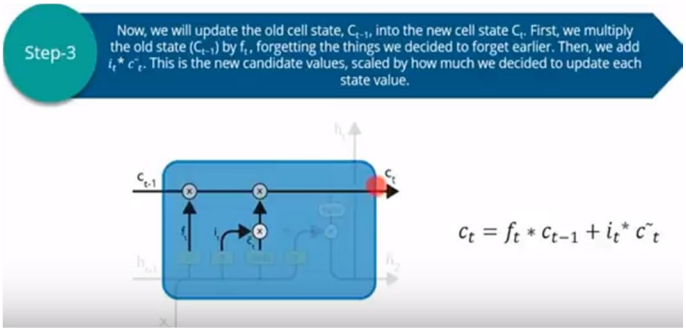

Figure 3.7: Step 3 of LSTM

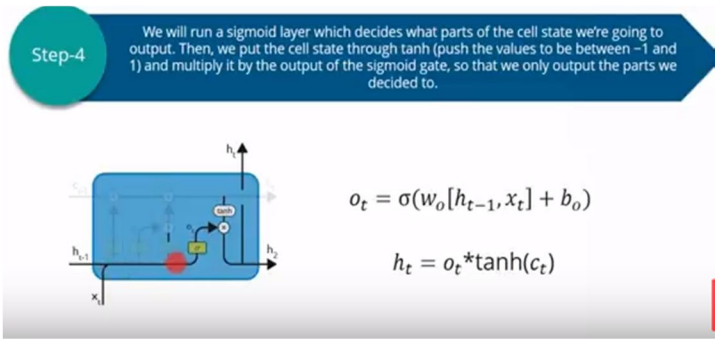

Figure 3.8: Step 4 of LSTM

For CNN since we are using inceptionV3 model, there are 32 layers of this convol, Relu and pooling. Max pooling is used for down sampling. Filter of 3*3 along with stride value of 1 is considered for first trial and based on the performance, these hyper parameters are changed.

Figure 3.9: Flow diagram

After generating different matrices of RGB, these matrices are mapped into single feature vector in order to get desire output. Fully connected layer is decision making layer. since here we are not doing any kind of object detection hence no image classifier is require to calculate the loss value Feature vector is directly input to RNN to get the sequence of words. After passing through different layers, loss is calculated. based on these loss values, back-propagation is done and gradient is calculated for each neuron in computational graph. Approximately after 50 such epoch we get desire output.

IV. CONCLUSION AND FUTURE WORK

A. *Conclusion*

We introduced a model that generates natural language descriptions of image regions based on weak labels in form of a dataset of images and sentences, and with very few hard coded assumptions. Our approach features a novel ranking model that aligned parts of visual and language modalities through a common, multi-modal embedding. We showed that this model provides state of the art performance on image-sentence ranking experiments. Second, we described a Multi-modal Recurrent Neural Network architecture that generates descriptions of visual data. We evaluated its performance on both full frame and region level experiments and showed that in both cases the Multi-modal RNN out performs retrieval baselines.

B. *Future Work*

Here we have deal with describing an image. Same model can also be use in other way round i.e. giving sentence and getting an image in output. We can also use this models for visual question answering which can be a good model for future reference.

BIBLIOGRAPHY

[1] O.Russakovsky,J.Deng,H.Su,J.Krause,S.Satheesh,S.Ma,Z.Huang,A.Karpathy,A.Khosla,M.Bernstein, A. C. Berg, and L. Fei-Fei, “ImageNet Large Scale Visual Recognition Challenge,” International Journal of Computer Vision (IJCV), vol. 115, no. 3, pp. 211–252, 2015.
[2] Szegedy, Christian, et al.” Going deeper with convolutions.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.
[3] Tran, Du, et al.” Learning spatiotemporal features with 3d convolutional networks.” arXiv preprint arXiv:1412.0767 (2014).
[4] Ren, Shaoqing, et al.” Faster r-cnn: Towards real-time object detection with region proposal networks.” Advances in neural information processing systems. 2015.
[5] Baccouche, Moez, et al.” Sequential deep learning for human action recognition.” International Workshop on Human Behaviour Understanding. Springer Berlin Heidelberg, 2011
[6] Delvingdeeper into convulational networks for learning video representations.Nicolas Ballas1, Li Yao1, Chris Pal2, Aaron Courville1,1MILA, Universit´e de Montr´eal.2 ´ Ecole Polytechnique de Mont´real.
[7] Karpathy, Andrej, and Li Fei-Fei.” Deep visual-semantic alignments for generating image descriptions.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.