VISVESVARAYA TECHNOLOGICAL UNIVERSITY, BELAGAVI

Internship Report on

"AUTOMATE IDENTIFICATION AND RECOGNITION OF HANDWRITTEN TEXT FROM AN IMAGE"

Submitted in partial fulfillment for the award of Degree of,

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE & ENGINEERING

By

JASLINE SHARON TAURO

4AL18CS029

Under the Supervision of

Mrs.Ankitha Shetty

Assistant Professor

ALVA'S
Education Foundation®

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

ALVA'S INSTITUTE OF ENGINEERING AND TECHNOLOGY

MOODBIDRI-574225, KARNATAKA

2021 – 2022

ALVA'S INSTITUTE OF ENGINEERING AND TECHNOLOGY

MIJAR, MOODBIDRI D.K. -574225

KARNATAKA

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

CERTIFICATE

This is to certify that the Internship report on "Automate Identification and Recognition of Handwritten texts from an image" submitted by JASLINE SHARON TAURO 4AL18CS029 is work done by him/her and is submitted during the academic year 2021 – 2022, in partial fulfilment of the requirements for the award of the degree of BACHELOR OF ENGINEERING in DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING of VISVESVARAYA TECHNOLOGICAL UNIVERSITY, BELAGAVI. It is certified that all corrections/suggestions indicated for Internal Assessment have been incorporated in the report deposited in the departmental library. The Internship report has been approved as it satisfies the academic requirements in respect of Internship work prescribed for the Bachelor of Engineering Degree.

Internship Guide
Department of CS&E

Internship Coordinator
Department of CS&E

Head of the Department
Department of CS&E

Examiners

Name of the Examiner **Signature with Date**

1.
2.

tcs iON

TATA

CERTIFICATE of INTERNSHIP

This is to certify that

Jasline Sharon Tauro

has successfully completed Remote Internship

for **210** ***hours*** *in project titled*

Automate Identification and Recognition of Handwritten Text from an Image

by TCS iON from 13 Apr 2021 *to* 05 Jul 2021.

TCS iON REMOTE INTERNSHIPS
Academic Credits with Industry Mentors
Cert. ID.: 358-11467763-1016
Dated: 05 Jul 2021

CERTIFIED

Mehul Mehta

Mehul Mehta
Global Delivery Head - TCS iON,
Tata Consultancy Services

ACKNOWLEDGEMENT

First, I would like to thank **Mr. MEHUL MEHTA, Global Delivery Head TCS iON** for giving me the opportunity to do an internship within the organization.

I would also like to thank **Mr. Debashis Roy** for mentoring me in **TCS iON** with their patience and openness they created an enjoyable working environment.

I am highly indebted to Managing Trustee **Mr. Vivek Alva** and **Principal Dr. Peter Fernandes, Alva's Institute of Engineering and Technology, Mijar** for the facilities provided to accomplish this internship.

I thank my Head of the Department **Dr. Manjunath Kotari,** ***Professor, Departmentof Computer Science and Engineering*** for his constructive criticism throughout my internship.

I am thankful to Internship coordinator, **Mrs. Merlyn Melita Mathias**, ***Assistant Professor, Department of Computer Science and Engineering*** for her valuable guidance, support and advices to get and complete internship in above said organization.

I would like to thank my internship guide **Mrs. Ankitha Shetty** ***Assistant Professor Department of Computer Science and Engineering*** for his constant help and support throughout the internship.

I am extremely grateful to my department staff members and friends who helped mein successful completion of this internship.

JASLINE SHARON TAURO

4AL18CS029

COMPANY PROFILE

Tata Consultancy Services is an Indian multinational information technology services and consulting company.TCS iON is a business unit of Tata Consultancy Services(TCS). It is focused on empowering people with tech-led education to transform themselves for the new world.TCS iON offers a plethora of multiple learning courses for individual learners imparting a varied range of skills and proficiencies. TCS iON provides technology by means of a unique IT-as-a-Service model, offering end-to-end business solutions. It caters to the needs of multiple industry segments, through innovative, easy-to-use, secured, integrated, hosted solutions in a build-as-you-grow, pay-as-you-use business model. TCS iON serves its clients with the help of best practices gained through TCS' global experience, domestic market reach, skills, and delivery capabilities. TCS iON is highly modular, scalable and configurable giving businesses and educational institutions the benefits of increased efficiencies, faster go to market, predictability of technology as well as spend and better business results.Actually ion is the subsidiary of T.C.S, which is cloud IT platform. T.C.S work of making and maintain software and websites.The TCS iON this is the cloud platform for conducting online assessment and publishing there results.The TCS iON is S.M.B project work. Whose aim is likely digital India. The company want to conduct all online.TCS iON serves its clients with the help of best practices gained through TCS' global experience, domestic market reach, skills, and delivery capabilities. TCS iON's Cloud Based Solution is highly modular, scalable and configurable giving businesses and educational institutions the benefits of increased efficiencies, faster go to market, predictability of technology as well as spend and better business results

ABSTRACT

Handwritten text acknowledgment is yet an open examination issue in the area of Optical Character Recognition (OCR). This paper proposes a productive methodology towards the advancement of handwritten text acknowledgment frameworks. The primary goal of this task is to create AI calculation to empower element and information extraction from records with manually written explanations, with an, expect to distinguish transcribed words on a picture. The main aim of this project is to extract text, this text can be handwritten text or it can machine printed text and convert it into computer understandable or we can say computer editable format. To implement this project we have used pytesseract which is an open-source OCR engine used to recognize handwritten text and opencv a library in python used to solve computer vision problems. So the input image is executed in various steps, first there is preprocessing of an image then there is text localization after that there is character segmentation and character recognition and finally we have post-processing of image. Further image processing algorithms can also be used to deal with the multiple characters input in a single image, tilt image, or rotated image. The prepared framework gives a normal precision of more than 95% with the concealed test picture.

TABLE OF CONTENTS

LIST OF FIGURES

ALVA'S INSTITUTE OF ENGINEERING & TECHNOLOGY
A Unit Of Alva's Education Foundation (R), Moodbidri
(Affiliated to VTU, Belgaum, Approved by AICTE, New Delhi, Recognized by Govt. Of Karnataka)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

1ST WEEK

DAY	DATE	NAME OF THE TOPIC/MODULE	COMPLETED
Tuesday	13-04-2021	Joining day of TCS ion completed the test given by the platform Basic introduction to AI and ML	Yes
Wednesday	14-04-2021	Introduction to pandas real world examples of AI	Yes
Thursday	15-04-2021	Introduction to pandas	Yes
Friday	16-04-2021	Enhanced image retrieval algorithm	Yes
Saturday	17-04-2021	Introduction to numpy	Yes

2nd WEEK

DAY	DATE	NAME OF THE TOPIC/MODULE	COMPLETED
Monday	19-04-2021	Collected some data set from IAM	Yes
Tuesday	20-04-2021	Introduction to CNN model	Yes
Wednesday	21-04-2021	Glimpse of CTC algorithm	Yes
Thursday	22-04-2021	Study / analysis on different image and its classification	Yes
Friday	23-04-2021	A application in industry and finance	Yes
Saturday	24-04-2021	Machine learning using scikit and Python	Yes

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

3rd WEEK

DAY	DATE	NAME OF THE TOPIC/MODULE	COMPLETED
Monday	26-04-2021	Identify the steps in the data science workflow.	Yes
Tuesday	27-04-2021	Differences between training and inference	Yes
Wednesday	28-04-2021	Glimpse about K nearest neighbor classifier	Yes
Thursday	29-04-2021	Introduction to machine learning technology	Yes
Friday	30-04-2021	Data representation and visualization of data	Yes
Saturday	01-05-2021	Steps in data science workflow	Yes

4th WEEK

DAY	DATE	NAME OF THE TOPIC/MODULE	COMPLETED
Monday	03-05-2021	New best classifier with Scikit	Yes
Tuesday	04-05-2021	Introduction to matplotlib	Yes
Wednesday	05-05-2021	Introduction to neural networks initialization the structure	Yes
Thursday	06-05-2021	Running neural network using Python	Yes
Friday	07-05-2021	Understanding the data analytics with Python technique data Munching	Yes
Saturday	08-05-2021	Understanding the use of numpy and implementation	Yes

ALVA'S INSTITUTE OF ENGINEERING & TECHNOLOGY

A Unit Of Alva's Education Foundation (R), Moodbidri

(Affiliated to VTU, Belgaum, Approved by AICTE, New Delhi, Recognized by Govt. Of Karnataka)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

5th WEEK

DAY	DATE	NAME OF THE TOPIC/MODULE	COMPLETED
Monday	10-05-2021	Basics of statistical learning for scientific data processing	Yes
Tuesday	11-05-2021	Supervised learning predicting an output variable from high dimension observations	Yes
Wednesday	12-05-2021	Model selection	Yes
Thursday	13-05-2021	Import libraries creating datasets	Yes
Friday	14-05-2021	Training of a model given	Yes
Saturday	15-05-2021	Recognition of sample datas	Yes

6th WEEK

DAY	DATE	NAME OF THE TOPIC/MODULE	COMPLETED
Monday	17-05-2021	Some concept of convolutional neural network	Yes
Tuesday	18-05-2021	Get transfer learning of fine tuning	Yes
Wednesday	19-05-2021	Data augmentation	Yes
Thursday	20-05-2021	Image analysis using tensorflow	Yes
Friday	21-05-2021	Checking for accuracy of the model	Yes
Saturday	22-05-2021	Then answer the message retrieval algorithm for detecting the text	Yes

ALVA'S INSTITUTE OF ENGINEERING & TECHNOLOGY

A Unit Of Alva's Education Foundation (R), Moodbidri

(Affiliated to VTU, Belgaum, Approved by AICTE, New Delhi, Recognized by Govt. Of Karnataka)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

7th WEEK

DAY	DATE	NAME OF THE TOPIC/MODULE	COMPLETED
Monday	24-05-2021	Working on project	Yes
Tuesday	25-05-2021	Working on project	Yes
Wednesday	26-05-2021	Working on project	Yes
Thursday	27-05-2021	Working on project	Yes
Friday	28-05-2021	Working on project	Yes
Saturday	29-05-2021	Working on project	Yes

8th WEEK

DAY	DATE	NAME OF THE TOPIC/MODULE	COMPLETED
Monday	31-05-2021	Working on project	Yes
Tuesday	01-06-2021	Working on project	Yes
Wednesday	02-06-2021	Working on project	Yes
Thursday	03-06-2021	Working on project	Yes
Friday	04-06-2021	Working on project	Yes
Saturday	05-06-2021	Working on project	Yes

9th WEEK

DAY	DATE	NAME OF THE TOPIC/MODULE	COMPLETED
Monday	07-06-2021	Working on project	Yes
Tuesday	08-06-2021	Working on project	Yes
Wednesday	09-06-2021	Working on project	Yes
Thursday	10-06-2021	Working on project	Yes
Friday	11-06-2021	Final report submission	Yes
Saturday	12-06-2021	Viva report submission	Yes

CHAPTER 1

INTRODUCTION

As we know nowadays there are different websites that are developed by the companies for various purposes, which require testing before deploying them to the market. This project is basically focused on identification and recognition of handwritten text from an image. It is based on enhancement of optical character recognition system. Optical character recognition or optical character reader is the electronic or mechanical conversion of images of typed, handwritten or printed text into machine-encoded text, whether from a scanned document, a photo of a document, a scene-photo or from subtitle text superimposed on an image. An optical character recognition problem is basically a type of image-based sequence recognition problem.

Extracting text of various sizes, shapes, and orientations from images is an essential problem in many contexts, especially in e-commerce, augmented reality assistance systems, and content moderation in social media platforms. To tackle this problem, one needs to accurately extract the text from images.

Basically, text extraction can be achieved into two steps, i.e., text detection and text recognition or by training a single model to achieve both text detection and recognition. Text detection helps identify the region in the image where the text is present. It takes in an image as an input, and the outputs bounding boxes.Text recognition extracts the text from the input image using the bounding boxes obtained from the text detection model. It takes in an image and some bounding boxes as inputs and outputs some raw text.

Text detection is very similar to the object detection task where the object which needs to be detected is nothing but the text. Much research has taken place in this field to detect text out of images accurately, and many of these detectors detect text at the word level. Few of the examples are:

Text Boxes

Text Boxes ++

EAST etc.

CHAPTER 2

SYSTEM ANALYSIS

System analysis is conducted for the purpose of studying a system or its parts in order to identify its objectives. It is a problem-solving technique that improves the system and ensures that all the components of the system work efficiently to accomplish their purpose.

2.1 Requirement Specification:

To be used efficiently, all computer software needs certain hardware components or the other software resources to be present on a computer. These prerequisites are known as (computer) system requirements and are often used as a guideline as opposed to an absolute rule. Most software defines two sets of system requirements: minimum and recommended. With increasing demand for higher processing power and resources in newer versions of software, system requirements tend to increase over time. Industry analysts suggest that this trend plays a bigger part in driving upgrades to existing computer systems than technological advancements.

2.1.1 Hardware Requirements:

Hardware components deal with the basic hardware requirement to develop and to run any system. It includes processor, memory etc.

Hardware Constraints: PROCESSOR: 1.4GHz Intel core i5

RAM: 4GB

2.1.2 Software Requirements:

Software Requirements deal with defining software resource requirements and pre-requisites that need to be installed on a computer to provide optimal functioning of an application. These requirements or pre-requisites are generally not included in the software installation package and need to be installed separately before the software is installed.

Software Constraints:

OPERATING SYSTEM: Linux or Windows

LANGUAGE: python

2.2 Language Used for Implementation:

Python is a high-level, interpreted, general-purpose programming language. Its design philosophy emphasizes code readability with the use of significant indentation. Python is dynamically-typed and garbage-collected. It supports multiple programming paradigms, including structured (particularly procedural), object-oriented and functional programming. It is often described as a "batteries included" language due to its comprehensive standard library. Python was conceived in the late 1980s by Guido van Rossum at Centrum Wiskunde & Informatica (CWI) in the Netherlands as a successor to the ABC programming language, which was inspired by SETL, capable of exception handling and interfacing with the Amoeba operating system.

2.3 Platform Used for Implementation:

- Google colab.

CHAPTER 3

OBJECTIVE ACHIEVEMENT

To develop machine learning algorithms in order to enable entity and knowledge extraction from documents with handwritten annotations, with an aim to first identify handwritten words on an image and then recognize the characters to transcribe the text. An optical character recognition problem is basically a type of image-based sequence recognition problem. And for sequence recognition problem, most suited neural networks are recurrent neural networks (RNN) while for an image-based problem most suited are convolution neural networks (CNN). To cop up with the OCR problems we need to combine both of these CNN and RNN. So, I used Convolutional Recurrent Neural Network (CRNN) to tackle the both the problems.

3.1 Skills Learnt:

3.1.1 Python

Python is a high-level, interpreted, general-purpose programming language. Its design philosophy emphasizes code readability with the use of significant indentation. Python is dynamically-typed and garbage-collected. It supports multiple programming paradigms, including structured (particularly procedural), object-oriented and functional programming. It is often described as a "batteries included" language due to its comprehensive standard library. Python was conceived in the late 1980s by Guido van Rossum at Centrum Wiskunde & Informatica (CWI) in the Netherlands as a successor to the ABC programming language, which was inspired by SETL, capable of exception handling and interfacing with the Amoeba operating system. Its implementation began in December 1989. Van Rossum shouldered sole responsibility for the project, as the lead developer, until 12 July 2018, when he announced his "permanent vacation" from his responsibilities as Python's "benevolent dictator for life", a title the Python community bestowed upon him to reflect his long-term commitment as the project's chief decision-maker. In January 2019, active Python core developers elected a five-member Steering Council to lead the project.

3.1.2 Open CV

OpenCV is the huge open-source library for the computer vision, machine learning, and image processing and now it plays a major role in real-time operation which is very important in today's systems. By using it, one can process images and videos to identify objects, faces, or even handwriting of a human. When it integrated with various libraries, such as NumPy, python is capable of processing the OpenCV array structure for analysis. To identify image pattern and its various features we use vector space and perform mathematical operations on these features. The first OpenCV version was 1.0. OpenCV is released under a BSD license and hence it's free for both academic and commercial use. It has C++, C, Python and Java interfaces and supports Windows, Linux, Mac OS, iOS and Android. When OpenCV was designed the main focus was real-time applications for computational efficiency. All things are written in optimized C/C++ to take advantage of multi-core processing.

3.1.3 TensorFlow

TensorFlow is a free and open-source software library for machine learning and artificial intelligence. It can be used across a range of tasks but has a particular focus on training and inference of deep neural networks. TensorFlow was developed by the Google Brain team for internal Google use in research and production. The initial version was released under the Apache License 2.0 in 2015. Google released the updated version of TensorFlow, named TensorFlow 2.0, in September 2019. TensorFlow can be used in a wide variety of programming languages, most notably Python, as well as JavaScript, C++, and Java.

3.1.4 Google colab

Google is quite aggressive in AI research. Over many years, Google developed AI framework called TensorFlow and a development tool called Collaboratory. Today TensorFlow is open sourced and since 2017, Google made Collaboratory free for public use. Collaboratory is now known as Google Colab or simply Colab. Another attractive feature that Google offers to the developers is the use of GPU. Colab supports GPU and it is totally free.

3.2 Step-by-Step guide to creating Google-Colab and Deploying TensorFlow

3.2.1Creating a document in a google colab

Google colab Chrome Extension If you want to create a machine learning model but say you don't have a computer that can take the workload, Google Colab is the platform for you. Even if you have a GPU or a good computer creating a local environment with anaconda and installing packages and resolving installation issues are a hassle. Collaborator is a free colab notebook environment provided by Google where you can use free GPUs and TPUs which can solve all these issues.

Opening Google Colab Notebook On opening the website you will see a pop-up containing following tabs

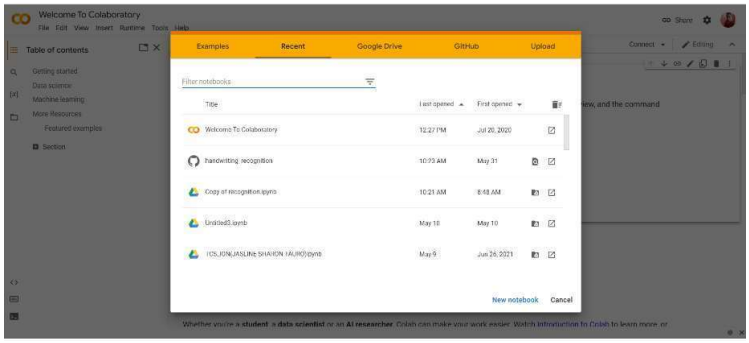


Fig 3.1: Google Colab Notebook

- EXAMPLES: Contain a number of Colab notebooks of various examples.
- RECENT: Colab notebook you have recently worked with.
- GOOGLE DRIVE: Colab notebook in your google drive
- . • GITHUB: You can add Colab notebook from your GitHub but you first need to connect Colab with GitHub
- . • UPLOAD: Upload from your local directory.

Else you can create a new Colab notebook by clicking New Python3 Notebook or New Python2 Notebook at the bottom right corner. On creating a new notebook, it will create a Colab notebook with Untitled0.ipynb and save it to your google drive in a folder named Colab Notebooks

CHAPTER 4

IMPLEMENTION

4.1 Tools and Technologies

TensorFlow

TensorFlow was developed by the Google Brain team for internal Google use in research and production. The initial version was released under the Apache License 2.0 in 2015. Google released the updated version of TensorFlow, named TensorFlow 2.0, in September 2019. TensorFlow can be used in a wide variety of programming languages, most notably Python, as well as Javascript, C++, and Java.

Python

Python is a high-level, interpreted, general-purpose programming language. Its design philosophy emphasizes code readability with the use of significant indentation. Python is dynamically-typed and garbage-collected. It supports multiple programming paradigms, including structured (particularly procedural), object-oriented and functional programming. It is often described as a "batteries included" language due to its comprehensive standard library.

Jupiter Notebook

Project Jupyter is a project and community whose goal is to "develop open-source software, open-standards, and services for interactive computing across dozens of programming languages". It was spun off from iPython in 2014 by Fernando Pérez and Brian Granger.

Google colab

Google is quite aggressive in AI research. Over many years, Google developed AI framework called TensorFlow and a development tool called Colaboratory. Today TensorFlow is open-sourced and since 2017, Google made Colaboratory free for public use. Colaboratory is now known as Google Colab or simply Colab

Synthetic Word Dataset

The exact data used to train our deep convolutional neural networks is available below. This is synthetically generated dataset which we found sufficient for training text recognition on real-world images. This dataset consists of 9 million images covering 90k English words, and includes the training, validation and test splits used in our work.

4.2 Instructions

1. Directory Structure:

Upload the Synthetic Word Dataset to the google drive.

2. Uploading the Dataset on Google Drive and accessing it.

After uploading the dataset accessing the dataset after unzipping.

Fig 4.2.1: Uploading the Dataset

3. Preprocessing the Data.

After fetching the dataset, we will preprocess the data.

4. Dividing the dataset into train, test and validation sets.

After splitting we have 760,895 train images, 256 test and validation images each.

Fig 4.2.2: Testing Dataset sets.

5.Creating the defining the model/network architecture.

Adding the Pooling Layers, CNN Layers, Activation Functions, etc.

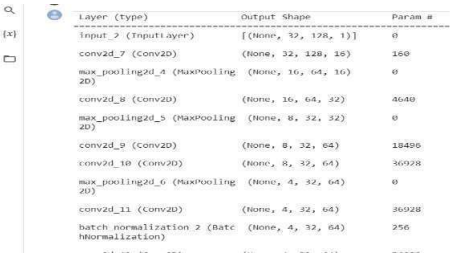
6.Training the model.

Using 30 epochs to train samples.

Fig 4.2.3: Training Dataset sets

7.Saving the model.

Saving the model using HDF5 file extension.



Layer (type)	Output Shape	Param #
input_2 (InputLayer)	[(None, 32, 128, 1)]	0
conv2d_7 (Conv2D)	(None, 32, 128, 16)	160
max_pooling2d_4 (MaxPooling 2D)	(None, 16, 64, 16)	0
conv2d_8 (Conv2D)	(None, 16, 64, 32)	4640
max_pooling2d_5 (MaxPooling 2D)	(None, 8, 32, 32)	0
conv2d_9 (Conv2D)	(None, 8, 32, 64)	18496
conv2d_10 (Conv2D)	(None, 8, 32, 64)	36928
max_pooling2d_6 (MaxPooling 2D)	(None, 4, 32, 64)	0
conv2d_11 (Conv2D)	(None, 4, 32, 64)	36928
batch_normalization_2 (Batc hNormalization)	(None, 4, 32, 64)	256

Fig 4.2.4: Saving model

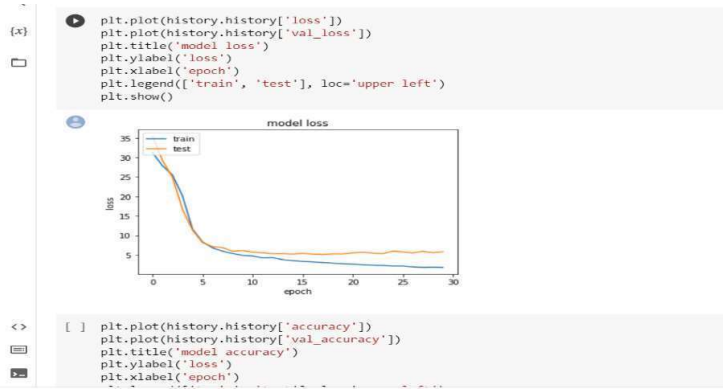
8. Testing the model

```
labels = Input(name='the_labels', shape=[max_label_len], dtype='float32')
input_length = Input(name='input_length', shape=[1], dtype='int64')
label_length = Input(name='label_length', shape=[1], dtype='int64')

def ctc_lambda_func(args):
    y_pred, labels, input_length, label_length = args
    return K.ctc_batch_cost(labels, y_pred, input_length, label_length)

loss_out = Lambda(ctc_lambda_func,
                  output_shape=(1,),
                  name='ctc')([outputs, labels, input_length, label_length])

#model to be used at training time
model = Model(inputs=[inputs, labels, input_length, label_length], outputs=loss_out)
```

```
file_path = "C_LSTM_best.hdf5"

model.compile(loss={'ctc': lambda y_true, y_pred: y_pred}, optimizer = 'adam', metrics=['accuracy'])


checkpoint = ModelCheckpoint(filepath=file_path,
                             monitor='val_loss',
                             verbose=1,
                             save_best_only=True,
                             mode='min')

callbacks_list = [checkpoint]
```

Fig 4.2.5: Testing model

9.Plotting the loss and accuracy plots.



```
plt.plot(history.history['loss'])
plt.plot(history.history['val_loss'])
plt.title('model loss')
plt.ylabel('loss')
plt.xlabel('epoch')
plt.legend(['train', 'test'], loc='upper left')
plt.show()
```



```
plt.plot(history.history['accuracy'])
plt.plot(history.history['val_accuracy'])
plt.title('model accuracy')
plt.ylabel('loss')
plt.xlabel('epoch')
```

Fig 4.2.6: Plotting of loss plot.

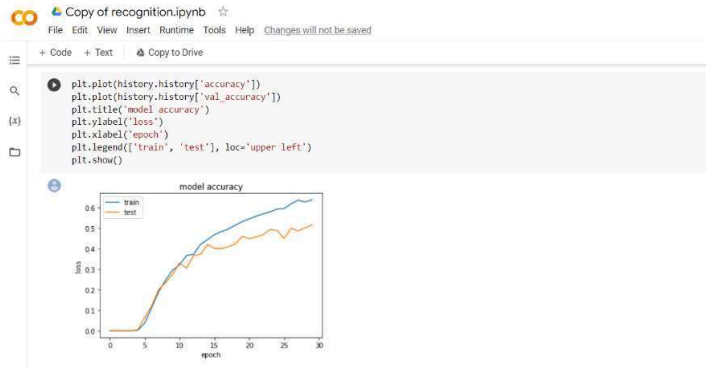


Fig 4.2.7: Plotting of accuracy plot

CHAPTER 5

CONCLUSION

In this project classification of characters takes place. The project is achieved through the conventional neural network. The accuracy we obtained in this is above 78.3%. This algorithm will provide both the efficiency and effective result for the recognition. The project gives best accuracy for the text which has less noise. The accuracy completely depending on the dataset if we increase the data, we can get more accuracy. If we try to avoid cursive writing then also it gives best results. Future Work: In future we are planning to extend this study to a larger extent where different embedding models can be considered on large variety of the datasets. The future is completely based on technology no one will use the paper and pen for writing. In that scenario they used write on touch pads so the inbuilt software which can automatically detects text which they writing and convert into digital text so that the searching and understanding very much simplified.

REFERENCES

1. https://www.leewayhertz.com/ethereum-smart-contract-tutorial/
2. https://www.intel.com/content/www/us/en/developer/learn/courseartificialintelligence.html
3. https://numpy.org/doc/
4. https://www.tensorflow.org/tutorials/images/classification