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 2021-2022.

(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

List of Figures

Chapter 1

Project Concept

1.1 Abstract

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. This paper 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 Recurrent Neural Network (RNN) decoder is used to translate the features and objects given by our image to obtain a sequential, meaningful description of the image. In deep learning, a convolutional neural network (CNN/ConvNet) is a class of deep neural networks, most commonly applied to analyse visual imagery. Now when we think of a neural network, we think about matrix multiplications but that is not the case with ConvNet. It uses a special technique called Convolution. Now in mathematics convolution is a mathematical operation on two functions that produces a third function that expresses how the shape of one is modified by the other. A recurrent neural network (RNN) is a class of artificial neural networks where connections between nodes form a directed or undirected graph along a temporal sequence. This allows it to exhibit temporal dynamic behaviour. Derived from feedforward neural networks, RNNs can use their internal state (memory) to process variable length sequences of inputs. This makes them applicable to tasks such as unsegmented, connected handwriting recognition or speech recognition. Recurrent neural networks are theoretically Turing complete and can run arbitrary programs to process arbitrary sequences of inputs.

1.2 Introduction

A large amount of information is stored in an image. Every day, image data that are generated are enormous in amounts. These data are generated by social media, CCTV footage, etc. generating captions manually thus becomes a tedious task. Deep learning can be used to automatically annotate these images, thus replacing the manual annotations done. This will greatly reduce 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 enable 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 . Image captioning can also be used for educational purposes for teaching pre-primary children to make them aware of what all entities are present within a picture. The image captioning model can be used for the 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. The field brings together state-of-the-art models in Natural Language Processing and Computer Vision, two of the major fields in Artificial Intelligence. There are many Natural Language Processing (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. One of the challenges is the availability of a large number of images with their associated text on the ever-expanding internet. Generating captions automatically from images is a complex task as it entails the model extracting features from the images and then forming a meaningful sentence from the available features . Basically, the feature extraction is done by training a Convolutional Neural Network (CNN) with a 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 .

1.3 Objectives

The goal of our project: Image Captioning is to convert a given input image into a natural language description. we will be using the concept of CNN and LSTM and build a model of Image Caption Generator which involves the concept of computer vision and Natural Language Process to recognize the context of images and describe them in natural language like English. The task of image captioning can be divided into two modules logically –

- Image based model — Extracts the features of our image.
- Language based model — which translates the features and objects extracted by our image based model to a natural sentence.

1.4 Literature Review

- Deep Learning based Automatic Image Caption Generation[1]- // 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 the 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 [1] makes use of a RNN which encodes the variable length input into a fixed dimensional vector and uses this representation to "decode" it to the desired output sentence [1] [4]. 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 [2]- The authors of this paper [2] have proposed a deep learning framework. A novel framework of multimodal deep learning where the CNNs with unlabeled data are 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[1] [2]. NUS-WIDE dataset is being used in the paper. The proposed framework consist of a unified two-stage learning path where (i) learning to fune-tune the parameters of deep neural network with respect to each individual modality, and (ii) learning to find the optimal combination of diverse modalities simultaneously in a coherent process[2].

- Automatic image annotation using DL representation[3]-
  In this paper [3], the authors propose a model for image annotation. They have used the Canonical Correlation Analysis (CCA) framework and have reported the results of all 3 variants of CCA i.e. linear CCA, kernel CCA and CCA with k-nearest neighbor (CCA-KNN) [3]. In the CNN based model (which is used for feature extraction and word embedding vectors for the representation of associated tags) 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. The advantage of this modeling 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 [3] [1]. Finally, the paper concludes by stating that CCA-KNN Model provides the best results.

- Show and Tell: A Neural Image Caption Generator [4]=-
  This paper [4] proposes a network of the same name. the model is used for the generation of captions from images using computer vision and machine translation. Here CNN is used for feature extraction and RNN helps in sentence formation [1] [4]. Pascal, MS-COCO [13], and Flickr30k [12] are some of the datasets that are being used. BLEU [10] scores are used as the evaluation metric.

- E. An Empirical Study of Language CNN for Image Captioning [5]-
  In this paper [5], the authors have introduced a language CNN model which is suitable for statistical language modelling tasks and shows competitive performance in image captioning. The primary contribution lies in incorporating a language CNN, which is very powerful for text representation [5] [8] [9], is capable of capturing long-range dependencies in sequences, with RNNs for image captioning. The model yields comparable performance with the state-of-the-art approaches on Flickr30k [12] and MS COCO [13] which validate the proposal and analysis of the experiments conducted. Performance improvements are clearly observed when compared with other image captioning methods.

1.5 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. Computer vision is a field of artificial intelligence (AI) that enables computers and systems to derive meaningful information from digital images, videos and other visual inputs. Image Captioning is the process of generating textual description of an image. 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.

1.6 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 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.

1.7 Technology Stack

- Deep learning
  Deep learning is a machine learning technique that teaches computers to do what comes naturally to humans. Deep learning is a key technology behind driverless cars, enabling them to recognize a stop sign, or to distinguish a pedestrian from a lamppost. Train the model using Convolutional Neural Networks and Recurrent Neural Networks (deep learning model) to detect features from image and predict the captions respectively.

- Kaggle
  Kaggle, a subsidiary of Google LLC, is an online community of data scientists and machine learning practitioners. Kaggle allows users to find and publish data sets, explore and build models in a web-based data-science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges. We have used Flickr30k dataset from Kaggle and also used the built-in Kaggle Juypter notebook. The Flickr30k dataset contains 31,000 images collected from Flickr, together with 5 reference sentences provided by human annotators. he Flickr30k dataset has become a standard benchmark for sentence-based image description. This paper presents Flickr30k Entities, which augments the 158k captions from Flickr30k with 244k coreference chains, linking mentions of the same entities across different captions for the same image, and associating them with 276k manually annotated bounding boxes. Such annotations are essential for continued progress in automatic image description and grounded language understanding. They enable us to define a new benchmark for localization of textual entity mentions in an image. We present a strong baseline for this task that combines an image-text embedding, detectors for common objects, a color classifier, and a bias towards selecting larger objects. While our baseline rivals in accuracy more complex state-of-the-art models, we show that its gains cannot be easily parlayed into improvements on such tasks as image-sentence retrieval, thus underlining the limitations of current methods and the need for further research. Jupyter notebooks consist of a sequence of cells, where each cell is formatted in either Markdown (for writing text) or in a programming language of your choice (for writing code). To start a notebook, click on “Create Notebook”, and select “Notebook”. This will open the Notebooks editing interface. Notebooks may be written in either R or Python.

- Programming language - Python
  Python is a general-purpose programming language started by Guido van Rossum that became very popular very quickly, mainly because of its simplicity and code readability. It enables the programmer to express ideas in fewer lines of code without reducing readability. Compared to languages like C/C++, Python is slower. That said, Python can be easily extended with C/C++, which allows us to write computationally intensive code in C/C++ and create Python wrappers that can be used as Python modules. This gives us two advantages: first, the code is as fast as the original C/C++ code (since it is the actual C++ code working in background) and second, it easier to code in Python than C/C++

  - Flask - An API of Python used to build web-applications(front - end)
    Flask is a micro web framework written in Python. It is classified as a microframework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions. However, Flask supports extensions that can add application features as if they were implemented in Flask itself. Extensions exist for object-relational mappers, form validation, upload handling, various open authentication technologies and several common framework related tools.

  - Pandas - Python library for data manipulation and analysis.
    Pandas is an open source Python package that is most widely used for data science/data analysis and machine learning tasks. It is built on top of another package named Numpy, which provides support for multi-dimensional arrays. As one of the most popular data wrangling packages, Pandas works well with many other data science modules inside the Python ecosystem, and is typically included in every Python distribution, from those that come with your operating system to commercial vendor distributions like ActiveState’s ActivePython.

- opencv - Python Library to load images.
  OpenCV-Python is a library of Python bindings designed to solve computer vision problems. OpenCV-Python makes use of Numpy, which is a highly optimized library for numerical operations with a MATLAB-style syntax. All the OpenCV array structures are converted to and from Numpy arrays. This also makes it easier to integrate with other libraries that use Numpy such as SciPy and Matplotlib.

- numpy - Python Library For mathematical operations
  NumPy is a Python library used for working with arrays. It also has functions for working in ‘domain of linear algebra, fourier transform, and matrices. NumPy was created in 2005 by Travis Oliphant. It is an open source project and you can use it freely. NumPy stands for Numerical Python.

- Keras Framework(Using Tenserflow Backend)- Used for building our model architecture for Image Captioning and also used for importing VGG-16 for Transfer Learning
  Keras is a high-level, deep learning API developed by Google for implementing neural networks. It is written in Python and is used to make the implementation of neural networks easy. It also supports multiple backend neural network computation. Keras is relatively easy to learn and work with because it provides a python frontend with a high level of abstraction while having the option of multiple back-ends for computation purposes. This makes Keras slower than other deep learning frameworks, but extremely beginner-friendly.

1.8 Benefits For Environment And Society

The AI-powered image captioning model is an automated tool that generates concise and meaningful captions for prodigious volumes of images efficiently. The model employs techniques from computer vision and Natural Language Processing (NLP) to extract comprehensive textual information about the given images.

- Recommendations in Editing Applications-
  The image captioning model automates and accelerates the close captioning process for digital content production, editing, delivery, and archival. Well-trained models replace manual efforts for generating quality captions for images as well as videos.

- Assistance for Visually Impaired-
  The advent of machine learning solutions like image captioning is a boon for visually impaired people who are unable to comprehend visuals. With AI-powered image caption generator, image descriptions can be read out to visually impaired, enabling them to get a better sense of their surroundings.

- Media and Publishing Houses-
  The media and public relations industry circulate tens of thousands of visual data across borders in the form of newsletters, emails, etc. The image captioning model accelerates subtitle creation and enables executives to focus on more important tasks.

- Social Media Posts-
  For social media, artificial intelligence is moving from discussion rooms to underlying mechanisms for identifying and describing terabytes of media files. It enables community administrators to monitor interactions and analysts to formulate business strategies.

Chapter 2

Project Design

2.1 Proposed System

- First we will import Flickr30k [7] [11] dataset and process. Flicker datasets are used for image captioning. 30k stands for around 30,000 images of various instances. The Flickr30k dataset has become a standard benchmark for sentence-based image description for beginners. It has up to five captions for each image. Note that the Flickr30K Dataset includes images obtained from Flickr. Use of the images must abide by the Flickr Terms of Use. We do not own the copyright of the images. They are solely provided for researchers and educators who wish to use the dataset for non-commercial research and/or educational purposes.

- We use VGG16 [14] model for image captioning. VGG16 is used for embedding of features within the image like identifying a person, thing, etc and LSTM [15] is used for encapsulating all features and describing it as a sentence. Embeddings are numerical representations of Natural Language Understanding (NLU) elements. They are expressed as fixed-dimensional vectors. We say that we embed a token, sentence, or utterance into a vector space called the embedding space. The embedding that we use in this project is called word embedding. We use the GLoVE file provided by google for this purpose. GLoVE is an unsupervised methods based on corpus statistics. There is another way of using embedding. It includes the use of tensorflow/keras packages that support "Embedding layers".

- We have considered our model with thresholds of both 0 (i.e. no threshold) and 10. A threshold is a frequency below which we do not consider a certain word. When the threshold is 10, it means that the frequency of words in the captions of the Flickr30k [7] dataset that are lesser than 10 are eliminated. Thresholds are kept for simplifying the computation of the model by removing unimportant, less recurring words.

2.2 Design(Flow Of Modules)

Our project is designed using the following steps-

- Read dataset- read the Flickr30k [7] dataset.
- Processing of data- firstly, create a dictionary of imageID and descriptions, then create a vocabulary and finally filter out words which are more frequent.
- Transfer learning- use VGG16 [14] for getting vectors for every image.
- Training and Prediction- combine image and caption as input and train the model then predict caption from learned weights during training of model.
- Application- using the Flask API of Python, we create a user-friendly website where the uploaded image by the user will receive a suitable caption.

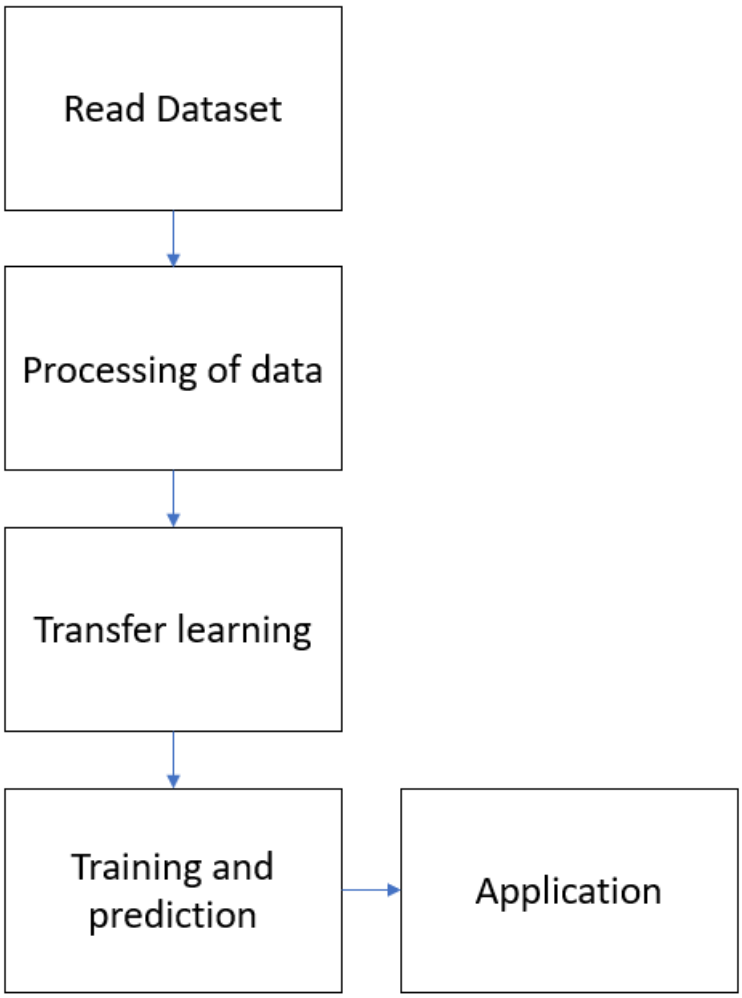

Figure 2.1: Design Flow

2.3 Description Of Use Case

The main implication of image captioning is automating the job of some person who interprets the image (in many different fields). 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 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.With AI-powered image caption generator, image descriptions can be read out to visually impaired, enabling them to get a better sense of their surroundings.

2.4 Class Diagram

The class diagram provides the detailed insight of how the project is implemented.

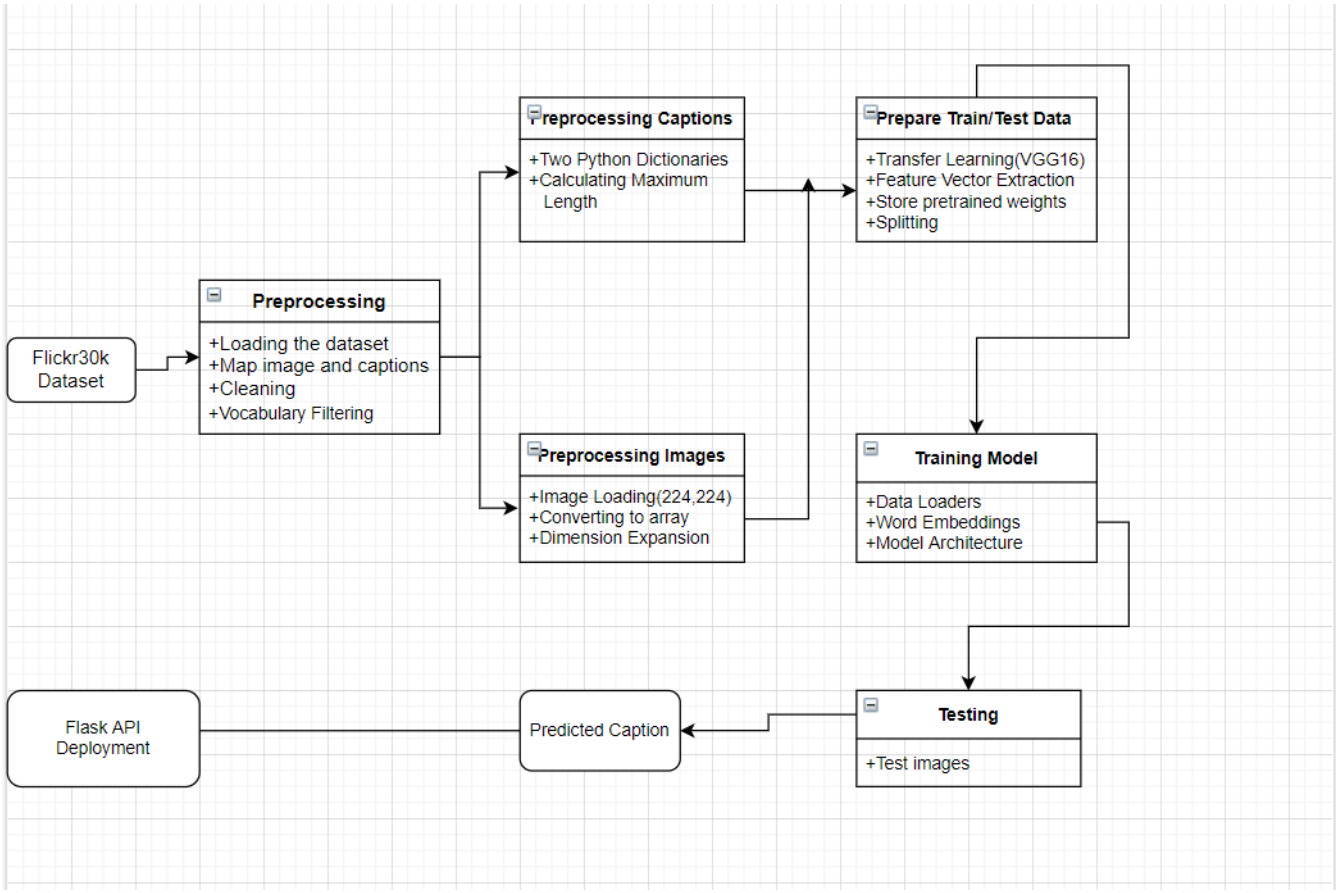

Figure 2.2: Class Diagram

2.5 Modules

2.5.1 Module 1-Read Dataset

We read the Flickr 30k dataset

2.5.2 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 another version of the model where the threshold is not present. This is done for evaluation of captions generated from both the versions for the same image.
- We create a train and test dataset.
- Now a start and end sequence has been created.

2.5.3 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.

2.5.4 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.

2.5.5 Module 5 - Application

Finally we can expect an output that would be apt sentence formation of the given input image. With the help of different features learned, the model will provide words relevant to the image. We create an application for this purpose. Our idea is that a user will upload an image and our model will be able to obtain suitable captions for the same. We use the flask framework of python for uploading the instance of the model. This instance will help the application to be a well-functioned one as it is analogous to a brain of a human body. HTML and CSS are used for the implementation of the user interface part.

(sidecap)

Chapter 3

Implementation

3.1 Proposed System Implementation

- 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.

- We have considered our model with thresholds of both 0 (i.e. no threshold) and 10. A threshold is a frequency below which we do not consider a certain word. When the threshold is 10, it means that the frequency of words in the captions of the Flickr30k dataset that are lesser than 10 are eliminated. Thresholds are kept for simplifying the computation of the model by removing unimportant, less recurring words.

3.1.1 Algorithms

- VGG16 for Transfer Learning:
  Transfer learning is a machine learning method where a model developed for a task is reused as the starting point for a model on a second task.

  It is a popular approach in deep learning where pre-trained models are used as the starting point on computer vision and natural language processing tasks given the vast compute and time resources required to develop neural network models on these problems and from the huge jumps in skill that they provide on related problems.

  VGG16 is a convolution neural net (CNN) architecture which was used to win ILSVR(Imagenet) competition in 2014. It is considered to be one of the excellent vision model architecture till date. Most unique thing about VGG16 is that instead of having a large number of hyper-parameter they focused on having convolution layers of 3x3 filter with a stride 1 and always used same padding and maxpool layer of 2x2 filter of stride 2. It follows this arrangement of convolution and max pool layers consistently throughout the whole architecture. In the end it has 2 FC(fully connected layers) followed by a softmax for output. The 16 in VGG16 refers to it has 16 layers that have weights. This network is a pretty large network and it has about 138 million (approx) parameters.

- 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. They have been used extensively for the task of object detection, self driving cars, image captioning etc. First convnet was discovered in the year 1990 by Yann Lecun and the architecture of the model was called as the LeNet architecture. A basic convnet is shown in the fig. below

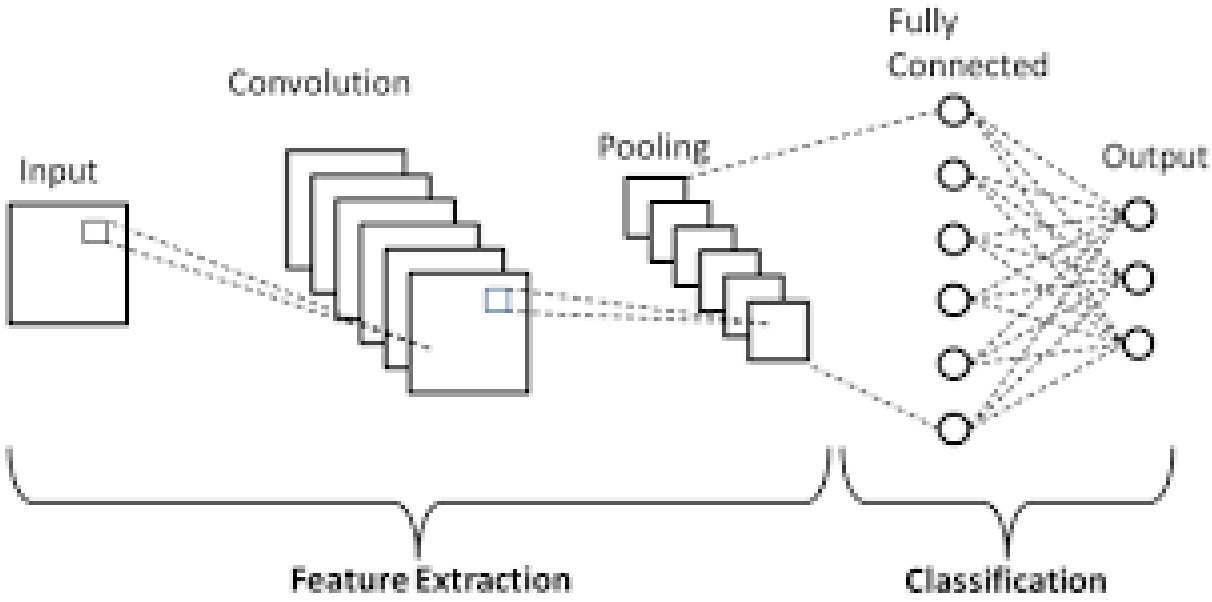

Figure 3.1: Simple CNN architecture

The convolution layer consists of 3*3 filters and the stride length is fixed at 1. Max pooling is done using 2*2-pixel window with a stride length of 2. All the images need to be converted into 224*224-dimensional image. A Rectified Linear Unit (ReLU) activation function is follows every convolution layer.

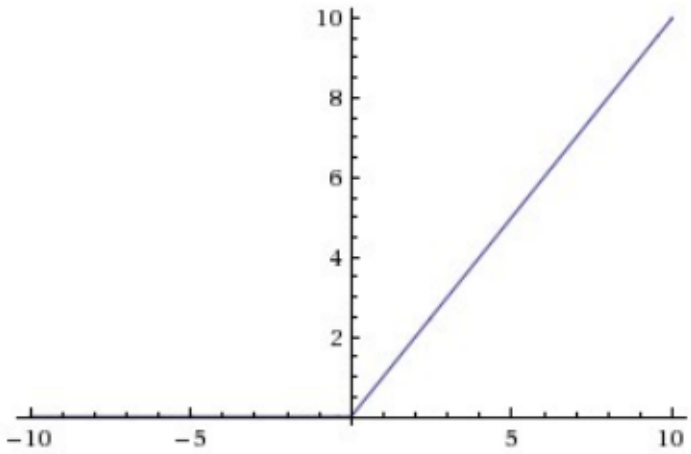

Figure 3.2: Relu Function

- Recurrent Neural Network:
  Recurrent neural nets are a type of artificial neural network in which connection between units form a directed cycle. The advantage of using RNN over conventional feed forward net is that the RNN can process arbitrary set of inputs using its memory. RNNs were discovered in the year 1980 by John Hopfield who gave the famous Hopfield model. Recurrent neural nets in simple terms can be considered as networks with loops which allows the information to persist in the network.

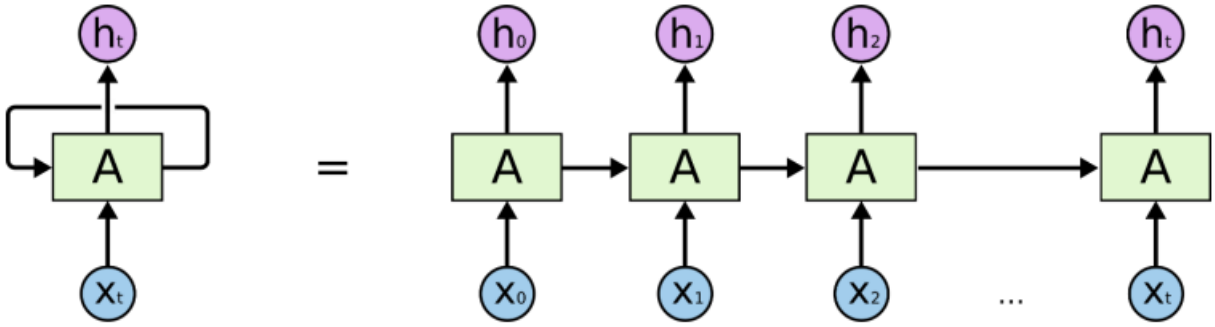

Figure 3.3: Simple RNN architecture

- Long Short-Term Memory(LSTM):
  One of the problems with RNNs is that they do not take long-term dependencies into account. Consider a machine that tries to generate sentences on its own. For instance, the sentence is "I grew up in England, I speak fluent English", if the machine is trying to predict the last word in the sentence i.e. English, the machine needs to know that the language name to be followed by fluent is dependent on the context of the word England. It is possible that the gap between the relevant information and the point where it is needed becomes very large in which case the conventional RNNs fail. To overcome the above-mentioned problem of "long term dependencies", Hochreiter and Schmidhuber proposed the Long Short-Term Memory (LSTM) networks in the year 1997. Since then LSTM networks have revolutionized the fields of speech recognition, machine translation etc. Like the conventional RNNs, LSTMs also have a chain like structure, but the repeating modules have a different structure in case of a LSTM network. A simple LSTM network is shown in Fig. 3.4

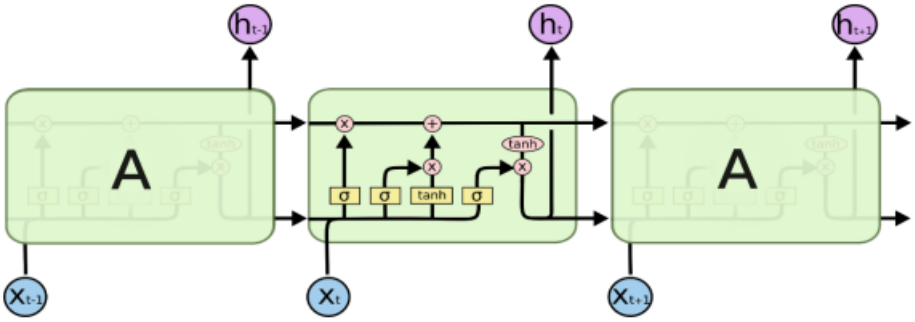

Figure 3.4: Simple LSTM architecture

Finally . The same LSTM network is repeated until an end token (.) is encountered by the network. The series of these word prediction generate the caption for a given image. The complete training process for the combined model (CNN encoder + RNN language generator) and the LSTM network in unravelled form is given below in Fig. 21. The LSTM model is trained to predict each word of the sentence after it has seen the image as well as all preceding words as defined by p(St—I, S0, . . . , St1).

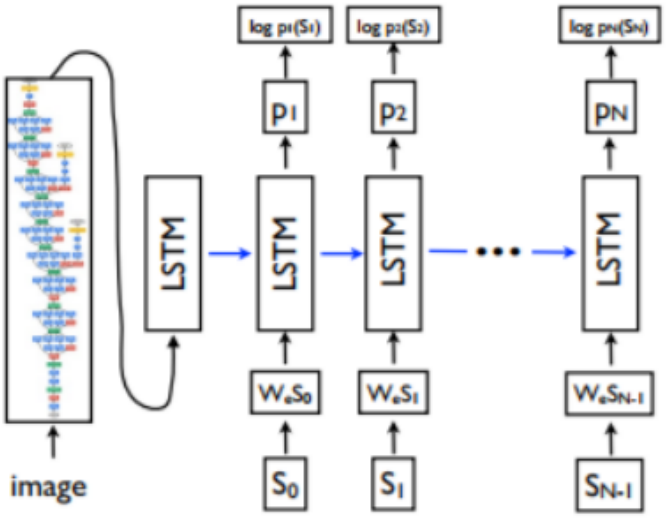

Figure 3.5: Image Captioning Model Layout

The final model architecture is given in fig 3.6

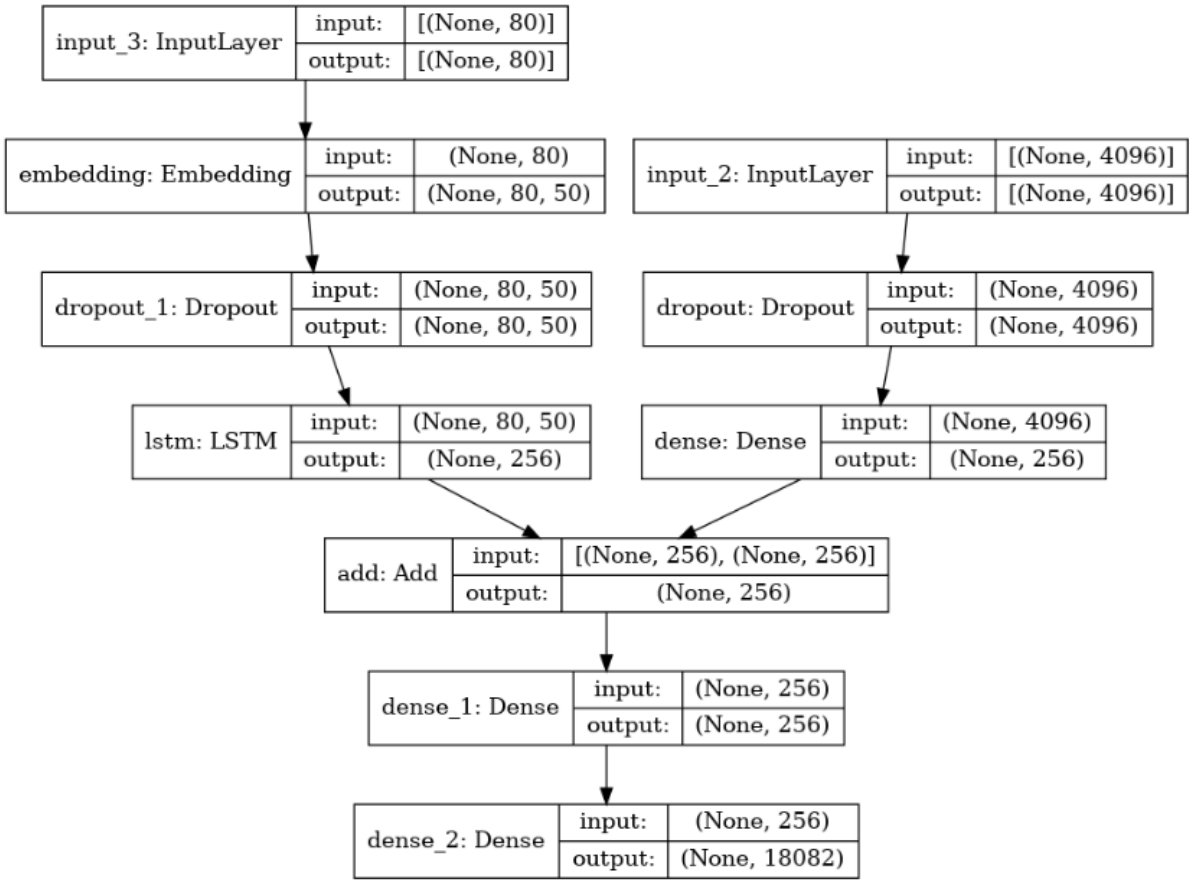

Figure 3.6: Model Architecture

3.1.2 Pseudo Code

- Read dataset- read the Flickr30k dataset.
- Processing of data- firstly, create a dictionary of imageID and descriptions.
- Transfer learning- use VGG16 for getting vectors for every image.
- Word embeddings- preprocess captions and put them into a fixed-length using glove.
- Set Threshold - In this step, set thresholds 0 and 10 and obtain vocabulary for both of them
- Training- combine image and caption as input (obtained from steps 2 and 3 above) and train the model.
- Testing- obtain caption from learned weights during training of the model.
- Deploy on Flask and test on any image of your choice

3.1.3 Platforms For Execution

- Jupyter Notebook : The Jupyter Notebook is an interactive computing environment that enables users to author notebook documents that include: - Live code - Interactive widgets - Plots - Narrative text - Equations - Images - Video

  These documents provide a complete and self-contained record of a computation that can be converted to various formats and shared with others using email, Dropbox, version control systems (like git/GitHub) or nbviewer.jupyter.org.

  Components The Jupyter Notebook combines three components:

  - The notebook web application: An interactive web application for writing and running code interactively and authoring notebook documents.
  - Kernels: Separate processes started by the notebook web application that runs users' code in a given language and returns output back to the notebook web application. The kernel also handles things like computations for interactive widgets, tab completion and introspection.
  - Notebook documents: Self-contained documents that contain a representation of all content visible in the notebook web application, including inputs and outputs of the computations, narrative text, equations, images, and rich media representations of objects. Each notebook document has its own kernel.

- Kaggle : Kaggle, a subsidiary of Google LLC, is an online community of data scientists and machine learning practitioners. Kaggle allows users to find and publish data sets, explore and build models in a web-based data-science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges.
  Kaggle got its start in 2010 by offering machine learning competitions and now also offers a public data platform, a cloud-based workbench for data science, and Artificial Intelligence education. Its key personnel were Anthony Goldbloom and Jeremy Howard. Nicholas Gruen was founding chair succeeded by Max Levchin. Equity was raised in 2011 valuing the company at 25 million dollars. On 8 March 2017, Google announced that they were acquiring Kaggle.
  We used kaggle for Training of the model as we need gpu because the dataset is too massive. You can also use google colab or set up gpu on jupyter notebook and do the same

- Pycharm : PyCharm is an integrated development environment (IDE) used in computer programming, specifically for the Python programming language. It is developed by the Czech company JetBrains (formerly known as IntelliJ). It provides code analysis, a graphical debugger, an integrated unit tester, integration with version control systems (VCSes), and supports web development with Django as well as data science with Anaconda.
  PyCharm is cross-platform, with Windows, macOS and Linux versions. The Community Edition is released under the Apache License,and there is also an educational version, as well as a Professional Edition with extra features (released under a subscription-funded proprietary license)
  We used Pycharm to run the python and flask codes of our front end part where users can upload and image and get the caption

- Chrome : To see our actual image captioning model. Use the server url given by terminal after running flask code

Chapter 4

Results

4.1 Expected Output

a. Expected output would be apt sentence formation of the given input image.We will provide an image to the flask model and it will give a caption for it. We can give unknown images also to the model and get the relevant captures for it.
b. With the help of different features learned, the model will first give relevant features using Transfer Learning , RNN model will provide captions in a proper grammar.

4.2 Obtained Results

For Threshold = 10

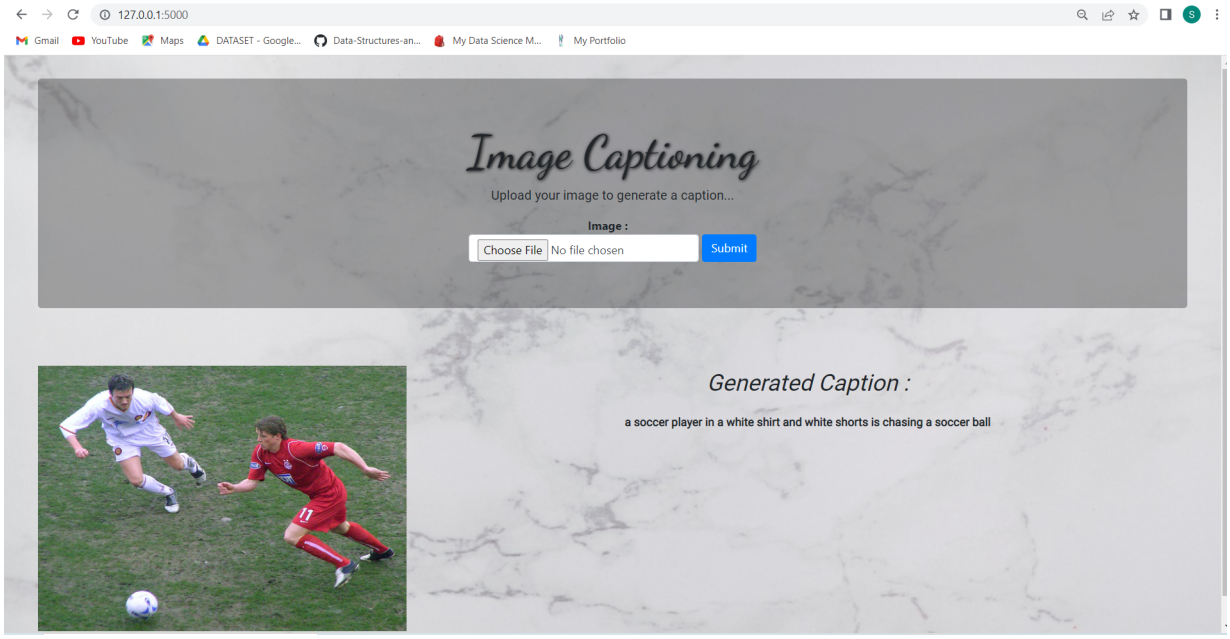

Figure 4.1: Predicted Caption Threshold = 10 img1

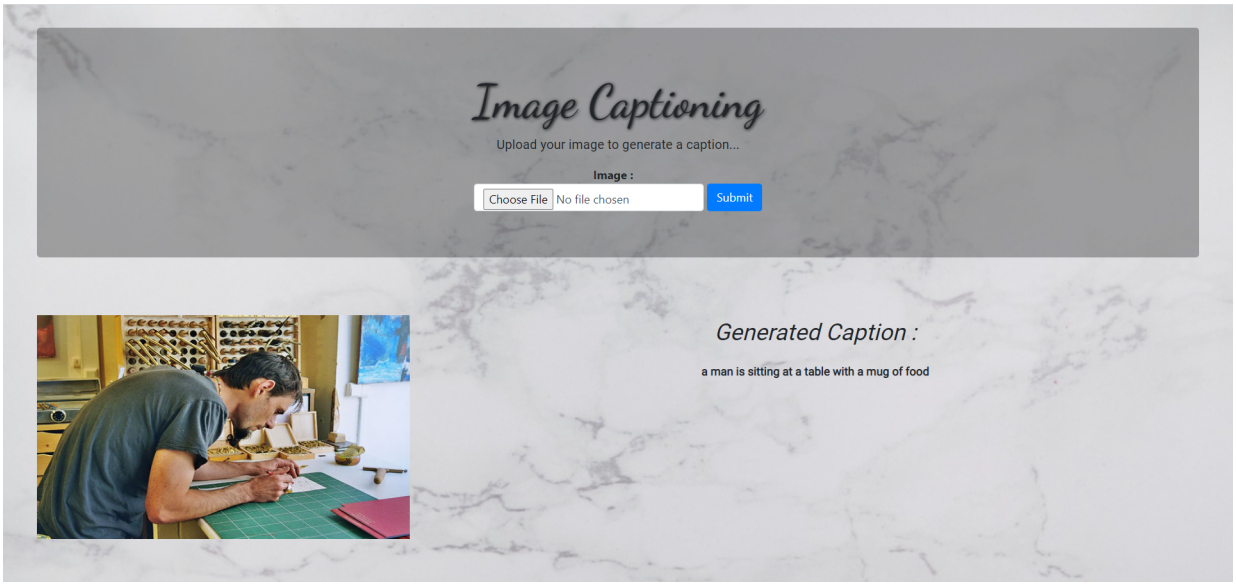

Figure 4.2: Predicted Caption Threshold = 10 img2

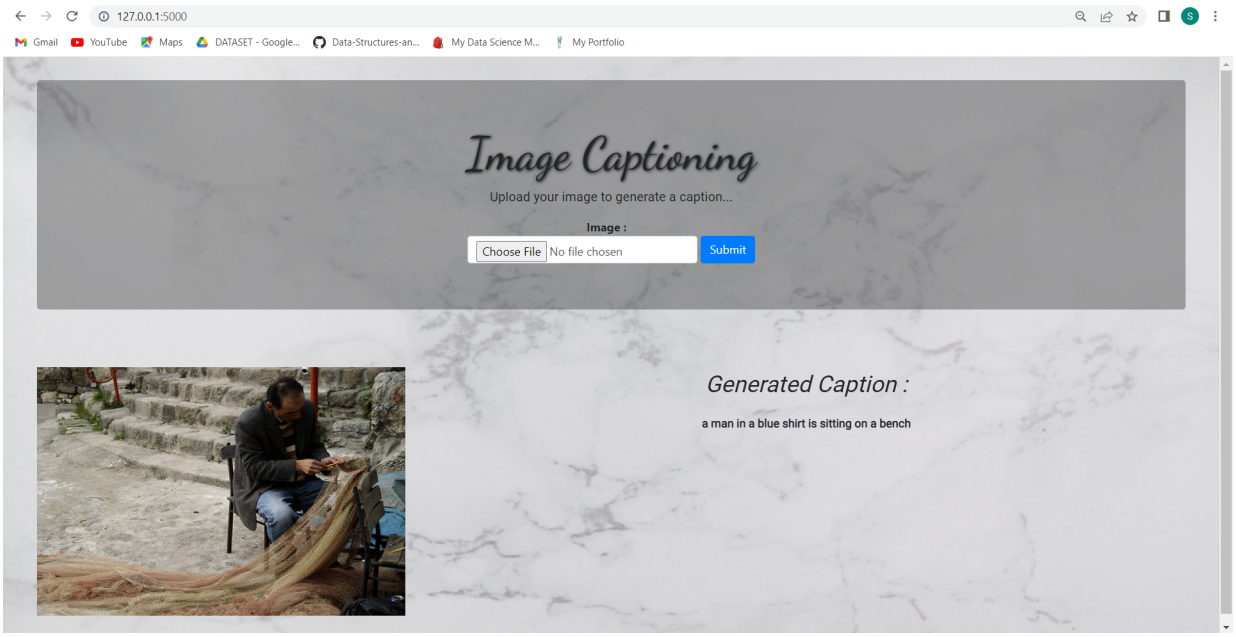

Figure 4.3: Predicted Caption Threshold= 10 img3

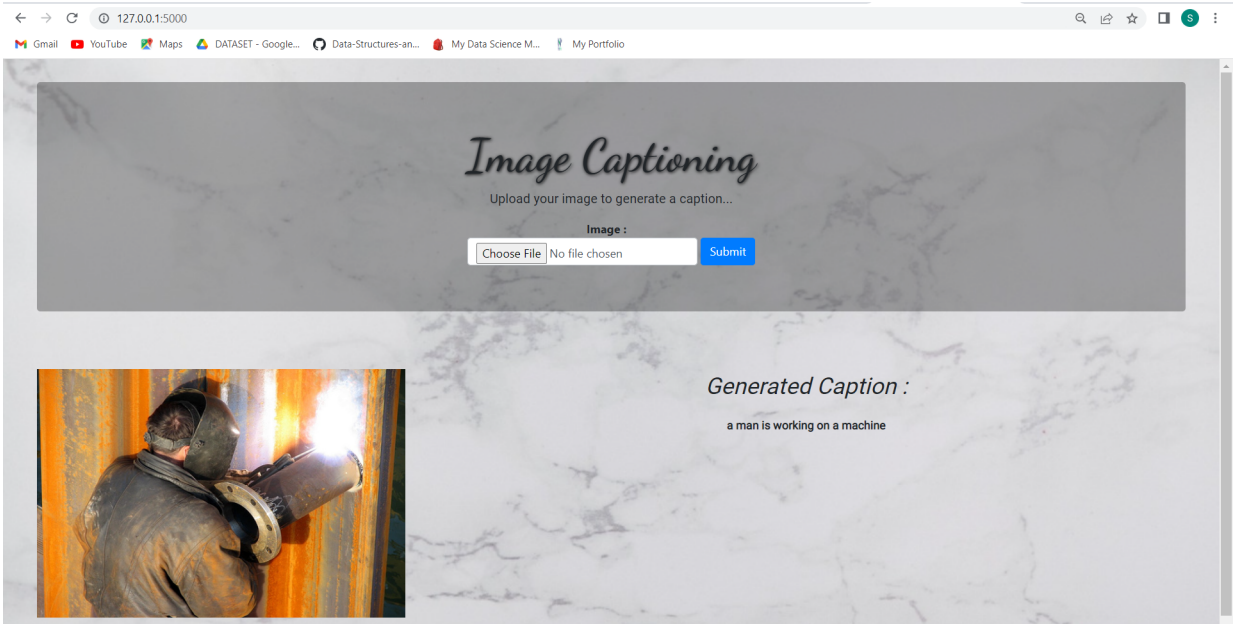

Figure 4.4: Predicted Caption Threshold = 10 img4

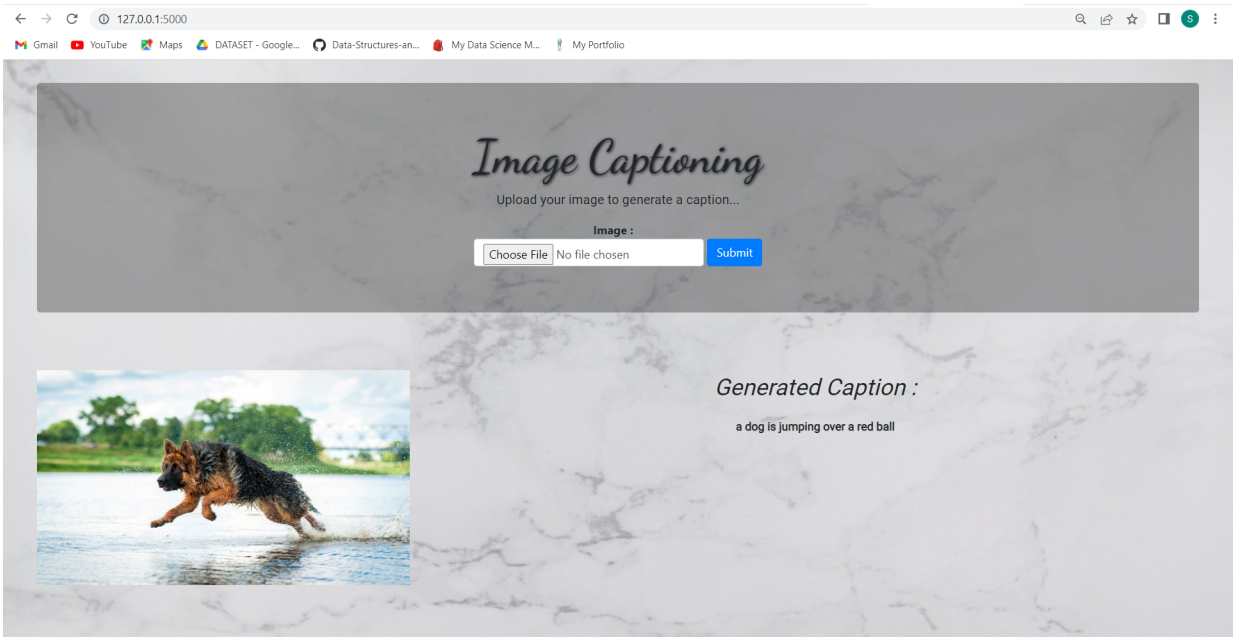

Figure 4.5: Predicted Caption Threshold = 10 img5

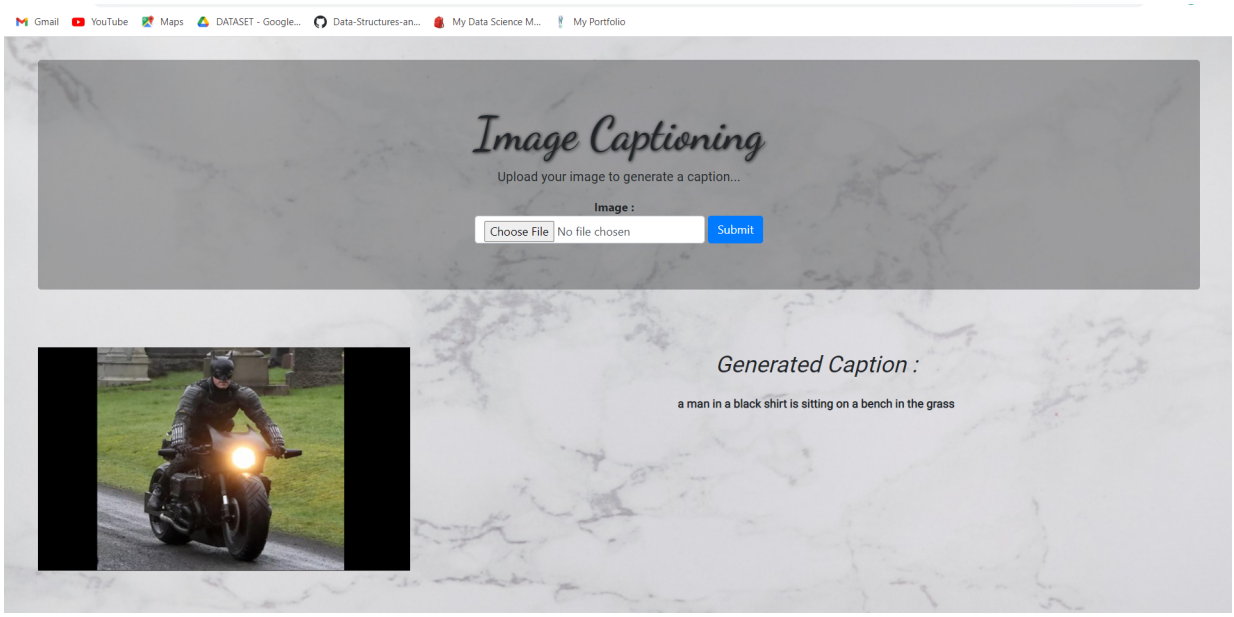

Figure 4.6: Predicted Caption Threshold = 10 img6

For Threshold = 0

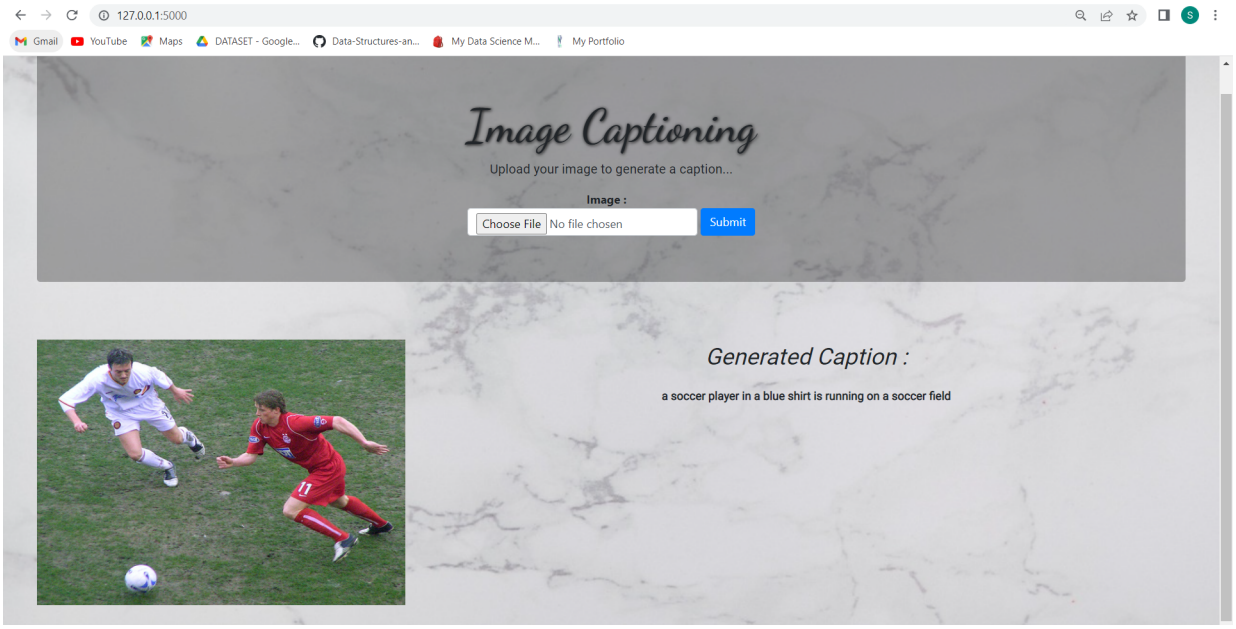

Figure 4.7: Predicted Caption Threshold = 0 img1

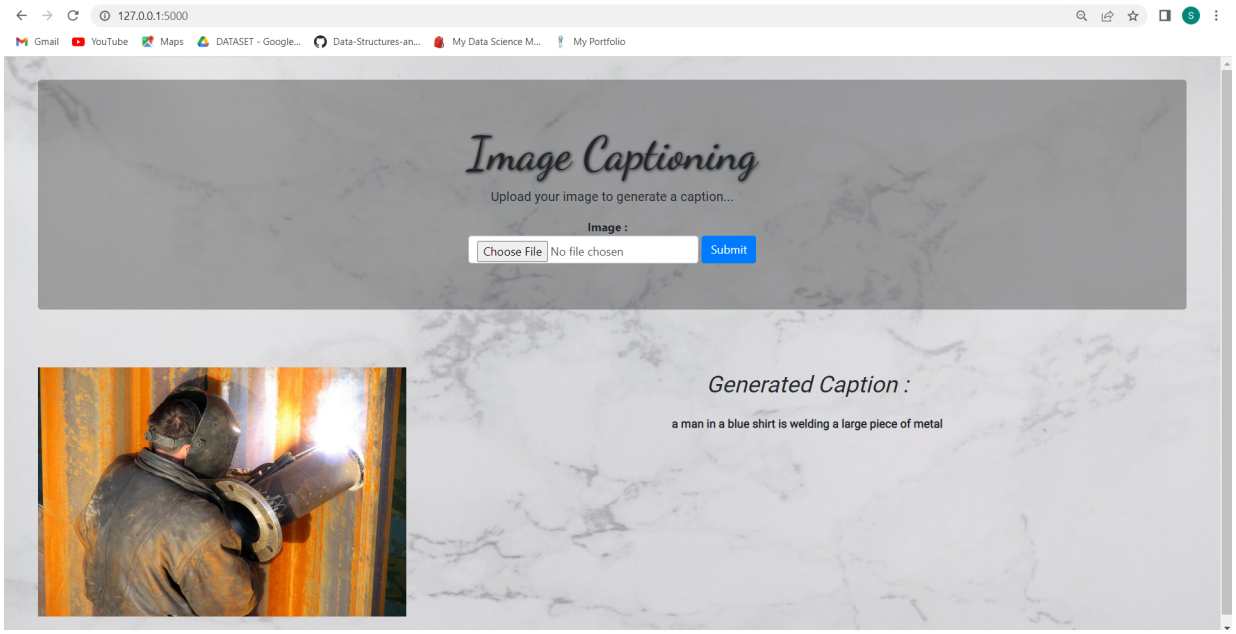

Figure 4.8: Predicted Caption Threshold = 0 img2

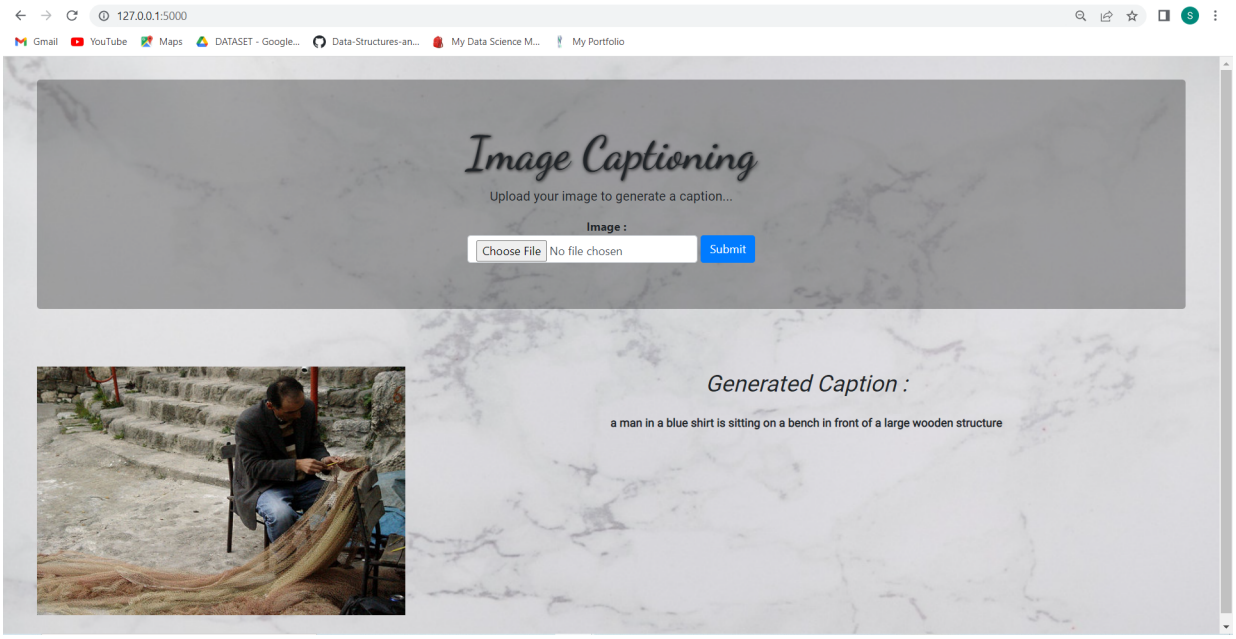

Figure 4.9: Predicted Caption Threshold = 0 img3

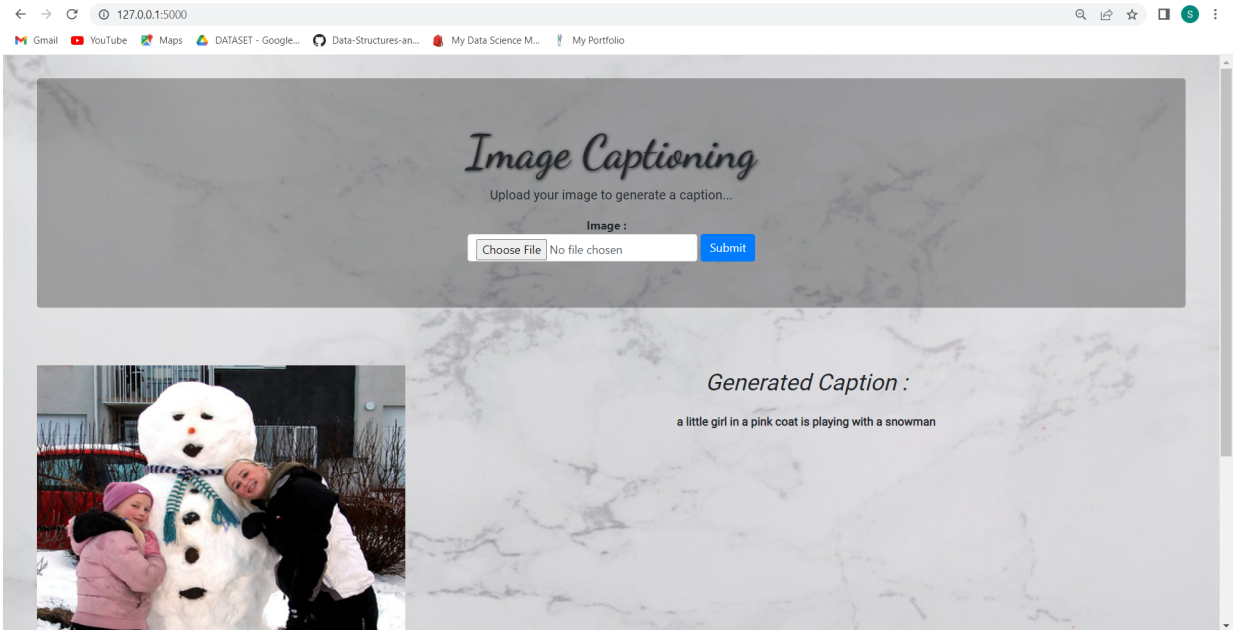

Figure 4.10: Predicted Caption Threshold = 0 img4

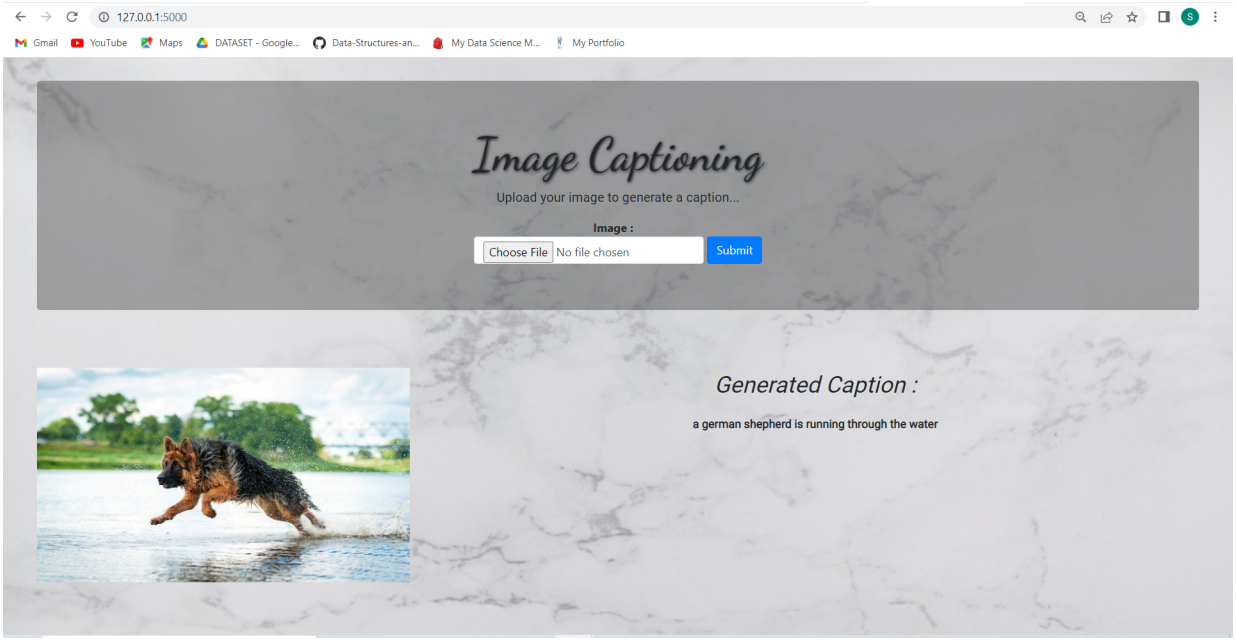

Figure 4.11: Predicted Caption Threshold = 0 img5

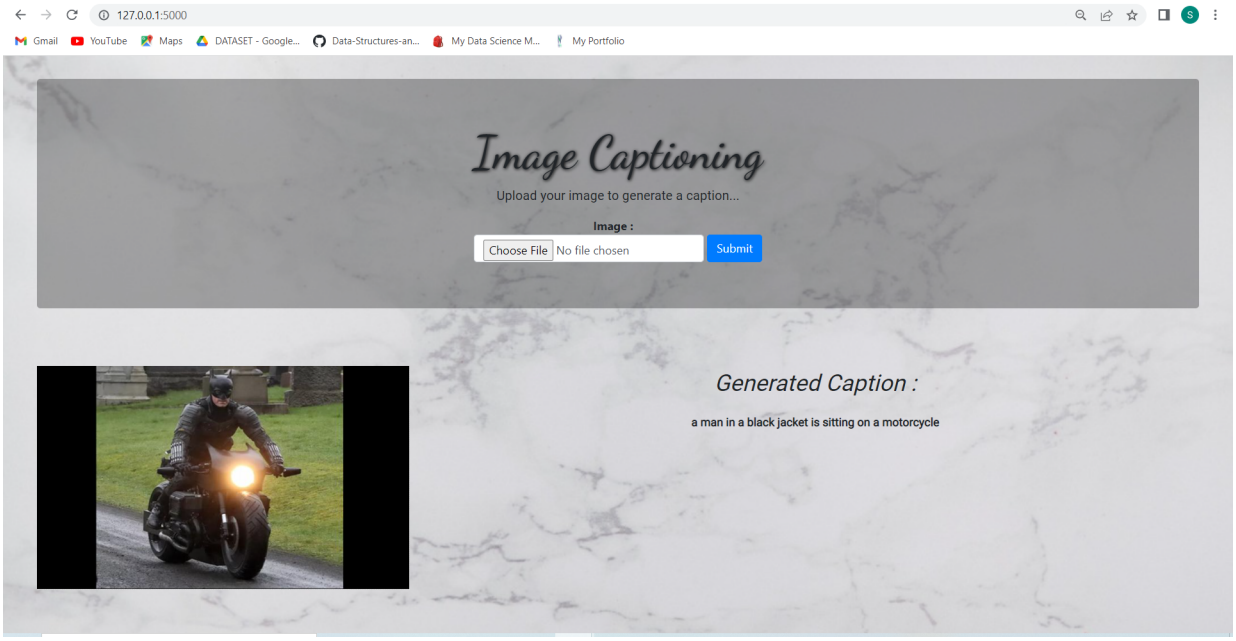

Figure 4.12: Predicted Caption Threshold = 0 img6

With BLEU Scores

```
1945324358
Predicted Caption: a man in a blue shirt and jeans is walking down the street

Reference 1: ['startseq boys wearing jeans and hooded sweatshirts stand on a sidewalk and one of them is holding a black plastic bag  endseq', 'startseq a young man is taking a bag o
'startseq a teenager in a gray sweatshirt is carrying a black trash bag next to a city street  endseq', 'startseq a boy walks down the sidewalk with a trash bag while two other boys
dseq']

BLEU-1: 0.834
```

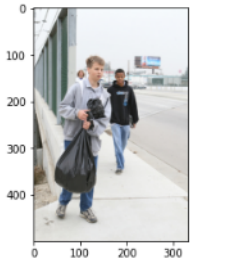

Figure 4.13: Predicted Caption with BLEU img1

```
x = 2558
pic = list(train_encoding.keys())[x]
train_content[pic]


reference = train_content[pic]
#pic = list(encoding_test.keys())[1]
print(pic)
img = 'twodogs.jpg'
e = train_encoding[pic].reshape(1,4096)
#image = encoding_test[pic].reshape((1,2048))
x=plt.imread(images+pic+'.jpg')
#x=plt.imread(img)
plt.imshow(x)
# plt.show()
caption = greedySearch(e)
print("Predicted Caption:",caption)
print()
print('Reference 1:',reference)

print()
print('BLEU-1:', round(sentence_bleu(reference, caption),3))
```

```
1945939774
Predicted Caption: a man in a black shirt is singing into a microphone

Reference 1: ['startseq a man wearing a green sleeveless t shirt red and white and black underpants and blue tights is singing or talking into a microphone  endseq', 'startseq a ma
dseq', 'startseq a young man in a colorful costume sings into a microphone  endseq', 'startseq a man sings on stage with his underwear over his pants  endseq', 'startseq a man sing

BLEU-1: 0.958
```

Figure 4.14: Predicted Caption with BLEU img2

194702176
Predicted Caption: a man is sitting on a boat in a body of water

Reference 1: ['startseq a young male child is floating on a makeshift boat upon a bright blue lake with rocks and light green grasses in the background endseq', 'startseq a boy fl
young boy drifting in small canoe on a calm blue river endseq', 'startseq negro boy in homemade boat on the water endseq', 'startseq a young boy in a small canoe on a lake endse

BLEU-1: 0.686

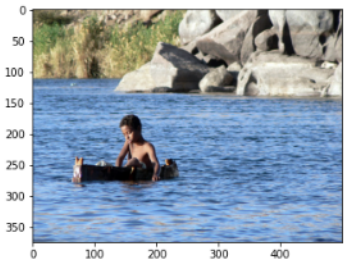

Figure 4.15: Predicted Caption with BLEU img3

194705585
Predicted Caption: a construction worker is working on a roof

Reference 1: ['startseq five construction men working on building the roof frame to a new building endseq', 'startseq construction workers hammer nails on the unfinished roof of a
endseq', 'startseq constructions workers are building a house endseq', 'startseq three men are building a roof endseq']

BLEU-1: 0.839

Figure 4.16: Predicted Caption with BLEU img4

Chapter 5

Conclusion

Image captioning has become a booming topic and how leveraging deep learning concepts has eased the process of annotations. Our paper has used both CNN and RNN for the generation of captions to the inputted image. We have used the glove file for word embedding purposes. Along with the creation of vocabulary, we have considered our model with thresholds of both 0 (i.e. no threshold) and 10. The advantage of our model is that we are able to obtain caption to the input images with few relevant features included in the caption.BLEU score has been used as an evaluation metric for our project.The Bilingual Evaluation Understudy Score, or BLEU for short, is a metric for evaluating a generated sentence to a reference sentence. A perfect match results in a score of 1.0, whereas a perfect mismatch results in a score of 0.0.Our findings show that certain features are not perfectly captured because those features have very little frequency in the dataset. For eg. The colors of T-shirts are captioned as either blue or black if any new color is given to the model. The disadvantage that we have observed is that the model requires hours to process and run to finally obtain image captions. Also, as the dataset consists of lesser images than those that are normally needed for image captioning, the results obtained for some images are not quite as expected. The main focus was to obtain relevant captions for the input image. In the future, the project can be improved to optimize the prediction of captions by training it on bigger datasets and using better computational resources like using GPUs. Usage of attention models can also benefit in the process of obtaining more relevant captions as the attention model will emphasize smaller details in the input image. As mentioned in the introduction, we can build an end-to-end application for visually impaired people who can benefit from our image captioning model by listening to the captions that our model predicts using text-to-speech for the predicted captions.

The model's accuracy can be boosted by deploying it on a larger dataset so that the words in the vocabulary of the model increase significantly. The use of relatively newer architecture, like GoogleNet can also increase the accuracy in the classification task thus reducing the error rate in the language generation. Apart from that the use of bidirectional LSTM network and Gated Recurrent Unit may help in improving the accuracy of the model

Chapter 6

References

[1] Shahar Banu , Seemakousar B , Sanchita S M , Nivedita A, Arun Joshi, Rajeshwari S.G, 2021, varsha, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH TECHNOLOGY (IJERT) Volume 10, Issue 08 (August 2021)
[2] S. Chengjian, S. Zhu and Z. Shi, "Image annotation via deep neural network," 2015 14th IAPR International Conference on Machine Vision Applications (MVA), 2015, pp. 518-521, doi: 10.1109/MVA.2015.7153244.
[3] Venkatesh N.Murty et al, Automatic image annotation using DL representation, ICMR '15: Proceedings of the 5th ACM on International Conference on Multimedia RetrievalJune 2015 Pages 603–606.
[4] O. Vinyals, A. Toshev, S. Bengio and D. Erhan, "Show and tell: A neural image caption generator," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3156-3164, doi: 10.1109/CVPR.2015.7298935.
[5] J. Gu, G. Wang, J. Cai and T. Chen, "An Empirical Study of Language CNN for Image Captioning," 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1231-1240, doi: 10.1109/ICCV.2017.138.
[6] Srinivasan, Lakshminarasimhan and Dinesh Sreekanthan. "Image Captioning-A Deep Learning Approach." (2018).
[7] Kaggle dataset for flickr30k-https://www.kaggle.com/datasets/adityajn105/flickr30k?select=Images [8] N. Kalchbrenner, E. Grefenstette, and P. Blunsom. A convolutional neural network for modelling sentences. ACL, 2014.
[9] M. Wang, Z. Lu, H. Li, W. Jiang, and Q. Liu. gen cnn: A convolutional architecture for word sequence prediction. ACL, 2015.
[10] BLEU: a Method for Automatic Evaluation of Machine Translation. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu IBM T. J. Watson Research Center Yorktown Heights, NY 10598, USA.
[11] M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: data, models and evaluation metrics. Journal of Artificial Intelligence Research, 2013.
[12] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. TACL, 2014.
[13] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick. Microsoft coco: Common objects in context. arXiv preprint arXiv:1405.0312, 2014.
[14] Simonyan, K. and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
[15] S. Hochreiter and J. Schmidhuber, "Long short-term memory", Neural Comput., vol. 9, no. 8, pp. 1735-1780, 1997.
[16] V. Kesavan, V. Muley and M. Kolhekar, "Deep Learning based Automatic Image Caption Generation," 2019 Global Conference for Advancement in Technology (GCAT), 2019, pp. 1-6, doi: 10.1109/GCAT47503.2019.8978293.
[17] https://python.engineering/getting-started-with-jupyter-notebook-python/

Chapter 7

Biblography

[1] https://www.analyticsvidhya.com/blog/2021/12/step-by-step-guide-to-build-image-caption-generator-using-deep-learning/

[2] https://data-flair.training/blogs/python-based-project-image-caption-generator-cnn/

[3] https://www.clairvoyant.ai/blog/image-caption-generator

[4] https://machinelearningmastery.com/develop-a-deep-learning-caption-generation-model-in-python/

[5] https://medium.com/swlh/automatic-image-captioning-using-deep-learning-5e899c127387

[6] https://keras.io/examples/vision/image$_{c}aptioning$/

Chapter 8

Paper Submitted

Image Captioning

[1]Shyamkrishna Menon, [2]Atharva Ranade, [3]Omkar Thavai, [4]Siddharth Nair, [5]Sofiya Mujawar
[1,2,3,4]Student, [5]Assistant Professor, A P Shah Institute of Technology, Thane, India,
[1]reachskmenon2017@gmail.com, [2]ranadeatharva2112@gmail.com,
[3]nair.siddharth01@rediffmail.com, [4]omkarthavai63@gmail.com, [5]ssmujawar@apsit.edu.in

Abstract--This paper 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 a sequential, meaningful description of the image [1] [16].

Keywords — image caption, convolution neural network(CNN), long short term model(LSTM), recurrent neural network(RNN), VGG16, Flickr30k.

I. INTRODUCTION

A large amount of information is stored in an image. Every day, image data that are generated are enormous in amounts. These data are generated by social media, CCTV footage, etc. generating captions manually thus becomes a tedious task. Deep learning can be used to automatically annotate these images, thus replacing the manual annotations done. This will greatly reduce 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 enable the automatic, cost-saving labeling 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 [1]. Image captioning can also be used for educational purposes for teaching pre-primary children to make them aware of what all entities are present within a picture. The image captioning model can be used for the 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. The field brings together state-of-the-art models in Natural Language Processing and Computer Vision, two of the major fields in Artificial Intelligence. There are many Natural Language Processing (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. One of the challenges is the availability of a large number of images with their associated text on the ever-expanding internet. Generating captions automatically from images is a complex task as it entails the model extracting features from the images and then forming a meaningful sentence from the available features [1]. Basically, the feature extraction is done by training a Convolutional Neural Network (CNN) with a 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 [1].

II. LITERATURE SURVEY

A. Deep Learning based Automatic Image Caption Generation [1]

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 the 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 [1] makes use of a RNN which encodes the variable length input into a fixed dimensional vector and uses this representation to "decode" it to the desired output sentence [1] [4]. 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

 DOI : 10.35291/2454-9150.2022.0100

for different models, deeper network constitutes to easier image captioning, etc.

B. Image Annotation via deep neural network [2]

The authors of this paper [2] have proposed a deep learning framework. A novel framework of multimodal deep learning where the CNNs with unlabeled data are 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[1] [2]. NUS-WIDE dataset is being used in the paper. The proposed framework consist of a unified two-stage learning path where (i) learning to fune-tune the parameters of deep neural network with respect to each individual modality, and (ii) learning to find the optimal combination of diverse modalities simultaneously in a coherent process[2].

C. Automatic image annotation using DL representation[3]

In this paper [3], the authors propose a model for image annotation. They have used the Canonical Correlation Analysis (CCA) framework and have reported the results of all 3 variants of CCA i.e. linear CCA, kernel CCA and CCA with k-nearest neighbor (CCA-KNN) [3]. In the CNN based model (which is used for feature extraction and word embedding vectors for the representation of associated tags) 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. The advantage of this modeling 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 [3] [1]. Finally, the paper concludes by stating that CCA-KNN Model provides the best results.

D. Show and Tell: A Neural Image Caption Generator [4]

This paper [4] proposes a network of the same name. the model is used for the generation of captions from images using computer vision and machine translation. Here CNN is used for feature extraction and RNN helps in sentence formation [1] [4]. Pascal, MS-COCO [13], and Flickr30k [12] are some of the datasets that are being used. BLEU [10] scores are used as the evaluation metric.

E. An Empirical Study of Language CNN for Image Captioning [5]

In this paper [5], the authors have introduced a language CNN model which is suitable for statistical language modelling tasks and shows competitive performance in image captioning. The primary contribution lies in incorporating a language CNN, which is very powerful for text representation [5] [8] [9], is capable of capturing long-range dependencies in sequences, with RNNs for image captioning. The model yields comparable performance with the state-of-the-art approaches on Flickr30k [12] and MS COCO [13] which validate the proposal and analysis of the experiments conducted. Performance improvements are clearly observed when compared with other image captioning methods.

F. Image Captioning - A Deep Learning Approach [6]

This paper [6] proposes a fusion between CNN and LSTM. Here CNN is used for building vocabulary and LSTM is used for forming meaningful sequence of words obtained. The efficiency of the proposed model is checked using Flickr30k and Flickr8k datasets and also by utilizing Bleu [10] metric gives superior results in comparison to other state-of-the-art models.

III. EXPERIMENT SET-UP

We use Python as our programming language as it is a popular language when it comes to using deep learning approaches and image processing. We use Deep learning for training the model using Convolutional Neural Networks and Recurrent Neural Networks (deep learning model) to detect features from image and predict the captions respectively. There are few python libraries that we will be using. We use pandas for data manipulation and analysis, opencv for loading images, numpy for mathematical operations, Keras Framework(Using Tenserflow Backend) is used for building our model architecture for Image Captioning and also used for importing VGG-16 for Transfer Learning. All these are implemented in Jupyter Notebook [17] enabling Python 3 language.

IV. PROPOSED SYSTEM

1) Explanation of proposed system

a. First we will import Flickr30k [7] [11] dataset and process. Flicker datasets are used for image captioning. 30k stands for around 30,000 images of various instances.

b. We use VGG16 [14] model for image captioning. VGG16 is used for embedding of features within the image like identifying a person, thing, etc and LSTM [15] is used for encapsulating all features and describing it as a sentence.

c. We have considered our model with thresholds of both 0 (i.e. no threshold) and 10. A threshold is a frequency below which we do not consider a certain word. When the threshold is 10, it means that the frequency of words in the captions of the Flickr30k [7] dataset that are lesser than 10 are eliminated. Thresholds are kept for simplifying

 DOI : 10.35291/2454-9150.2022.0100

the computation of the model by removing unimportant, less recurring words.

2) Block diagram

Fig 1 depicts the block diagram.

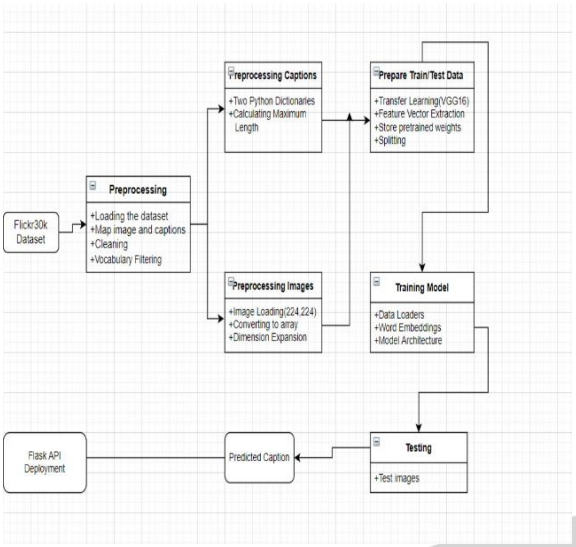

Fig. 1. Block diagram.

This figure represents the workflow of our project starting from loading the dataset to deploying it using Flask.

3) Model algorithm diagram

Fig 2 depicts the Model algorithm diagram

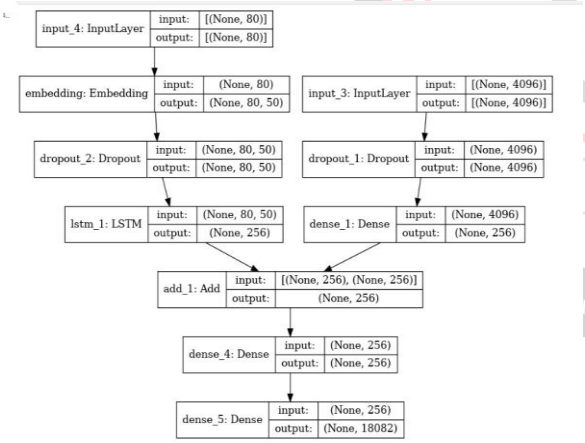

Fig. 2. Model algorithm diagram.

This figure shows the complete architecture of our model.

4) Pseudocode

a. Read dataset- read the Flickr30k [7] dataset.

b. Processing of data- firstly, create a dictionary of imageID and descriptions, then create a vocabulary and finally filter out words that are more frequent.

c. Transfer learning- use VGG16 [14] for getting vectors for every image.

d. Word embeddings- preprocess captions and put them into a fixed-length using glove.

e. Training- combine image and caption as input (obtained from steps c and d) and train the model.

f. Testing- obtain caption from learned weights during training of the model.

5) Expected output

a. Expected output would be apt sentence formation of the given input image.

b. With the help of different features learned, the model will provide words relevant to the image.

V. RESULTS

Fig 3 shows the final results of few images that are obtained by our paper.

Fig. 3. a

Fig. 3. b

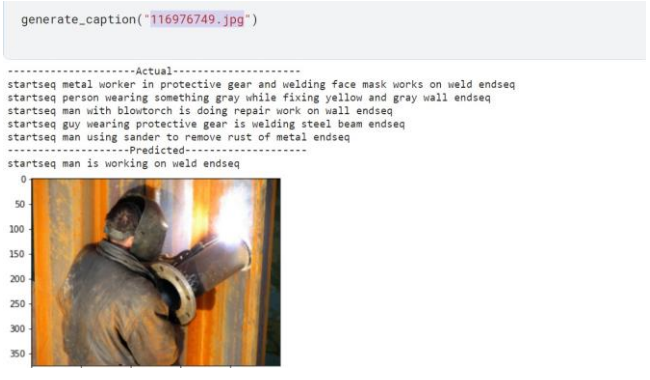

Fig. 3. c

 DOI : 10.35291/2454-9150.2022.0100

Fig. 3. d

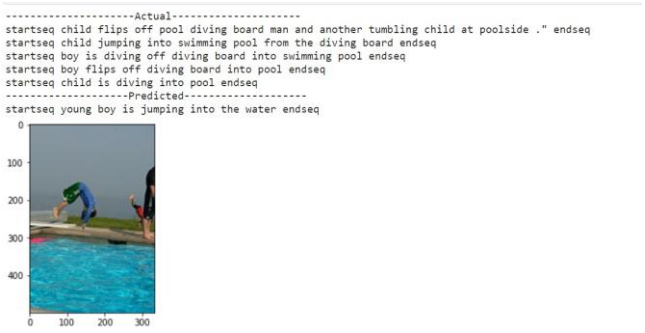

Fig. 3. e

Fig. 3. Few Screenshots of the Final Results obtained through our model.

As seen in figures 3. a,b,c,d, and e, our model has captioned images as per the features that are available in the image.

VI. CONCLUSION

Image captioning has become a booming topic and how leveraging deep learning concepts has eased the process of annotations. Our paper has used both CNN and RNN for the generation of captions to the inputted image. We have used the glove file for word embedding purposes. Along with the creation of vocabulary, we have considered our model with thresholds of both 0 (i.e. no threshold) and 10. The advantage of our model is that we are able to obtain caption to the input images with few relevant features included in the caption. Our findings show that certain features are not perfectly captured because those features have very little frequency in the dataset. For eg. The colors of T-shirts are captioned as either blue or black if any new color is given to the model. The disadvantage that we have observed is that the model requires hours to process and run to finally obtain image captions. Also, as the dataset consists of lesser images than those that are normally needed for image captioning, the results obtained for some images are not quite as expected. The main focus was to obtain relevant captions for the input image. In the future, the project can be improved to optimize the prediction of captions by training it on bigger datasets and using better computational resources like using GPUs. Usage of attention models can also benefit in the process of obtaining more relevant captions as the attention model will emphasize smaller details in the input image. As mentioned in the introduction, we can build an end-to-end application for visually impaired people who can benefit from our image captioning model by listening to the captions that our model predicts using text-to-speech for the predicted captions.

REFERENCES

[1] Shahar Banu , Seemakousar B , Sanchita S M , Nivedita A, Arun Joshi, Rajeshwari S.G, 2021, varsha, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH TECHNOLOGY (IJERT) Volume 10, Issue 08 (August 2021)

[2] S. Chengjian, S. Zhu and Z. Shi, ”Image annotation via deep neural network,” 2015 14th IAPR International Conference on Machine Vision Applications (MVA), 2015, pp. 518-521, doi: 10.1109/MVA.2015.7153244.

[3] Venkatesh N.Murty et al, Automatic image annotation using DL representation, ICMR ’15: Proceedings of the 5th ACM on International Conference on Multimedia RetrievalJune 2015 Pages 603–606.

[4] O. Vinyals, A. Toshev, S. Bengio and D. Erhan, ”Show and tell: A neural image caption generator,” 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3156-3164, doi: 10.1109/CVPR.2015.7298935.

[5] J. Gu, G. Wang, J. Cai and T. Chen, ”An Empirical Study of Language CNN for Image Captioning,” 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1231-1240, doi: 10.1109/ICCV.2017.138.

[6] Srinivasan, Lakshminarasimhan and Dinesh Sreekanthan. “Image Captioning-A Deep Learning Approach.” (2018).

[7] Kaggle dataset for flickr30k- https://www.kaggle.com/datasets/adityajn105/flickr30k?select=Images

[8] N. Kalchbrenner, E. Grefenstette, and P. Blunsom. A convolutional neural network for modelling sentences. ACL, 2014.

[9] M. Wang, Z. Lu, H. Li, W. Jiang, and Q. Liu. gen cnn: A convolutional architecture for word sequence prediction. ACL, 2015.

[10] BLEU: a Method for Automatic Evaluation of Machine Translation. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu IBM T. J. Watson Research Center Yorktown Heights, NY 10598, USA.

[11] M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: data, models and evaluation metrics. Journal of Artificial Intelligence Research, 2013.

[12] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. TACL, 2014.

[13] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick. Microsoft coco: Common objects in context. arXiv preprint arXiv:1405.0312, 2014.

[14] Simonyan, K. and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[15] S. Hochreiter and J. Schmidhuber, ”Long short-term memory”, Neural Comput., vol. 9, no. 8, pp. 1735-1780, 1997.

[16] V. Kesavan, V. Muley and M. Kolhekar, ”Deep Learning based Automatic Image Caption Generation,” 2019 Global Conference for Advancement in Technology (GCAT), 2019, pp. 1-6, doi: 10.1109/GCAT47503.2019.8978293.

[17] https://python.engineering/getting-started-with-jupyter-notebook-python/

 DOI : 10.35291/2454-9150.2022.0100

Chapter 9

Certificate


The Editorial Board of
INTERNATIONAL JOURNAL FOR RESEARCH IN ENGINEERING APPLICATION & MANAGEMENT
ISO 3297:2007 CERTIFIED JOURNAL | ISSN : 2454-9150
Certificate of Publication
This is to Certify that Paper ID : IJREAMV07I1284117
Image Captioning
Authored By
Shyamkrishna Menon
has been published in Volume 07, Issue 12, Mar 2022.
The mentioned paper has gone through Peer Review Process & measured upto the required standard.
*Corresponding Author
DOI : 10.35291/2454-9150.2022.0100
Editor, IJREAM
Certificate No. 245491502200220
IJREAM
International Journal for Research in Engineering Application & Management
http://www.ijream.org


The Editorial Board of
INTERNATIONAL JOURNAL FOR RESEARCH IN ENGINEERING APPLICATION & MANAGEMENT
ISO 3297:2007 CERTIFIED JOURNAL | ISSN : 2454-9150
Certificate of Publication
This is to Certify that Paper ID : IJREAMV07I1284117
Image Captioning
Authored By
Atharva Ranade
has been published in Volume 07, Issue 12, Mar 2022.
The mentioned paper has gone through Peer Review Process & measured upto the required standard.
*Corresponding Author
DOI : 10.35291/2454-9150.2022.0100
Editor, IJREAM
Certificate No. 245491502200221
IJREAM
International Journal for Research in Engineering Application & Management
http://www.ijream.org


The Editorial Board of
INTERNATIONAL JOURNAL FOR RESEARCH IN ENGINEERING APPLICATION & MANAGEMENT
ISO 3297:2007 CERTIFIED JOURNAL | ISSN : 2454-9150
Certificate of Publication
This is to Certify that Paper ID : IJREAMV07I1284117
Image Captioning
Authored By
Siddharth Nair
has been published in Volume 07, Issue 12, Mar 2022.
The mentioned paper has gone through Peer Review Process & measured upto the required standard.
*Corresponding Author
DOI : 10.35291/2454-9150.2022.0100
Editor, IJREAM
7.427
Certificate No. 245491502200223
IJREAM
International Journal for Research in Engineering Application & Management
http://www.ijream.org
Crossref
Google scholar
THOMSON REUTERS


The Editorial Board of
INTERNATIONAL JOURNAL FOR RESEARCH IN ENGINEERING APPLICATION & MANAGEMENT
ISO 3297:2007 CERTIFIED JOURNAL | ISSN : 2454-9150
Certificate of Publication
This is to Certify that Paper ID : IJREAMV07I1284117
Image Captioning
Authored By
Omkar Thavai
has been published in Volume 07, Issue 12, Mar 2022.
The mentioned paper has gone through Peer Review Process & measured upto the required standard.
*Corresponding Author
DOI : 10.35291/2454-9150.2022.0100
Editor, IJREAM
7.427
Certificate No. 245491502200222
IJREAM
International Journal for Research in Engineering Application & Management
http://www.ijream.org
Crossref
Google scholar
THOMSON REUTERS

Chapter 10

LogBook

A. P. SHAH INSTITUTE OF TECHNOLOGY
(Approved by AICTE New Delhi & Govt. of Maharashtra, Affiliated to University of Mumbai)
(Religious Jain Minority)

Department Of Computer Engineering

Academic Year: 2021 to 2022

Year (SE / TE / BE): BE SEM : VIII

Project Title: IMAGE CAPTIONING

Team Member 1: (Team Leader)

Name: Shyam Krishna Menon

College ID: 18102014 Email: reachskmenon2017@gmail.com

Mobile No: 9819677210

Team Member 2:

Name: Atharva Ranade

College ID: 18102016 Email: ranadeatharva2112@gmail.com

Mobile No: 7045275441

Team Member 3:

Name: Siddharth Nair

College ID: 18102044 Email: nair.siddharth01@rediffmail.com

Mobile No: 8879714128

Team Member 4:

Name: Omkar Thavi

College ID: 18102061 Email: omkarthavi63@gmail.com

Mobile No: 9833436265

Project Guide **Project Coordinator** **Head of Department**

Guide Name: Sofiya Mujawar

Parivarnath Charitable Trust's

A. P. SHAH INSTITUTE OF TECHNOLOGY

(Approved by AICTE New Delhi & Govt. of Maharashtra, Affiliated to University of Mumbai)
(Religious Jain Minority)

Department Of Computer Engineering

Academic Year: 2021 to 2022

Year (SE / TE / BE): BE SEM : VIII WEEK No.: 1

Date: From 3/1/22 to 7/1/22

Progress Planned	Progress Achieved
Initiated research work and explored domains	Shortlisted Image Captioning, Image Processing domain
Referred to five papers published in IEEE journals	

Guides Review: I have checked the shortlisted research papers for Image captioning and suggested the group to proceed with the same.

Signature
Team Member 1:
Team Member 2:
Team Member 3:
Team Member 4:

Project guide:

Signature: 7/1/22

Date:

Parshvanath Charitable Trust's

A. P. SHAH INSTITUTE OF TECHNOLOGY

(Approved by AICTE New Delhi & Govt. of Maharashtra, Affiliated to University of Mumbai)
(Religious Jain Minority)

Department Of Computer Engineering

Academic Year: 2021 to 2022

Year (SE / TE / BE): BE SEM : VIII WEEK No.: 2

Date: From 7/2/22 to 11/2/22

Progress Planned	Progress Achieved
Finalized Image captioning topic and began research	Found several papers and selected five to study.
Learned different approaches of Image Captioning	Referred seperate papers focusing on different approaches

Guides Review: Had a discussion with the group about dataset to be chosen and selected the appropriate dataset.

Signature
Team Member 1:
Team Member 2:
Team Member 3:
Team Member 4:

Project guide:

Signature: 11-2-22

Date:

Parshvanath Charitable Trust's

A. P. SHAH INSTITUTE OF TECHNOLOGY

(Approved by AICTE New Delhi & Govt. of Maharashtra, Affiliated to University of Mumbai)
(Religious Jain Minority)

Department Of Computer Engineering

Academic Year: 2021 to 2022

Year (SE / TE / BE): BE SEM : VIII WEEK No.: 3

Date: From 21/2/22 to 25/2/22

Progress Planned	Progress Achieved
Began initial phase	Started working on finding datasets.
Simultaneously began working on front end	Started designing GUI for website
Initial coding was done on Jupyter notebook	
Looked for several datasets	Finalized Flickr30K dataset.

Guides Review: Discussed with the group about the complete work flow of the project and instructed them to start working on the frontend simultaneously.

Signature
Team Member 1:
Team Member 2:
Team Member 3:
Team Member 4:

Project guide:

Signature: 25-2-22

Date:

Parshvanath Charitable Trust's

A. P. SHAH INSTITUTE OF TECHNOLOGY

(Approved by AICTE New Delhi & Govt. of Maharashtra, Affiliated to University of Mumbai)
(Religious Jain Minority)

Department Of Computer Engineering

Academic Year: 2021 to 2022

Year (SE / TE / BE): BE SEM : VIII WEEK No.: 4

Date: From 7/3/22 to 11/3/22

Progress Planned	Progress Achieved
Started with preprocessing of our dataset.	Used preprocessing methods for captions & images
Started to research on various transfer learning methods	Executed VGG16 model

Guides Review: I had a meeting with the group and discussed about various algorithms for the image captioning model and instructed them to go with VGG16 model.

Signature
Team Member 1:
Team Member 2:
Team Member 3:
Team Member 4:

Project guide:

Signature: 11-3-22

Date:

Parshvanath Charitable Trust's

A. P. SHAH INSTITUTE OF TECHNOLOGY

(Approved by AICTE New Delhi & Govt. of Maharashtra, Affiliated to University of Mumbai)
(Religious Jain Minority)

Department Of Computer Engineering

Academic Year: 2021 to 2022

Year (SE / TE / BE) BE SEM VIII WEEK No: 5

Date: From 14/3/22 to 18/3/22

Progress Planned	Progress Achieved
Finalized algorithms for model building for image captioning	Used CNN - RNN Model along with LSTM
Integrated our ML model with Flask	Created a web application using Flask.

Guides Review: Checked their progress regarding the front end section. Had a discussion and shortlisted various journals for research paper publications.

Signature
Team Member 1:
Team Member 2:
Team Member 3:
Team Member 4:

Project guide:

Signature: 18-3-22

Date:

Parshvanath Charitable Trust's
A. P. SHAH INSTITUTE OF TECHNOLOGY
(Approved by AICTE New Delhi & Govt. of Maharashtra, Affiliated to University of Mumbai)
(Religious Jain Minority)

Department Of Computer Engineering

Academic Year: 20 21 to 20 22

Year (SE / TE / BE): BE SEM : VIII WEEK No.: 6

Date: From 4/4/22 to 8/4/22

Progress Planned	Progress Achieved
Tested our model on new images	Achieved decent Bleu Score on new images
Looked for various journals for publishing our paper	Submitted paper on IJREAM journal
Made some changes requested by the IJREAM panel	Published our research paper successfully for March edition

Guides Review: Checked execution of the complete working model for Image captioning and also had a meeting regarding changes to be made requested by the IJREAM journal panel and instructed them to publish the paper.

Signature
Team Member 1:
Team Member 2:
Team Member 3:
Team Member 4:

Project guide:

Signature:

Date: