A Synopsis on

Image Captioning

Submitted in partial fulfilment of the requirements in

**Computer Engineering**

By

**Shyamkrishna Menon – 18102014**

**Siddharth Nair – 18102044**

**Atharva Ranade - 18102016**

**Omkar Thavai – 18102061**

Under the Guidance of

**Guide: Prof. S.S. Mujawar**



**Department of Computer Engineering**

A.P. Shah Institute of Technology G.B. Road, Kasarvadavli, Thane(W), Mumbai - 400615

University Of Mumbai

**Academic Year 2020-2021**

Table of Contents

[Chapter 1 6](#_Toc81601059)

[Introduction: 6](#_Toc81601060)

[Chapter 2:Project Concept 7](#_Toc81601061)

[Chapter 2.1: Abstract: 7](#_Toc81601062)

[Chapter 2.2:Objectives: 7](#_Toc81601063)

[Chapter 2.3:Literature Review: 7](#_Toc81601064)

[Chapter 2.4:Problem Definition: 8](#_Toc81601065)

[Chapter 2.5:Scope: 9](#_Toc81601066)

[Chapter 2.6:Technology Stack: 9](#_Toc81601067)

[Chapter 2.7:Benefits for environment and society: 9](#_Toc81601068)

# Chapter 1

## Introduction:

A large amount of information is stored in an image. Everyday huge image data is generated on social media and observatories. Deep learning can be used to automatically annotate these images, thus replacing the manual annotations done. This will greatly reduce the human error as well as the efforts by removing the need for human intervention The generation of captions from images has various practical benefits, ranging from aiding the visually impaired, to enabling the automatic, cost-saving labelling of the millions of images uploaded to the Internet every day, recommendations in editing applications, beneficial in virtual assistants, for indexing of images, for visually challenged people, for social media, and several other natural language processing applications. The field brings together state-of-the-art models in Natural Language Processing and Computer Vision, two of the major fields in Artificial Intelligence. One of the challenges is availability of large number of images with their associated text ever expanding internet.

Generating captions automatically from images is a complex task as it entails the model to extract features from the images and then form a meaningful sentence from the available features. Basically, the feature extraction is done by training a CNN (Convolutional Neural Network) with huge number of images and the correct weights are identified by multiple forward and backward iterations. With the help of RNN (Recurrent Neural Network) and the extracted features, a sentence is generated.

# Chapter 2:Project Concept

## Chapter 2.1:Abstract:

This project aims at generating automated captions by learning the contents of the image. At present images are annotated with human intervention and it becomes nearly impossible task for huge commercial databases. The image database is given as input to a deep neural network (Convolutional Neural Network (CNN)) encoder for generating “thought vector” which extracts the features and nuances out of our image and RNN (Recurrent Neural Network) decoder is used to translate the features and objects given by our image to obtain sequential, meaningful description of the image. In this project, we systematically analyze different deep neural network-based image caption generation approaches and pretrained models to conclude on the most efficient model with fine-tuning.

## Chapter 2.2:Objectives:

The goal of image captioning is to automatically generate descriptions for a given image, i.e., to capture the relationship between the objects present in the image, generate natural language expressions , and judge the quality of the generated descriptions. The problem, therefore, is seemingly more difficult than popular computer vision tasks, e.g., object detection or segmentation, where the emphasis is solely on identifying the different entities present in the image. With recent advancements in training neural networks, the availability of GPU computing power, and large datasets, neural network driven approaches are the most popular choice for handling the caption generation problem. However, humans are still better at interpreting images and constructing useful and meaningful captions, with or without a particular application context, which renders it an interesting applications.In this project we aim to implement a transfer learning approach to generate automated captions for any given image. In this model the encoder used is pre-trained VGG16 model. This model makes use of a recurrent neural network which encodes the variable length input into a fixed dimensional vector and uses this representation to “decode” it to the desired output sentence.

## Chapter 2.3:Literature Review:

* **Deep Learning based Automatic Image Caption Generation**

The aim of the paper [1] is to generate captions to the image which is normally, manually annotated by data annotators. It first creates feature vectors with the help of CNN and later uses RNN for creation of sentences with the help of features gained before. For the purpose of automated captioning, a pre-trained model called VGG16 model is being used. This model makes use of a recurrent neural network which encodes the variable length input into a fixed dimensional vector and uses this representation to “decode” it to the desired output sentence. An encoder is a process of extracting vectors which describe contents of an image. A decoder reverses the process of encoding. Decoder process uses layers like tokenizer, embedding, GRU and dense layer. The paper also points few previous works done on image captioning. The paper[1] uses 2 approaches for obtaining image captioning with the same dataset i.e. MS-COCO, one without using Attention Model and one using Attention Model. Finally, the paper concludes with important points like different epochs used for different models, deeper network constitutes to easier image captioning, etc.

* **Image Annotation via deep neural network**

This paper proposes a novel framework of multimodal deep learning where the convolutional neural networks (CNN) with unlabeled data is utilized to pre-train the multimodal deep neural network to learn intermediate representations and provide a good initialization for the network then use backpropagation to optimize the distance metric functions on individual modality

* **Automatic image annotation using DL representation**

In this paper, the last layer of CaffeNet of the CNN based model is replaced with a projection layer to perform regression and the resulting network is trained for mapping images to semantically meaningful word embedding vectors. Advantage of this modelling is: firstly, it does not require dozens of handcrafted features and secondly, the approach is simpler to formulate than any other generative or discriminative models.

* **Show and Tell: A Neural Image Caption Generator**

This paper proposes a network of the same name. In this network, deep convolutional network is used for image classification and sentence generation is done by a powerful Recurrent Neural Network which is trained with the visual input so that RNN can keep track of the objects explained by the text.

## Chapter 2.4:Problem Definition:

The problem introduces a captioning task, which requires a computer vision system to both localize and describe salient regions in images in natural language. The image captioning task generalizes object detection when the descriptions consist of a single word. Given a set of images and prior knowledge about the content find the correct semantic label for the entire images. First, it is necessary to detect objects on the scene and determine the relationships between them and then, express the image content correctly with properly formed sentences. The generated description is still much different from the way people describe images because people rely on common sense and experience, point out important details and ignore objects and relationships that they imply .

## Chapter 2.5:Scope:

The main implication of image captioning is automating the job of some person who interprets the image (in many different fields). Probably, will be useful in cases/fields where text is most used and with the use of this, you can infer/generate text from images. As in, use the information directly from any particular image in a textual format automatically. There are many NLP applications right now, which extract insights/summary from a given text data or an essay etc. The same benefits can be obtained by people who would benefit from automated insights from images. A slightly (not-so) long term use case would definitely be, explaining what happens in a video, frame by frame. Would serve as a huge help for visually impaired people. Lots of applications can be developed in that space. Social media platforms like Facebook can infer directly from the image, where you are ( beach, cafe etc), what you wear (colour) and more importantly what you’re doing also (in a way). It will also be helpful to improve search results of google image search.

## Chapter 2.6:Technology Stack:

1. Keras – For Deep Learning Tasks
2. Tenserflow – For Deep learning Tasks
3. Pandas – To load the dataset and perform various tasks
4. Numpy – To do mathematical functions
5. Flask – For Hosting on Web
6. OpenCV – For Handling Images
7. Matplotlib and Seaborn – For plotting and visualization purposes

## Chapter 2.7 : Benefits for environment and society:

Image Captioning can play a big role for society. Our project can be used for educational purpose for teaching pre-primary children to make them aware with what all entities are present within a picture. Our image captioning model can be used for enhancement of products like Google Lens. Google Lens is used by users to identify objects and provide relative e-commerce links. With our project imbibed, Lens can also explain the scenario to a confused user.