



# KONKAN GYANPEETH COLLEGE OF ENGINEERING, KARJAT

Affiliated to University of Mumbai, Approved by AICTE, New Delhi.

## Automatic Image Captioning using Deep Learning.

- Project Members:

- Dhruv Solanki(Roll No – 50).
- Tejal Supe (Roll No – 51) .
- Siddhi Undirkar (Roll No – 55).

Under the Guidance of:  
Prof. J.P.Patil



# CONTENTS:

- Abstract
- Introduction
- Objectives
- Scope
- CNN(Convolution Neural Networks)
- LSTM (Long short term memory)
- VGG16 model
- Phases
- Literature Survey
- Conclusion
- References



# Abstract

Image captioning means automatically generating a caption for an image. Automatically creating the description of an image using any natural language sentences is a very challenging task. It requires expertise of both image processing as well as natural language processing. Sharing photos through the internet (e.g. Instagram, Facebook, etc.), which becomes a common practice, leads to archives in the order of millions of images. Manual annotation is time consuming and extremely labour-intensive work. Thereafter, computer-assisted system are desired to lesson these difficulties by automation using machine learning. The annotated image can be used for retrieval purposes, in image processing algorithms, intelligent scanner, digital cameras, photocopiers, and printers. Our goal is to create system which will annotate images based on previously learned datasets.





# Introduction

Image annotation refers to the tagging of images with appropriate keywords. The process of annotating images manually is time-consuming and extremely labor-intensive work. Thereafter, computer-assisted systems are desired to lessen these difficulties by automation. The automatic image annotation is targeted to assign a few relevant text keywords to an input image that reflects its visual content. Machine learning techniques are used to develop the image annotation systems to map the low-level (visual) features to high-level concepts or semantics. There are two main approaches to Image Captioning: bottom-up and top-down. Bottom-up approach generate items observed in an image, and then attempt to combine the items identified into a caption. We are going to use two neural network algorithm which is Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) based Long Short Term Memory (LSTM). In the project CNN will be used to train the images as well as to detect the objects in the image with the help of various pre trained models and RNN based LSTM will be used to generate captions from the generated object keyword



# Objectives

- The objective of our project is to extract the features of a given image and to automatically caption the objects present in the image.
- To build a system that will identify multiple objects from a given image.
- To create a system that will be able to caption the objects present in the image.
- To be able to train the model created to identify and caption the images.
- To be able to test and evaluate the performance of the system of both the training and testing phase.



# Scope

- The scope of this project is to classify the images from the different classes available which makes it easy to search an image based on a keyword.
- The system developed in this project is such that it will add a caption to multiple objects present in an image.
- The images after been identified and captioned will be saved into the separate database

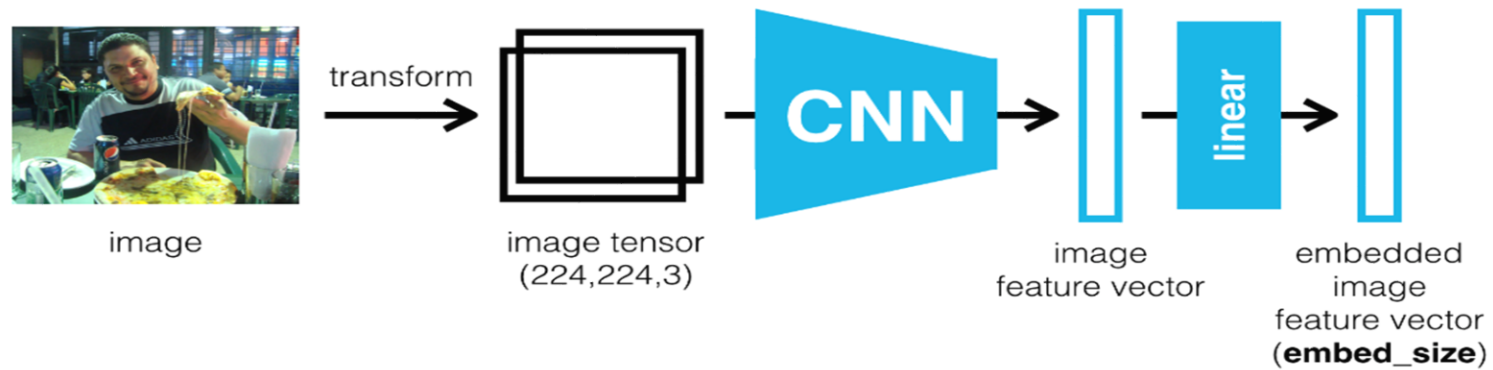




# CNN(Convolution Neural Networks):

- The convolutional neural network, or CNN for short, is a specialized type of neural network model designed for working with two-dimensional image data, although they can be used with one-dimensional and three-dimensional data.
- Convolutional neural networks do not learn a single filter; they, in fact, learn multiple features in parallel for a given input.
- CNN is often called as encoder because it encodes the content of the image into a smaller feature vector.

# CNN Model:



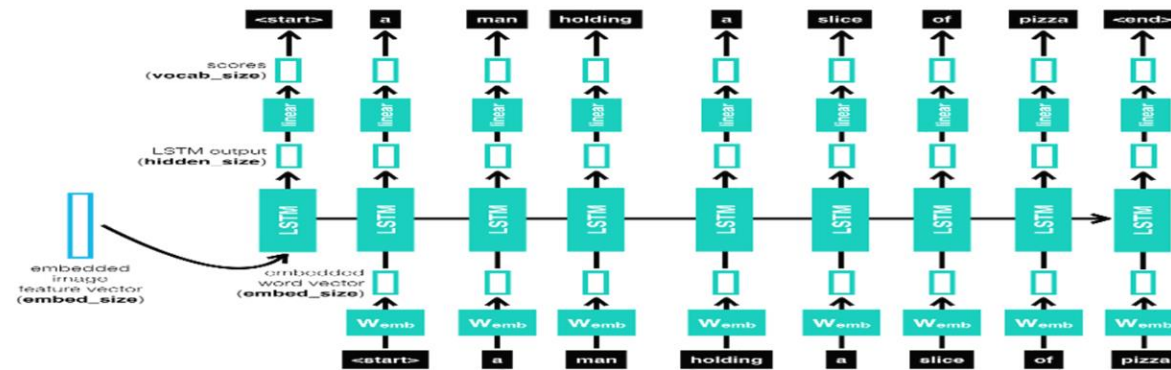




# LSTM(Long Short Term Memory):

- Recurrent neural networks (RNN) are the state of the art algorithm that remembers its input, due to an internal memory, which makes it perfectly suited for machine learning problems that involve sequential data.
- RNN remembers things for just small durations of time, i.e. if we need the information after a small time it may be reproducible, but once a lot of words are fed in, this information gets lost somewhere.
- This issue can be resolved by applying a slightly tweaked version of RNNs – the Long Short-Term Memory Networks. Long Short-Term Memory (LSTM) networks are a type of recurrent neural network capable of learning order dependence in sequence prediction problems.
- Lstm is called as decode because it decodes the vector and try to turn into a caption.

# LSTM Model:



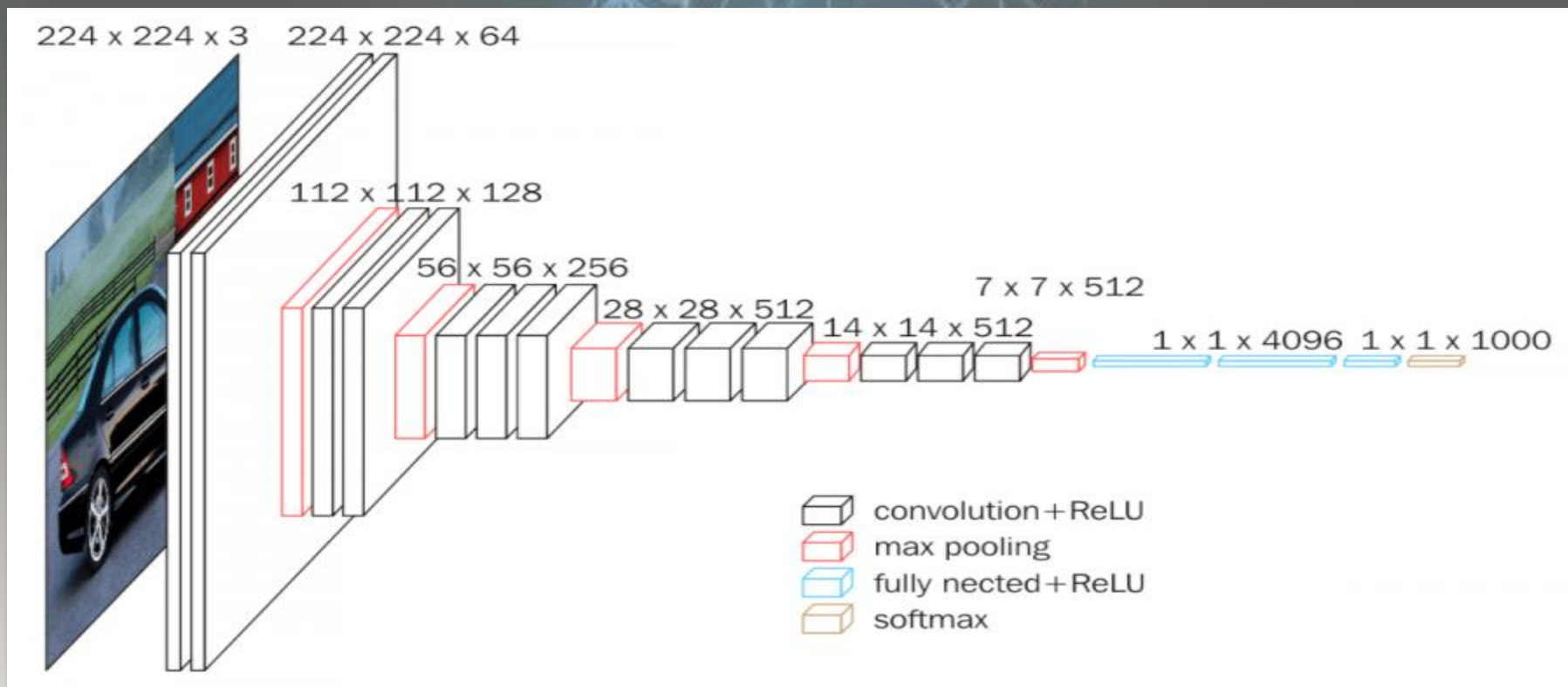


# VGG16

- There are various pre-trained Convolutional Neural Network Trained model such as VGG19, VGG16, Inceptionv3, ResNet, Mobile Net etc. but in our project we are going to use VGG16 pre-trained CNN model. VGG16 has 13 convolutional layers, some of which are followed by maximum pooling layers and then 4 fully connected layers and finally a 1000-way softmax classifier.
- At last there are 3 fully connected layer followed by last softmax layer. In our project we are going to cut-off the last layer because vgg16 is used to classify the image but we don't want to classify the image but just get fixed-length information vector of image.



# VGG16 Model





## Phases:

- Phase 1: Prepare Photo Data
- Phase 2: Prepare Text Data
- Phase 3: Develop Deep Learning Model
- Phase 4: Evaluate Model
- Phase 5: Generate New Captions

# Literature survey

Sr. No.	Paper Title	Author's Name	Techniques Used
1	Image Captioning Using CNN and LSTM	Ali Ashraf Mohamad	Technique - CNN , LSTM Dataset – Flickr8k Accuracy – BLEU-0.46
2	Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering	Peter Anderson, Damien Teney, Mark Johnson , Stephen Gould, Lei Zhang, Xiaodong He, Chris Buehler	Technique - Faster R-CNN Dataset – MSCOCO Evaluation Matrix – BLEU-4 – 36.9 CIDEr – 117.9 SPICE – 21.5
3	Transfer Learning Using VGG16 with Deep CNN for classifying Image	Srikanth Tammina	Model : VGG16



Sr. No	Paper Title	Author's	Techniques used
4	Deep Learning based Automatic Image Caption Generation	Varsha Kesavan, Vaidehi Muley , Megha Kolhekar	Techniques – CNN(Encoder), RNN(Decoder) Dataset – MSCOCO Pre Trained Model - VGG16 model(Encoder) GRU network(Decoder)
5	Guiding the Long-Short Term Memory model for Image Caption Generation	Xu Jia, Efstratios Gavves, Basura Fernando, Tinne Tuytelaars.	Techniques- LSTM , gLSTM Dataset - <i>Flickr8K</i> , <i>Flickr30K</i> and <i>MS COCO</i> <i>Evaluation Matrix – BLEU</i> , <i>METEOR</i>
6	Image Captioning with Semantic Attention	Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, Jiebo Luo	Techniques – RNN Dataset – MSCOCO , <i>Flickr30K</i> <i>Evaluation Matrix – BLEU</i> , <i>METEOR</i> , ROUGE-L, CIDEr



## Conclusion

- In this stage we did a deep study about the various datasets and methods for the best accuracy of image captioning. We did study about the CNN, RNN based LSTM architecture as well as various output methods used in different models, to finding out the suitable architecture and output method for this task. We will evaluate using evaluation metrics such as BLEU , ROUGE , CIDEr , METEOR , SPICE.
- We have used python version 3.5 and have Run our code in Google Colab.



## References

- [www.researchgate.net/publication/332109738](http://www.researchgate.net/publication/332109738) A Survey in Deep Learning Model for Image Annotation.
- [www.researchgate.net/publication/328993377](http://www.researchgate.net/publication/328993377) Building Detection and Segmentation Using a CNN with Automatically Generated Training Data
- Claudio Cusano, Gianluigi Ciocca. *Image annotation using SVM*
- Mahdia Bakalem, Nadjia Benblidia. *A Comparative Image Annotation*. 978-1-4799-4796-6/13/c 2013 IEEE. (2012).





THANK YOU!

