Optical Character Recognition

A documentation Submitted
in Complete Fulfillment of
the Requirements for the
Course of

Summer Internship in Celebal Technologies

In Third year – Sixth Semester,
B-Tech in Computer Science
Engineering

Under

Mr. Bikram Pratim Bhuyan (Faculty Mentor (UPES))

Mr. Anurag Sharma (Team Lead (Celebal))

By

500067783	**R177218086**	**Shreya Joshi**
500067500	**R177218089**	**Shubham Sanghi**
500068029	**R177218096**	**Suvansh Kapur**
500069768	**R133218012**	**M. Sweety Reddy**

DEPARTMENT OF INFORMATICS
SCHOOL OF COMPUTER SCIENCE

UNIVERSITY OF PETROLEUM AND ENERGY STUDIES, BIDHOLI, DEHRADUN, UTTRAKHAND, INDIA

7th June 2021- 22nd July 2021

Table of Contents

-

Acknowledgement

We would like to express our special thanks of gratitude to our UPES faculty mentor Mr. Bikram Pratim Bhuyan and Celebal mentor Mr. Anurag Sharma for guiding us during the entire course of our internship and for providing the necessary information regarding the project.

We would also like to thank the supervisors from Celebal Mr. Sharthak Acharjee and Ms. Anjali Arora who time to time helped in solving our doubts and queries through virtual meetings and gave their full efforts in guiding the team in achieving the goal as well as provided encouragement to maintain our progress in track.

Last but not least we extend our gratitude to the Company Celebal Technologies for giving us this opportunity.

Company Overview

Company Name : Celebal Technologies

Celebal Technologies is a premier software services company in the field of Data Science, Big Data and Enterprise Cloud. Celebal Technologies helps you to discover the competitive advantage by employing intelligent data solutions using cutting-edge technology solutions that can bring massive value to your organization. The core offerings are around "Data to Intelligence", wherein company leverage data to extract intelligence and patterns thereby facilitating smarter and quicker decision making for clients.

Celebal Technologies solutions are powered by Robotics, Artificial Intelligence and Machine Learning algorithms which offer improved business efficiency in the interconnected world. Company’s clients are Product ISVs and delivery organizations across the globe who use company’s niche expertise in product development as well as Enterprise project implementations. Company have adopted the highest standards of service quality and operational excellence, enabling its clients across a wide range of industries to transform into a data-driven enterprise. Company’s tailor-made solutions help enterprises maximize productivity, improve speed and accuracy.

With Celebal Technologies, who understands the core value of modern analytics over the enterprise, we help the business in improving business intelligence and more data-driven in architecting solutions. Company’s analytics team caters to ad-hoc and define business analytics.

Project Title : Optical Character Recognition

Internship Duration : 45 days

Project Scope

In this project we had to
predict the next character from the previous
set of words or sentences. For that we required a dataset
which contains images of printed sentences.
Some uses of optical character Recognition:-

- **Searchability**: Once your scanned file has been converted to machine-readable text, you can save it in a formatsuch as .doc, .rtf, .txt (simplest), .pdf etc. These files are internally searchable using Ctrl+F in Windows (Command+F in Mac). By uploading them to a suitable database, like Google Drive (for your private use) or Archive.org (accessible to anyone), you make these documents available globally.
- **Editability** : You might want to make amends to an old term paper you wrote, or revise an old will. Once your document is digitised with OCR, this is easily done using a word processor, rather than have to type the whole document again.
- **Accessibility** : Once a document is scanned by OCR and made available on a common database, it is available to anyone with access to that database. This is particularly useful for banks for example, which can access a customer’s past cheques anytime, anywhere to investigate their credit history. Another immediately obvious application is in making government archives available, so you can find your land ownership record or your grandfather’s birth certificate from anywhere.

Limitation : If the original document is of poor quality or the handwriting is difficult to read then more mistakes will occur.

Project Description

Optical Character Recognition

1. **PURPOSE:** The principle behind Optical Character Recognition (OCR) framework dependent on a lattice foundation is to perform Document Image Analysis, report preparing of electronic record designs changed over from paper arranges all the more viably and productively. This works on the exactness of perceiving the characters during record preparing contrasted with different existing accessible person acknowledgment techniques. Here OCR procedure determines the importance of the characters, their textual style properties from their spot planned pictures. The essential goal is to accelerate the cycle of character acknowledgment in report preparing. The characters, words, sentences and anything which an image contains, OCR helps to recognize them and can even cut and copy the words that image contains in it. The main scope of our project Optical Character Recognition is to give a productive and improved programming apparatus for the clients to perform Document Image Analysis, archive handling by perusing and perceiving the characters in research, scholarly, administrative and business associations that are having enormous pool of reported, checked pictures. The goal of my project is to create a reliable OCR Predicting the next character from the previous set of sentences and words.

2. **PseudoCode/ Algorithm (0CR)**

def OCRModel():

(1) //taking image as input and defining 1st convolutional layer

```
image=keras.layers.Input((32,784,1)
conv1=keras.layers.Conv2D(16,
(3,3),activation='relu',padding='same')(image)
```

(2) //applying Max pooling and defining 2nd convolutional layer

```
mp1=keras.layers.MaxPooling2D((2,2),padding='same')(conv1)
conv2=keras.layers.Conv2D(32,
(3,3),activation='relu',padding='same')(mp1)
```

(3) //applying Max pooling and defining 3rd convolutional layer

```
mp2=keras.layers.MaxPooling2D((2,2),padding='same')(conv2)
conv3=keras.layers.Conv2D(64,
(3,3),activation='relu',padding='same')(mp2)
```

(4) //applying Max pooling and defining 4th convolutional layer

```
mp3=keras.layers.MaxPooling2D((2,2),padding='same')(conv3)
conv4=keras.layers.Conv2D(128,
(3,3),activation='relu',padding='same')(mp3)
```

(5) //applying Max pooling and defining 5th convolutional layer

```
mp4=keras.layers.MaxPooling2D((2,1),padding='same')(conv4)
conv5=keras.layers.Conv2D(256,
(3,3),activation='relu',padding='same')(mp4)
```

(6) //applying Max pooling and defining 6th convolutional layer

```
mp5=keras.layers.MaxPooling2D((2,1),padding='same')(conv5)
conv6=keras.layers.Conv2D(256,
(3,3),activation='relu',padding='same')(mp5)
```

(7) //applying Batch Normalization

```
bn=keras.layers.BatchNormalization()(conv6)

sq=keras.backend.squeeze(bn,axis=1)
```

(8) //Using LSTM model

```
rn1=keras.layers.Bidirectional(keras.layers.LSTM(256,return_sequences=True))(sq)
rn2=keras.layers.Bidirectional(keras.layers.LSTM(256,return_sequences=True))(rn1)
```

(9) // using Softmax Activation Function

```
exd=keras.backend.expand_dims(rn2,axis=2)
maping=keras.layers.Conv2D(len(symbols),(2,2),activation='relu',padding='same')(exd)
maping=keras.backend.squeeze(maping,axis=2)

maping = tf.keras.layers.Softmax()(maping)
```

(10) // using Adam optimizer to compute loss

```
model=keras.Model(image,maping)

model.compile(loss='categorical_crossentropy', optimizer='adam')
```

(11) return model

3. Methodology:

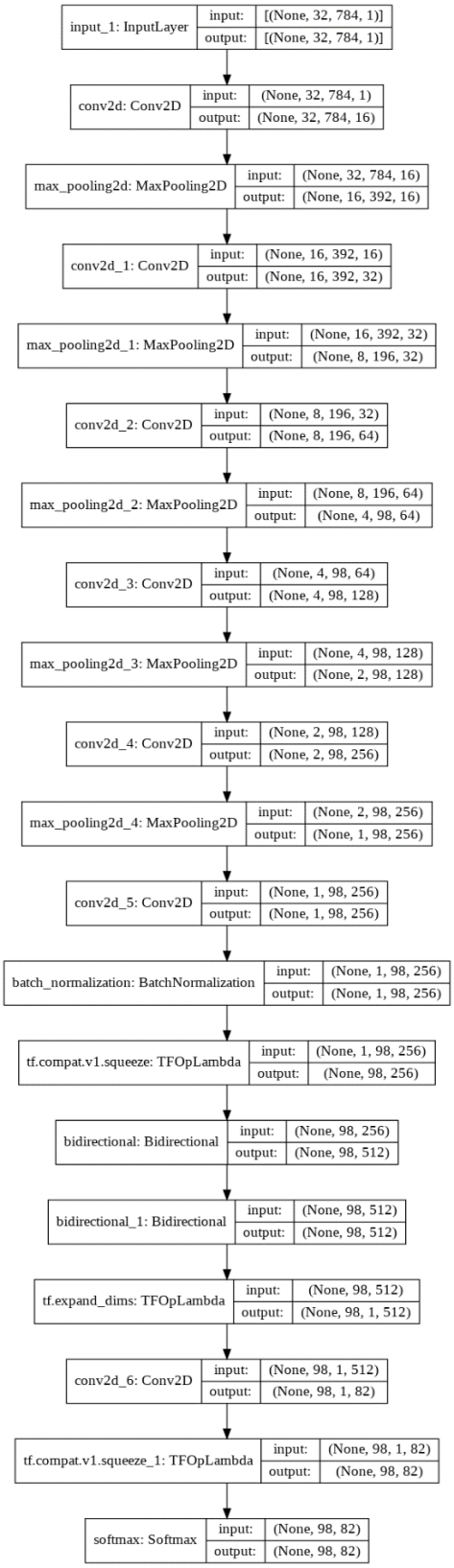
1. Training: Training One part of the program trains the network to recognize the characters. This network takes input-output vector pairs during training. The network trains its weight array to minimize the selected performance measure, i.e., error using back propagation algorithm.

2. Preprocessing: The raw data is subjected to a number of preliminary processing steps to make it usable in the descriptive stages of character analysis. Pre-processing aims to produce data that are easy for the OCR systems to operate accurately. The various tasks performed on the image in pre-processing stage includes batch normalization (also known as **batch norm**) is a method used to make artificial neural networks faster and more stable through normalization of the layers' inputs by re-centering and re-scaling. Batch normalization is applied after every convolutional block to speed up the training.

3. Segmentation: During the recognition phase, when a word needs to be recognized, it is divided into separate characters by removing its headline by horizontal scanning. This is done by inverting the image and replacing the rows with all white pixels into black ones. After doing this, segmentation is performed in the same way as of training using the same functions. This can be easily accomplished due to bounding box as number of components in any image does not affect it. Each individual character is uniformly resized into 5x7 pixels for extracting its features.

4. Feature Extraction: This step is the core of the framework. It characterizes each character by the presence or nonappearance of key highlights and these key highlights may height, width, thickness, circles, lines, stems and other characteristics. Each character is passed judgement on dependent on these attributes and fundamental highlights. The information is limited by saving just the required data. This will give us a vector with scalar qualities. Highlights are removed similarly as in the preparation part and the highlights of the character to be perceived is stored in independent vector.

5. Model flowchart:


input_1: InputLayer | input: [(None, 32, 784, 1)] | output: [(None, 32, 784, 1)]
conv2d: Conv2D | input: (None, 32, 784, 1) | output: (None, 32, 784, 16)
max_pooling2d: MaxPooling2D | input: (None, 32, 784, 16) | output: (None, 16, 392, 16)
conv2d_1: Conv2D | input: (None, 16, 392, 16) | output: (None, 16, 392, 32)
max_pooling2d_1: MaxPooling2D | input: (None, 16, 392, 32) | output: (None, 8, 196, 32)
conv2d_2: Conv2D | input: (None, 8, 196, 32) | output: (None, 8, 196, 64)
max_pooling2d_2: MaxPooling2D | input: (None, 8, 196, 64) | output: (None, 4, 98, 64)
conv2d_3: Conv2D | input: (None, 4, 98, 64) | output: (None, 4, 98, 128)
max_pooling2d_3: MaxPooling2D | input: (None, 4, 98, 128) | output: (None, 2, 98, 128)
conv2d_4: Conv2D | input: (None, 2, 98, 128) | output: (None, 2, 98, 256)
max_pooling2d_4: MaxPooling2D | input: (None, 2, 98, 256) | output: (None, 1, 98, 256)
conv2d_5: Conv2D | input: (None, 1, 98, 256) | output: (None, 1, 98, 256)
batch_normalization: BatchNormalization | input: (None, 1, 98, 256) | output: (None, 1, 98, 256)
tf.compat.v1.squeeze: TFOpLambda | input: (None, 1, 98, 256) | output: (None, 98, 256)
bidirectional: Bidirectional | input: (None, 98, 256) | output: (None, 98, 512)
bidirectional_1: Bidirectional | input: (None, 98, 512) | output: (None, 98, 512)
tf.expand_dims: TFOpLambda | input: (None, 98, 512) | output: (None, 98, 1, 512)
conv2d_6: Conv2D | input: (None, 98, 1, 512) | output: (None, 98, 1, 82)
tf.compat.v1.squeeze_1: TFOpLambda | input: (None, 98, 1, 82) | output: (None, 98, 82)
softmax: Softmax | input: (None, 98, 82) | output: (None, 98, 82)

Literature Review

On reviewing some of the papers published by IEEE we were able to conclude:

- Given the universality of manually written archives in human exchanges, Optical Character Recognition (OCR) of records have significant commonsense worth.
- Optical person acknowledgment is a science that empowers to decipher different sorts of archives or pictures into analyzable, editable and accessible information.
- During last decade, analysts have utilized computerized reasoning/AI devices to consequently examine written by hand and printed reports to change over them into electronic organization.
- The goal of this survey paper is to sum up research that has been led on character acknowledgment of transcribed archives and to give research headings.
- In this Systematic Literature Review (SLR) we gathered, blended and examined research articles on the subject of transcribed OCR (and firmly related points) which were distributed between year 2000 to 2019.
- We followed generally utilized electronic information bases by following pre-characterized survey convention. Articles were looked through utilizing catchphrases, forward reference looking and in reverse reference looking to look through every one of the articles identified with the subject.
- After cautiously following investigation determination measure 176 articles were chosen for this SLR. This survey article effectively presents best in class results and procedures on OCR

and furthermore give research headings by featuring research holes.