**IMAGE CAPTIONING**

**MINOR PROJECT II**

Submitted by:

**NANDAN SHARMA (9916103113)**

**RITIKA SINGH (9916103153)**

**VRIDHI SAGAR (9916103029)**

**AYUSHI JAIN (9916103226)**

Under the supervision of

**Dr.Gyan Singh**

**Assistant Professor(Department of CSE)**

****

**Department of CSE/IT**

**Jaypee Institute of Information Technology University, Noida**

**March 2018**

**ACKNOWLEDGEMENT**

The success and final outcome of this project required a lot of guidance and assistance from many people and we are extremely privileged to have got this all along the completion of our project. All that we have done is only due to such supervised and assistance and we would not forget to thank them.

We would like to place on record our deep sense of gratitude to **Dr. Gyan** **Singh**–Assistance Professor, Jaypee Institute of information technology, India for her generous guidance, help and useful

Suggestions. The valuable advice and suggestions for the corrections, modification and improvement did enhance the perfection in performing our job well.

We are thankful for her simulating guidance, continuous encouragement and supervision throughout the course of present work.

We also wish to extend thanks to our classmates for their insightful comments and constructive suggestion to improve the quality of this project work.

**Signature of students Signature of supervisor**

Nandan Sharma (9916103113) Dr. Gyan Singh

Ritika Singh (9916103153)

Vridhi Sagar (9916103029)

Ayushi Jain (9916103226)

**Abstract**

Computer vision has become ubiquitous in our society, with applications in several fields. In this project, we focus on one of the visual recognition facets of computer vision, i.e image captioning. The problem of generating language descriptions for visual data has been studied from a long time but in the field of videos. In the recent few years emphasis has been lead on still image description 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 was to train convolutional neural networks with several hundreds of hyperparameters and apply it on a huge dataset of images , and combine the results of this image classifier with a recurrent neural network to generate a caption for the classified image. In this report we present the detailed architecture of the model used by us.

**TABLE OF CONTENTS**

Page No.

*Abstract* *i.*

*List of Tables* *iii.*

*List of Figures* *iv.*

*List of Abbreviations v.*

**Chapter 1: INTRODUCTION** 1.

**Chapter 2: BACKGROUND STUDY** 2

**Chapter 3: REQUIREMENT ANALYSIS** 4

3.1 Software and hardware requirements

3.2 Functional requirements

3.3 Non-Functional Requirement

**Chapter 4: DETAILED DESIGN** 5

**Chapter 5: IMPLEMENTATION** 7

**Chapter 6: TESTING** 9

**Chapter 7: CONCLUSION** 10

**Gantt Chart** 10

**References** 11

**Power Point Presentation** 12-16

**LIST OF TABLES**

**Table Title Page**

3.1 Requirements 4

6.1 Testing 9

**LIST OF FIGURES**

**Figure Title Page**

4.1 A schematic of the VGG-16 deep CNN layer 5

4.2 Image (in green) and Filter (in orange) 6

4.3 Convolution operation 6

5.1 Sample image and corresponding captions from the Flickr8k dataset 7

5.2 The output of our CNN encoder would thus be passed to the language

generating RNN 8

6.1 Encoding 9

6.2 Start and End token added 9

**ABBREVATIONS**

* CNN Convolution Neural Network
* VGG Visual Geometry Group 16
* LSTM Long Short Term Memory
* RNN Recurrent Neural Network
* ReLu Rectified Linear Unit

**1.INTRODUCTION**

**1.1 Overview**

Artificial Intelligence(AI) is now at the heart of innovation economy and thus the base for this project is also the same. In the recent past a field of AI namely Deep Learning has turned a lot of heads due to its impressive results in terms of accuracy when compared to the already existing Machine learning algorithms. The task of being able to generate a meaningful sentence from an image is a difficult task but can have great impact, for instance helping the visually impaired to have a better understanding of images.

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

The aim of the project is to train convolutional neural networks with several hundreds of hyperparameters and apply it on the Flickr8k dataset and combine the results of this image classifier with a recurrent neural network to generate a caption for the classified image. In this report we present the detailed architecture of the model used by us.

1. **BACKGROUND STUDY**

Apart from the technical studies for building the main project, the planning session required a lot of knowledge about the image recognition. We had to read an insight about the different algorithms to reach to the best applicable algorithm and technique.

**2.1) Very deep convolutional neural network based image classification using small training sample size [1]**

In this paper, a modified VGG-16 network is proposed and used to fit CIFAR-10(a small dataset). By adding stronger regularizer and using Batch Normalization, this work achieves 8.45% error rate on CIFAR-10 without severe overfitting. The results show that the very deep CNN can be used to fit small datasets with simple and proper modifications and don't need to re-design specific small networks and if a model is strong enough to fit a large dataset, it can also fit a small one.

**2.2) CIDEr: Consensus-based Image Description Evaluation [2]**

This work proposes a novel paradigm for evaluating image descriptions that uses human consensus. This paradigm consists of three main parts: a new triplet-based method of collecting human annotations to measure consensus, a new automated metric that captures consensus, and two new datasets: PASCAL-50S and ABSTRACT-50S that contain 50 sentences describing each image. Ther simple metric captures human judgment of consensus better than existing metrics across sentences generated by various sources. This work also evaluates five state-of-the-art image description approaches using this new protocol and provide a benchmark for future comparisons.

**2.3) Image Captioning with Deep Bidirectional LSTMs [3]**

This paper presents an end-to-end trainable deep bidirectional LSTM (Long-Short Term Memory) model for image captioning. This model builds on a deep convolutional neural network (CNN) and two separate LSTM networks. It is capable of learning long term visual-language interactions by making use of history and future context information at high level semantic space. Two novel deep bidirectional variant models are proposed to learn hierarchical visual-language embeddings. Data augmentation techniques such as multi-crop, multi-scale and vertical mirror are proposed to prevent overfitting in training deep models. In this work the evolution of bidirectional LSTM internal states over time is visualised and qualitatively analysis of how the models "translate" image to sentence is also done. This work demonstrates that bidirectional LSTM models achieve highly competitive performance to the state-of-the-art results on caption generation even without integrating additional mechanism.

**2.4) Show and Tell: A Neural Image Caption Generator [4]**

In this paper, a generative model based on a deep recurrent architecture that combines recent advances in computer vision and machine translation and that can be used to generate natural sentences describing an image is presented. The model is trained to maximize the likelihood of the target description sentence given the training image. Experiments on several datasets show the accuracy of the model and the fluency of the language it learns solely from image descriptions. This model is often quite accurate, which we verify both qualitatively and quantitatively. For instance, while the current state-of-the-art BLEU-1 score (the higher the better) on the Pascal dataset is 25, this work’s our approach yields 59, to be compared to human performance around 69. It also shows BLEU-1 score improvements on Flickr30k, from 56 to 66, and on SBU, from 19 to 28. Lastly, on the newly released COCO dataset, it achieves a BLEU-4 of 27.7, which is the current state-of-the-art.

**3. REQUIREMENT ANALYSIS**

**3.1 Minimum System Requirements-**

|  |  |
| --- | --- |
| **Particulars** | **Recommended Configuration** |
| **Processor** | Intel(R) core(TM) i7-7500U CPU |
| **RAM** | 4 GB or more\* |
| **Hard Disk** | 1 GB of free space |
| **Monitor Resolution** | 1024 x 768 or higher\* |
| **Operating system** | Windows 10 or Linux or macOS |

**Table 3.1: Requirements**

**3.2 Software Requirements-**

-Python 2.7 or above

-libraries : Tensorflow,Keras,Numpy,h5py,Pandas,Pillow,Pyttsx

**3.3 Functional Requirements-**

-project will have the ability to encode images using VGG16 model.

-system should be able to remove noise from text and image data.

-project should be able to generate caption for fed image.

**3.4 Non-Functional Requirements-**

-there should be atleast 5 captions per image.

-data set should consist of minimum 8000 images.

-each image should have a unique id.

1. **DETAILED DESIGN**
   1. **Data Pre processing**

In this step,all the punctuation and numbering from the text dataset is removed. Then a start and an end token is added to each text in the data set.On execution this step creates new text files in Flickr8K\_Text folder.

* 1. **Encoding Image**

This step creates a file image\_encodings.p which generates image encodings by feeding the image to VGG16 model.VGG-16 is a convolutional neural network that is trained on more than a million images from the ImageNet database. The network is 16 layers deep and can classify images into 1000 object categories, such as keyboard, mouse, pencil, and many animals. As a result, the network has learned rich feature representations for a wide range of images. The network has an image input size of 224-by-224

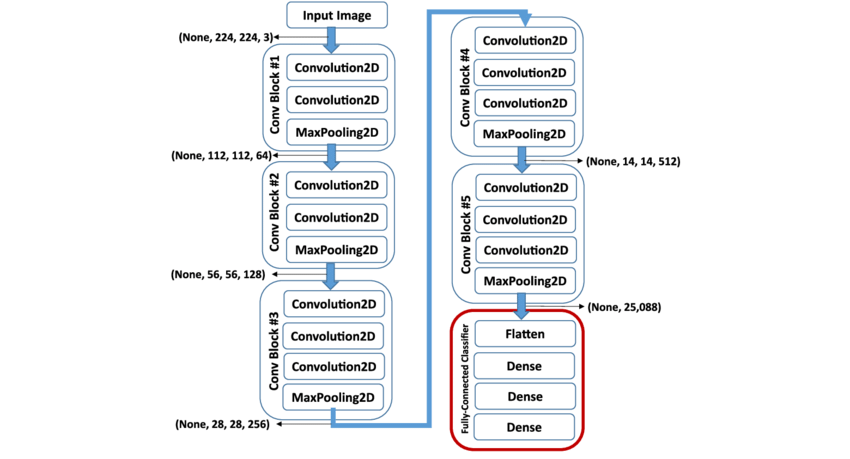


Fig.4.1:A schematic of the VGG-16 deep CNN layer

**4.1.1 Convolutional Neural Network**

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.

The entire architecture of a convnet can be explained using four main operations namel

**1. Convolution**: The purpose of convolution operation is to extract features from an image. We consider filters of size smaller than the dimensions of image.The entire operation of convolution can be understood with the example below.

Consider a small 2-dimensional 5\*5 image with binary pixel values. Consider another 3\*3 matrix.

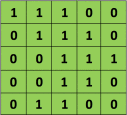
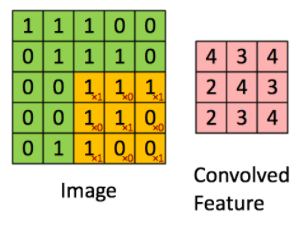
  

Fig. 4.2: Image (in green) and Filter (in orange) Fig. 4.3: Convolution operation

We slide this orange 3\*3 matrix over the original image by 1 pixel and calculate element-wise multiplication of the orange matrix with the sub-matrix of the original image and add the final multiplication outputs to get the final integer which forms a single element of the output matrix which is shown in the Fig. 8 by the pink matrix.

The 3\*3 matrix is called a filter or kernel or feature detector and the matrix formed by sliding the filter over the image and computing the dot product is called the Convolved Feature or Activation Map or the Feature Map. The number of pixels by which we slide the filter over the original image is known as stride.

**2. Non- Linearity (ReLU)**: An additional operation is applied after every convolution operation. The most commonly used non-linear function for images is the ReLU which stands for Rectified Linear Unit. The ReLU operation is an element-wise operation which replaces the negative pixels in the image with a zero.Since most of the operations in real-life relate to non-linear data but the output of convolution operation is linear because the operation applied is elementwise multiplication and addition

**3. Pooling or Sub Sampling**: The pooling operation reduces the dimensionality of the image but preserves the important features in the image. The most common type of pooling technique used is max pooling. In max pooling you slide a window of n\*n where n is less than the side of the image and determine the maximum in that window and then shift the window with the given stride length.

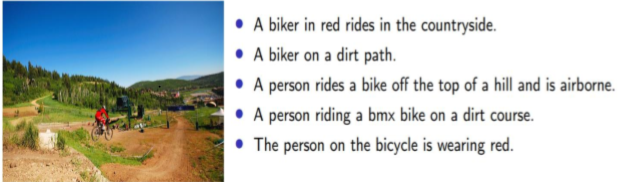
**4. Classification (Fully Connected Layer)**:The fully connected layer is the multi-layer perceptron that uses the SoftMax activation function in the output layer. The term “fully-connected” refers to the fact that all the neurons in the previous layer are connected to all the neurons of the next layer. The convolution and pooling operation generate features of an image. The task of the fully connected layer is to map these feature vectors to the classes in the training data.

1. **IMPLEMENTATION**

**5.1 Dataset:**

For the task of image captioning we use **Flickr8k dataset**. 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.

The dataset contains 6000 training images, 1000 development images and 1000 test images. A sample from the data is given in figure.



**Fig.5.1 : Sample image and corresponding captions from the Flickr8k dataset**

We have used Flickr8k rather Flicker30k and MSCOCO for image captioning because both these datasets have more than 30,000 images thus processing them becomes computationally very expensive.

**5.2 Preprocessing:**

Since the image taken by VGG16 model for classification needs to be a 224-by-224 image. So all the images needed to be converted into 224-by-224 dimensional image, the only preprocessing done is by subtracting the mean RGB values from each pixel determined from the training images.

**5.3 Deep CNN Architecture:**

CNN architecture used in this particular step is **VGG16 (Visual Geometry Group-16).** It is a convolutional neural network that is trained on more than a million images from the ImageNet database. The network is 16 layers deep and can classify images into 1000 object categories. The network has 41 layers. There are 16 layers with learnable weights: 13 convolutional layers, and 3 fully connected layers.

The convolution layer consists of 3\*3 filters and max pooling is done using 2\*2-pixel. A Rectified Linear Unit (ReLU) activation function is used after every convolution layer.



VGG16 model is used to encode an image to its features, with its pretrained weights loaded. For our purpose however, we need not classify the image and hence we remove the last softmax layer of VGG16 and the vector of dimension(4096) is obtained from the second last layer.

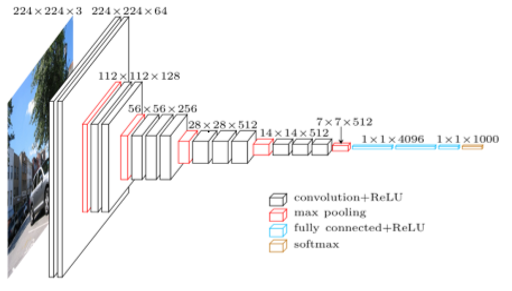


Fig.5.2:The output of our CNN encoder would thus be passed to the language generating RNN.

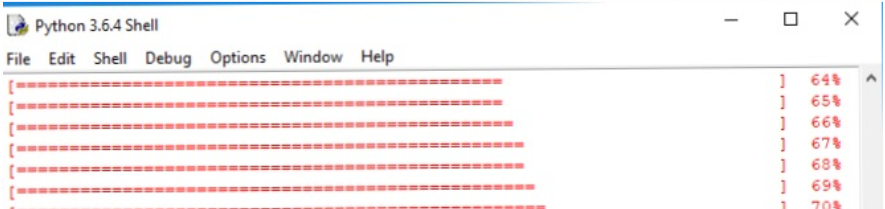
**5.4 Image Captioning Model:**

The image encodings are passed on to the Language Generating Recurrent Neural Network (RNN) which helps in generating a meaningful sentence for the image. A Long Short-Term Memory (LSTM) network follows the pretrained VGG16. The LSTM network is used for language generation. LSTM differs from traditional Neural Networks as a current token is dependent on the previous tokens for a sentence to be meaningful and LSTM networks take this factor into account.

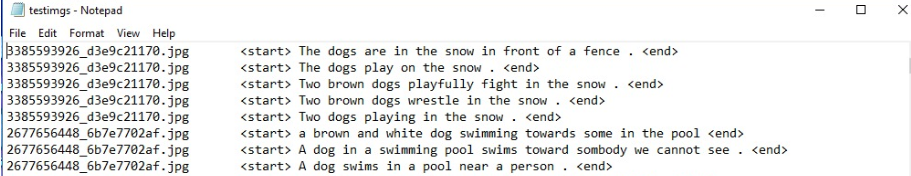
**6. TESTING**

To get detailed information about the various errors the following tests were carried out and detailed descriptions are given as follows.

|  |  |  |  |
| --- | --- | --- | --- |
| **Functionality** | **Input** | **Expected Output** | **Remark** |
| Preprocessing (resizing / cleaning of text files) | Multiple Images and text | Resize successful and start and end token added successfully. | Preprocessing successful |
|  |  | Resize  unsuccessful. | File path may be incorrect. |
| Encoding | Multiple Images and text | Image\_encoding file formed. | Encoding successful. |
|  |  | Image\_encoding file not formed. | Encoding not successful. |



**Fig.6.1 : Encoding**



**Fig.6.2: Start and End tokens added to every sentence.**

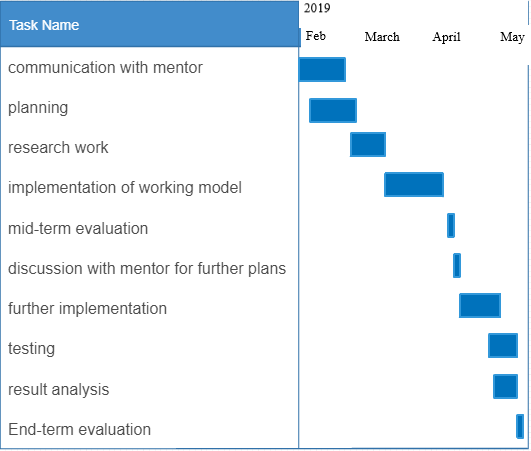
**7. CONCLUSION**

In this report we discussed about VGG16 CNN model and why it is used among so many other models, and discussed abouts its limitations.

We were able to successfully preprocess all the images in the Flickr8K dataset, so that it can be passed on to the deep learning model for further processing. And our deep CNN architecture VGG16 was capable to encode the images to its features.

Further we hope to implement LSTM (Long Short-Term Memory) network to generate captions.

**GANTT CHART**



**8.REFERENCES**

[1] Liu, S., & Deng, W. (2015, November). Very deep convolutional neural network based image classification using small training sample size. In 2015 3rd IAPR Asian conference on pattern recognition (ACPR) (pp. 730-734). IEEE.

[2] Vedantam, R., Lawrence Zitnick, C., & Parikh, D. (2015). Cider: Consensus-based image description evaluation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4566-4575).

[3] Wang, C., Yang, H., Bartz, C., & Meinel, C. (2016, October). Image captioning with deep bidirectional LSTMs. In Proceedings of the 24th ACM international conference on Multimedia (pp. 988-997). ACM.

[4] Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3156-3164).