IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 4, April 2023

DOI: 10.17148/IJARCCE.2023.124190

IMAGE DESCRIPTION GENERATOR USING DEEP LEARNING

A. Sricharitha[1], S.V. Amit[2], Md B Sultan Bahyal[3]

Bachelor of Engineering, Information Technology, Matrusri Engineering College, Hyderabad, India[1]

Bachelor of Engineering, Information Technology, Matrusri Engineering College, Hyderabad, India[2]

Bachelor of Engineering, Information Technology, Matrusri Engineering College, Hyderabad, India[3]

Abstract: Image description generator using Deep Learning can create an image's content using properly constructed, meaningful English sentences. The user's camera or mobile phone is continuously used to capture images in real-time. Our models extract features from an image using a convolutional neural network (CNN). To create a reliable image description in English, these Features Are Provided to A Recurrent Neural Network (RNN) Or A Long Short-Term Memory (LSTM) Network. We Use A CNN To Extract Features from The Image. CNNs are the state-of-the-art methods for object recognition and detection and have been used and studied for a variety of image tasks. In more detail, we extract features from the Fc7 Layer of the VGG-16 Network that has been trained on ImageNet and is well suited for object detection for all input images. Due to computational limitations in LSTM, we first obtain a 4096-dimensional image feature vector and then reduce it using principal component analysis (PCA) to a 512-dimensional image feature vector. These characteristics are fed into the LSTM network to produce a description of the image in accurate English, which might then be converted to audio using text-to-speech technology.

Keywords- Caption Generator, Feature Extraction, LSTM, Neural Network, Object Detection

I. INTRODUCTION

Any person would not choose to have a visual impairment, and there is no temporary fix for it. Visual impairment is nothing more than a visual disability that prevents temporary solutions from working. Worldwide, there are 285 million visually impaired people, including over 39 million blind people, according to the World Health Organisation. In a technologically developing world where even the smallest piece of work would require sight, it is quite difficult to live without one of the most useful sensory organs.

There may be numerous developments that could enhance the lives of those who are blind or visually impaired as we live in a time when the technology sector is booming. The Use of Technology to Assist the Visually Impaired Can Be Done in several Ways, one method is to identify the objects in an image and then provide a meaningful caption that would be spoken out to the person using the system, assisting them in connecting all the objects in the image. The difficulties include the inability to automatically translate the image's content into properly formed English sentences and the requirement that the caption express how the objects in the image relate to one another as well as their characteristics and the activities they are engaged in. These Descriptions Must Be Said in a Natural Language Like English, Which Requires a Language Model. Convolution neural networks, recurrent neural networks, the ImageNet dataset, and text to speech converters can all be used in conjunction with one another to achieve this.

II. DATASET

Dataset is an assortment of connected items that may be used by the system to classify various item types and group them together by comparison with the dataset. Microsoft COCO and ImageNet are the two types of datasets we use in this paper. While ImageNet uses images with variable resolution, our system needs input with a constant dimensionality. We downscale the image as a result to a specific resolution. In order to remove the central 256x256 patch from a rectangular image, we first rescale the image so that the shorter side was of length 256.

We trained our network using the raw RGB values of pixels as a result. The Microsoft COCO Dataset's 2014 Release, which has evolved into the industry's standard testbed, Will Be Used for This Exercise. Dataset is a group of connected items that might aid with image captioning. Each image in the dataset has five captions written by Amazon Mechanical Turk workers, with 80,000 training images and 40,000 validation images total.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 4, April 2023

DOI: 10.17148/IJARCCE.2023.124190

III. CONVOLUTIONAL NEURAL NETWORKS

In Machine Learning, A Convolutional Neural Network (CNN, Or Convnet) Is A Class of Deep, Feed- Forward Artificial Neural Networks That Has Successfully Been Applied to Analyzing Visual Imagery. CNN Compares Any Image Piece by Piece and The Pieces That It Looks for In an Image While Detection Are Called as Features. By Finding Rough Feature Matches in Roughly, The Same Position in Two Images CNN Gets Trained. Every Neuron in CNN Will Be Connected to Small Region of Neurons Below It This Would Allow Handling Number of Weights and Number of Neurons Required Will Also Be Less.

3.1 CONVOLUTION LAYER

The Convolution Layer Is the Core Building Block of a Convolutional Network That Does Most of The Computational Heavy Lifting. It Preserves Spatial Relationship Between Pixels Thereby Extracting and Learning Features Out of Them. The Image Is Represented as A Matrix and A Filter, Which Is Also a Matrix Is Used to Obtain the Convolved Feature Map or Activation Map by Sliding the Feature Matrix Over the Image Matrix As Shown In Fig. 1. We Can Perform Operations Such as Edge Detection, Sharpen and Blur Just by Changing Values in The Filter Matrix. It Captures the Local Dependencies in The Original Image [3].

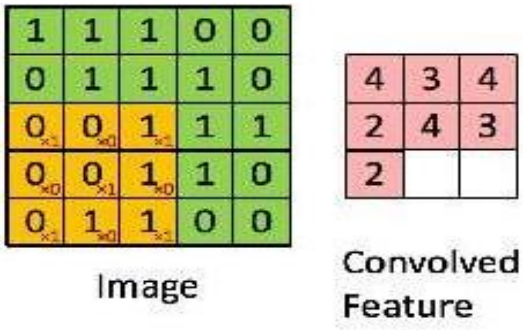


FIGURE 1: OBTAINING CONVOLVED FEATURE MAP FROM IMAGE MATRIX BY SLIDING FILTER MATRIX SEQUENTIALLY

3.2. Rectified Linear Unit Layer

Rectified Linear Unit (Relu) is a type of activation function that activates a node if an input value exceeds a predetermined threshold, while the output is zero if the input value is zero. However, there is a linear relationship between the input and the dependent variable if it is above a certain threshold.

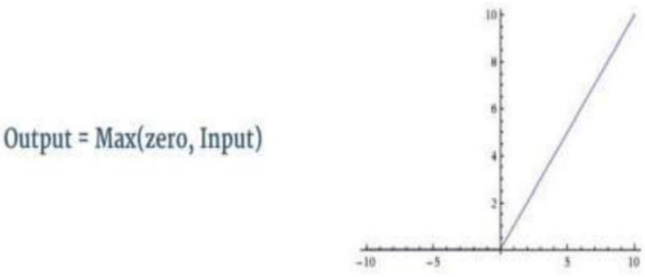


FIGURE 2: RELU OPERATION TO OBTAIN RECTIFIED FEATURE MAPS

It Replaces All the Negative Values in The Feature Map by Zero and Generates a Rectified Feature Map as Shown in Fig. 2. Relu Introduces Non-Linearity in The Convnet Since Most of The Image Data Are Non-Linear in Nature in The Real World.

3.3. Pooling Layer

In this layer, we shrink or reduce the size of the feature map in order to obtain smaller maps that would require fewer computations and parameters. Pooling from the downsized and rectified feature map can be done as a maximum, average, or sum pooling. As shown in Fig. 3, the number of output maps from pooling is the same as the number of filters in the convolution layer. Additionally, it renders the network insensitive to minor distortions, translations, and transformations in the input image.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 4, April 2023

DOI: 10.17148/IJARCCE.2023.124190

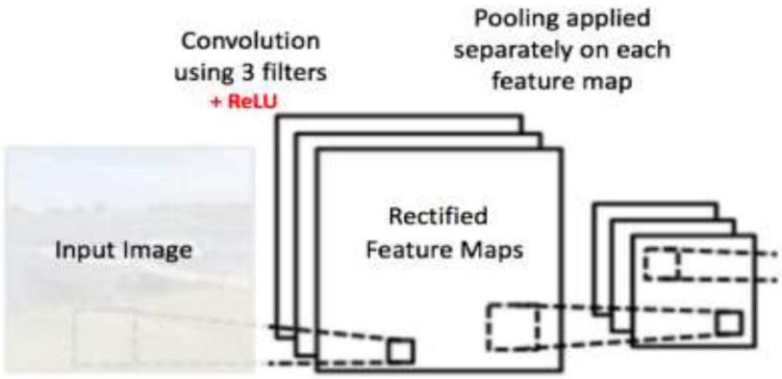


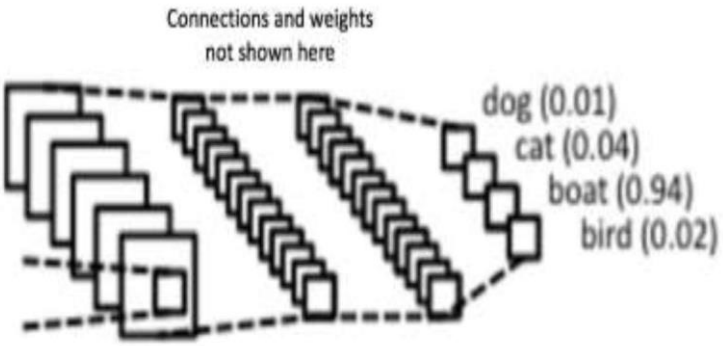
Figure 3: Pooling Applied to Rectified Feature Maps.

3.4. Fully Connected Layer

The final layer, where classification takes place, is where we put our downsized or shrunk images that we obtained after processing through the convolution, reluctance, and pooling layers into a single list or a vector.

It uses a SoftMax Activation Function and is a Traditional Multi-Layer Perceptron. High-Level Features are produced via convolution and pooling layers. The Fully Connected Layers' goal is to use these features to divide the data into different classes according to the dataset.

Training Of CNN Using Back-Propagation



- Initialize All Filters and Parameters with Random Variables
- Take Input Images for Training, Go Through the Forward Propagation and Find Output Probability for Each Class.
- Calculate Total Error by The Formula:

Total error= Total Error =$\sum$1/2(target probability *-Output Probability)* ²

Calculate Gradients of The Error with Respect to The Weights and Use Gradient Descent to Update the Filter Values and Parameters to Minimize the Output Error.
Repeat 2-4 For All Images in The Training Set.

IV. RECURRENT NEURAL NETWORK

The main purpose of RNN is to use sequential information. To predict future words in a sentence, it is very necessary to know the previous words. As the name suggests, RNNs are recursive because they perform the same operations on each element of the sequence and the current result depends on the results of previous calculations.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 4, April 2023

DOI: 10.17148/IJARCCE.2023.124190

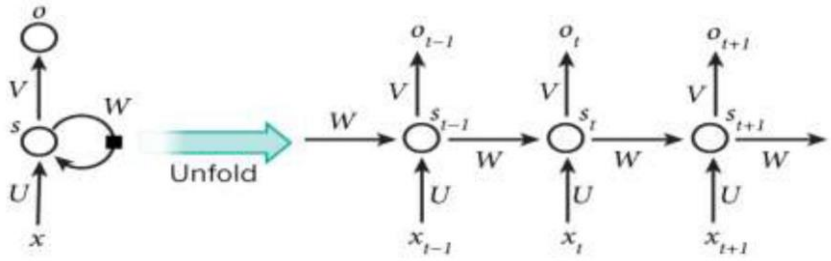


Figure 5 shows that the RNN is open to the entire network.

Development is just a network representation of each element of the set. For example, if the sequence is a sentence with 7 words, the network is divided into 7 layers, one layer for each word. The formulas that control the RNN process are: Xt represents the input in stage T. St represents the hidden state at time t represents the output of stage T. Hidden state St is a storage network memory The data from the previous time steps and are calculated as the previous hidden state and the current state as input:

St=F (Uxt Wst-1) (2)

The function F is a nonlinearity function, like Tanh or Relu. S-1, which is required to calculate the first hidden state, is reset to zero at Ot, the output at time T is calculated entirely from memory at time T.

Ot = SoftMax (Vst) (3)

Unlike the traditional approach that uses different parameters in each layer, RNN uses the same (U, V, W) vectors in each layer, which means that each layer performs the same process with different inputs. RNNs have shown great success in natural language processing. The most used RNN is LSTMs, which captures long-term dependencies much better than vanishing gradient problems. The development method is the same, but the hidden space calculation method is different.

V. LONG-SHORT TERM MEMORY

LSTMs are special recurrent neural networks, and these recurrent neural networks are capable of learning long-term dependencies. LSTM also has a chain-like structure like RNN [5]. The network takes 3 inputs: Xt is the current input, Ht-1 is the output of the previous LSTM device, and Ct-1 is the memory of the previous device, which is very important. It gives 2 outputs: Ht is the output of the current network and Ct is the memory of the current unit. A single LSTM device decides based on current input and past output and memory. Figure 6 shows the LSTM architecture and its different gates. Forget Gate: This removes unnecessary information from the cell state that is not needed. This is done with a multiplicative filter. It is driven by a single-layer neural network with a sigmoidal activation function, the output of which is applied to the old memory to decide whether it should be retained or forgotten.

New Memory Gate: This is also a single-layer neural network with the same inputs as the Forgetting layer. It controls the effect of new memory on old memory. Another neural network whose activation function is Tanh creates a new memory. The output of this network is multiplied by the new memory gate and complements the old memory to form the new memory

Output port: The output valve is controlled by the new memory, the previous output Ht-1, the input Xt and the bias vector. This controls the amount of new memory flowing into the next LSTM unit.

The following formulas are used to derive the value of each port:

I_t $Wx_ix_t + Wh_ih_{t-1} + Wc_ic_{t-1} + B_i)$ (4)

F_t $Wx_fx_t + Wh_fh_{t-1} + Wc_fc_{t-1} + B_f)$ (5)

$C_t = F_tc_{t-1} + I_ttanh\ (Wx_cx_t + Wh_ch_{t-1} + B_c)$ (6)

O_t $Wx_ox_t + Wh_oh_{t-1} + Wc_oc_{t-1} +$ (7)

$B_o)\ H_t = O_ttanh\ (C_t)$ (8)

Were,

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 4, April 2023

DOI: 10.17148/IJARCCE.2023.124190

I_t Is Input Gate Vector
F_t Is the Forget Gate Vector
C_t Is the Cell State Vector
H_t Is Output of LSTM Cell

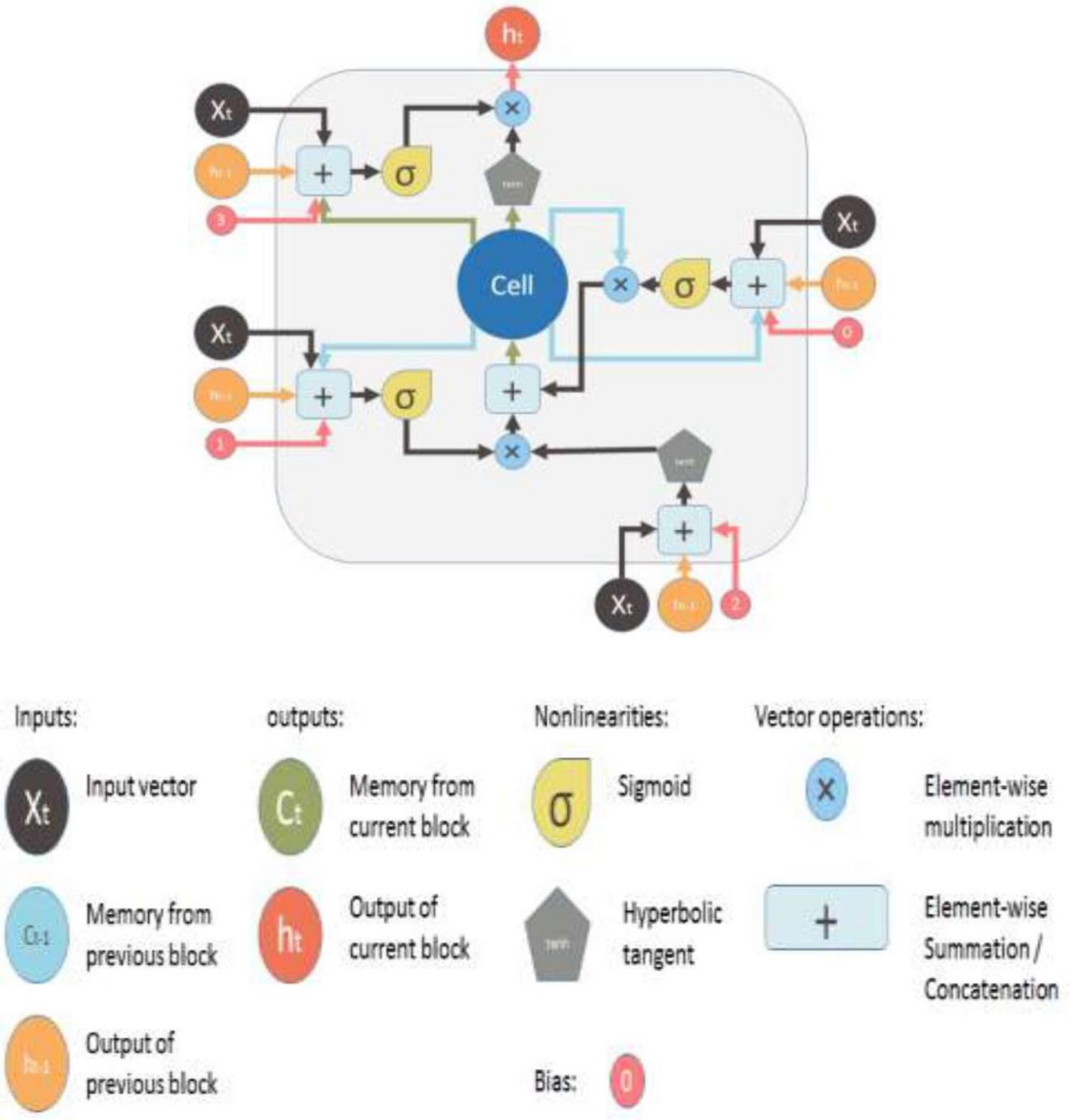


Figure 6: LSTM Architecture with Gates

This is the input gate vector Ft is the forget gate vector Ct is the state vector of the cell. Ht is the output of the LSTM cell Ot is the output of 3 different single layer neural networks with values of -1 and 1, the steps to operate the LSTM cell are:

- Step 1: The first step in LSTM is to decide what information we discard from the state of the cells. This decision is made by a sigmoid layer called the Forgetting Gate layer
- Step 2: In This Step We Decide What New Information We Are Going to Store in The Cell State.
- This Has 2 Parts, First, Sigmoid Layer or Input Gate Layer Decides Which Values Will Be Updated. Then, The Tanh Layer Creates a Vector of New Values That Could Be Added to The State.
- Step 3: Now We Will Update the Old Cell State into The New Cell State.
- Step 4: We Will Run a Sigmoid Layer, Which Decides the Parts of The Cell State We Are Going to Output. Then We Put the Cell Through Tanh and Multiply It by The Output of The Sigmoid Layer, So That We Only Output the Parts We Decided To [5].

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 4, April 2023

DOI: 10.17148/IJARCCE.2023.124190

VI. TECHNICAL FRAMEWORK

The Proposed Architecture Involves All the Above Stated Techniques to Achieve an Almost 80% Accuracy in Scene Description. Fig. 7 Shows the Basic Block Architecture of The Current System.

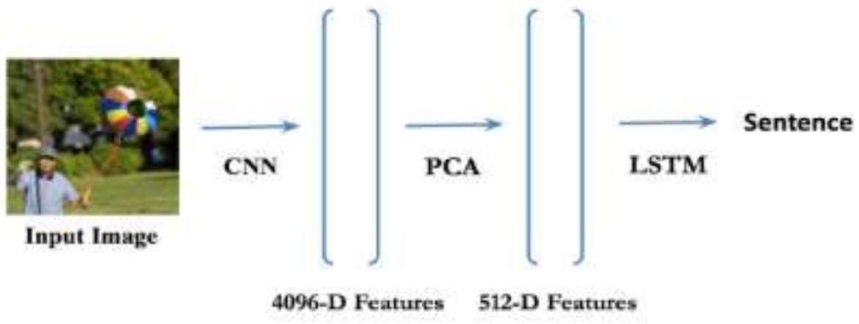


Figure 7: Image Extraction and Language Generation Pipeline Architecture

- The features are taken from images using CNN and are taken from the Fc7 Layer of the VGG-16 Network that has previously been trained on the ImageNet Dataset.

- Due to computational limitations for the LSTM input, a 4096-dimensional feature vector is obtained and reduced using principal component analysis (PCA) to a 512-dimensional vector.

- Vanilla RNNs are good at text generation and speech recognition, but it's hard to teach them to learn long-term dependencies because of the disappearing and exploding gradient problem that happens when gradients are propagated over many layers of recurrent networks [2].

- As a result, LSTM offers a solution by including memory units that enable the network to learn when to update or forget previous and current hidden states in response to new information. LSTM is the best recurrent neural network Solution Against the Vanilla RNN to Get the An Almost Perfect Caption [2]

VII. CONCLUSION

We introduced this book to help visually impaired people using deep learning techniques. Techniques such as Convolutional Neural Networks (CNN) and feature maps created using such neural networks help us recognize objects and then generate sentences using recurrent networks such as LSTM (Long-Short Term Memory).

CNN and LSTM are currently state-of-the-art techniques for object detection, scene representation and scene description, so the generated captions describe the objects shown in the images very well. Thanks to the high quality of the image descriptions created, the visually impaired can benefit greatly from text-to-speech technology and better understand their surroundings.

Further research and exploration of this project is to create real-time video texts to provide the user with a real-time understanding and description of the scene, instead of capturing still images that can only provide information to the blind at one specific time.

ACKNOWLEDGEMENT

We Sincerely Thank Our Project Guide **Dr. J Srinivas** for His Guidance and Encouragement in Carrying Out This Research. We Wish to Express Our Sincere Gratitude to The Principal and The Head of Department of Information technology, Matrusri engineering college for Providing Us with this wonderful opportunity.

"Image description generator using Deep Learning" is a Research that Bears an Imprint of Many People. Finally, We Would Like to Thank Our professor's and Friends Who Helped Us in Every Possible Way for Completion of This Research Paper.

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 4, April 2023
DOI: 10.17148/IJARCCE.2023.124190

REFERENCES

[1] Bourne RRA, Flaxman SR, Braithwaite T, Cicinelli MV, Das A, Jonas JB, Et Al, "Magnitude, Temporal Trends, And Projections of The Global Prevalence of Blindness and Distance and Near Vision Impairment: A Systematic Review and Meta-Analysis," The Lancet Global Health, Sep. 2017 Vol. 5
[2] Christopher Elamri, Teun De Planque, "Automated Neural Image Caption Generator for Visually Impaired People," Stanford University, Department of Computer Science CS224d, 2016
[3] Ujjwalkarn An Intuitive Explanation of Convolutional Neural Networks [2016 Online]. Available: Https://Ujjwalkarn.Me/2016/08/11/Intuitive-Explanation-Convnets/. [Accessed: 28-Feb-2018]
[4] WildMl.com Recurrent Neural Networks Tutorial Online]. Available: Http://Www.Wildml.Com/2015/09/Recurrent- Neural-Networks-Tutorial-Part-1-Introduction-To-Rnns/. [Accessed: 17-Sep-2015]
[5] Medium.com Understanding LSTM Online]. Available: Https://Medium.Com/Mlreview/Understanding-Lstm-And-Its-Diagrams-37e2f46f1714. [Accessed: 14-Mar-2016]

