A Project Report on

IMAGE CAPTIONING

Submitted in partial fulfillment of the requirements for the award of the degree of

Bachelor of Engineering

in

Computer Engineering

by

Shyamkrishna Menon(18102014)
Atharva Ranade(18102016)
Siddharth Nair(18102044)
Omkar Thavai(18102061)

Under the Guidance of

Prof.Sofiya Mujawar

Department of Computer Engineering
NBA Accredited
A.P. Shah Institute of Technology
G.B.Road,Kasarvadavli, Thane(W), Mumbai-400615
UNIVERSITY OF MUMBAI

Academic Year 2021-2022

Approval Sheet

This Project Report entitled ***"IMAGE CAPTIONING"*** Submitted by ***"Shyamkrishna Menon"(18102014), "Atharva Ranade"(18102016), "Siddharth Nair"(18102044), "Omkar Thavai"(18102061)***is approved for the partial fulfillment of the requirement for the award of the degree of ***Bachelor of Engineering*** in ***Computer Engineering*** from ***University of Mumbai***.

(Prof.Sofiya Mujawar)
Guide

Prof. Sachin Malave
Head Department of Computer

Place:A.P.Shah Institute of Technology, Thane
Date:

CERTIFICATE

This is to certify that the project entitled ***"IMAGE CAPTIONING"*** submitted by ***"Shyamkrishna Menon" (18102014),"Atharva Ranade" (18102016),"Siddharth Nair" (18102044),"Omkar Thavai" (18102061)*** for the partial fulfillment of the requirement for award of a degree ***Bachelor of Engineering*** in ***Computer Engineering***,to the University of Mumbai,is a bonafide work carried out during academic year 2020-2021.

(Prof Sofiya Mujawar)
Guide

Prof. Sachin Malave
Head Department of Computer

Dr. Uttam D.Kolekar
Principal

External Examiner(s)

1.

2.

Place:A.P.Shah Institute of Technology, Thane
Date:

Acknowledgement

We have great pleasure in presenting the report on **IMAGE CAPTIONING.** We take this opportunity to express our sincere thanks towards our guide **Prof.Sofiya Mujawar** , APSIT thane for providing the technical guidelines and suggestions regarding line of work. We would like to express our gratitude towards his constant encouragement, support and guidance through the development of project.

We thank **Prof. Sachin Malave** Head of Department,Computer, APSIT for his encouragement during progress meeting and providing guidelines to write this report.

We thank **Prof. Amol Kalugade** BE project co-ordinator, Department of Computer, APSIT for being encouraging throughout the course and for guidance.

We also thank the entire staff of APSIT for their invaluable help rendered during the course of this work. We wish to express our deep gratitude towards all our colleagues of APSIT for their encouragement.

Student Name1:Shyamkrishna Menon
Student ID1:18102014

Student Name2:Atharva Ranade
Student ID2:18102016

Student Name3:Siddharth Nair
Student ID3:18102044

Student Name4:Omkar Thavai
Student ID4:18102061

Declaration

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, We have adequately cited and referenced the original sources. We also declare that We have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

_____

(Signature)

_____

(Shyamkrishna Menon 18102014)
(Atharva Ranade 18102016)
(Siddharth Nair 18102044)
(Omkar Thavai 18102061)

Date:

Contents

Chapter 1

Introduction

Caption generation is an interesting artificial intelligence problem where a descriptive sentence is generated for a given image. It involves the dual techniques from computer vision to understand the content of the image and a language model from the field of natural language processing to turn the understanding of the image into words in the right order. Image captioning has various applications such as recommendations in editing applications, usage in virtual assistants, for image indexing, for visually impaired persons, for social media, and several other natural language processing applications. Recently, deep learning methods have achieved state-ofthe-art results on examples of this problem. It has been demonstrated that deep learning models are able to achieve optimum results in the field of caption generation problems. Instead of requiring complex data preparation or a pipeline of specifically designed models, a single end-to-end model can be defined to predict a caption, given a photo.

Chapter 2

Project Concept

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

2.2 Objectives

The goal of this image captioning project 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 predict a caption , and deploy it on web using Flask API.

2.3 Literature Review

Deep Learning based Automatic Image Caption Generation The aim of the paper 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.

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.

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 .

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.

- 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.

2.6 Technology Stack

- Deep learning
  - Train the model using Convolutional Neural Networks and Recurrent Neural Networks (deep learning model) to detect features from image and predict the captions respectively.
- Google Colab
- Programming language - Python
  - Flask - An API of Python used to build web-applications(front - end)
  - Pandas - Python library for data manipulation and analysis.
  - opencv - Python Library to load images.
  - numpy - Python Library For mathematical operations
- Keras Framework(Using Tenserflow Backend)
  - Used for building our model architecture for Image Captioning and also used for importing VGG-16 for Transfer Learning

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.

Chapter 3

Project Design

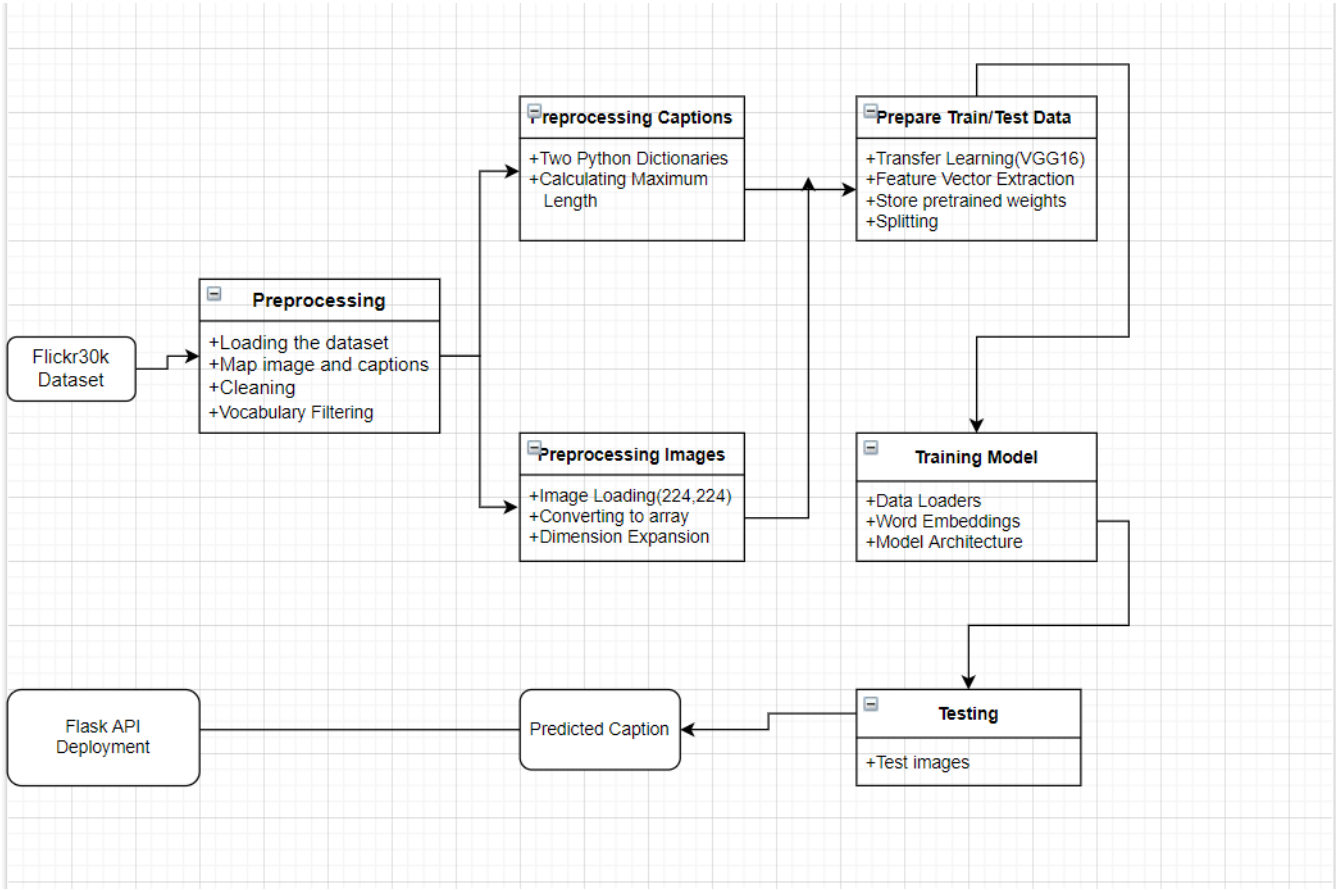
3.1 Proposed System

- First we will import Flickr 30k dataset and process. Flicker datasets are used for image captioning. 30k stands for 30,000 images of various instances.

- We use VGG16 model for image captioning. VGG16 is used for embedding of features within the image like identifying a person, thing, etc and LSTM is used for encapsulating all features and describing it as a sentence.

3.2 Design(Flow Of Modules)

3.3 Class Diagram



3.4 Modules

3.4.1 Module 1-Read Dataset

We read the Flickr 30k dataset

3.4.2 Module 2- Processing of data

- There are few sub-steps for implementing processing of the dataset
- We first input the .csv file. This file will have ImageID of the image and associated captions. An image can have multiple captions.
- A description (dictionary) is created where ImageID is considered as key and caption is value.
- Since caption can have few unwanted characters eg. , etc, we remove them. All the cleaned captions are inputted into a text file. We create a vocabulary i.e. count of each word in a caption and count of each word in the whole text document.
- A threshold is set so as to keep relevant words only, relevance is found if word count is greater than or equal to threshold.
- We create a train and test dataset.
- Now a start and end sequence has been created.

3.4.3 Module 3- Transfer Learning

We use VGG16 for feature capture. VGG16 is a convolution neural net (CNN) architecture. It is considered to be one of the excellent vision model architecture till date.

VGG16 : VGG16 is a convolutional neural network model proposed by K. Simonyan and A. Zisserman from the University of Oxford in the paper "Very Deep Convolutional Networks for Large-Scale Image Recognition". The model achieves 92.7 percentage top-5 test accuracy in ImageNet, which is a dataset of over 14 million images belonging to 1000 classes. It makes the improvement over AlexNet by replacing large kernel-sized filters (11 and 5 in the first and second convolutional layer, respectively) with multiple 3×3 kernel-sized filters one after another.

3.4.4 Module 4 - Training and prediction

- We first load our data into batches for word embedding process. Word embedding will help us find feature vectors for the words in the training set.

- Then a model architecture is been formed. We will use LSTM for sentence formation from the features of caption.

- Lastly we predict test sets captions.

NEED FOR LSTM:

In every data point, it's not just the image which goes as input to the system, but also, a partial caption which helps to predict the next word in the sequence.Since we are processing sequences, we will employ a Recurrent Neural Network to read these partial captions.LSTM is a more powerful RNN architecture, so we use it in this project So, VGG16 will give us the feature vector, which together with captions will go as input and then our model will predict the caption.

3.5 References

- V.Keshavan et al., "Deep Learning based Automatic Image Caption Generation"

- O.Vinyals et al., "Show and Tell: A Neural Image Caption Generator"

- S.Chengjian et al., "Image Annotation Via Deep Neural Network"

- V.Murthy et al ., "Automatic Image Annotation using Deep learning representations".

Chapter 4

Planning for next semester

- We will incorporate our model into web. In this manner, anyone can upload an image and a suitable caption will be provided to it.

- We will find perfect set of parameters during model architecture.