CSCE 5290

Natural Language Processing

Project Proposal

Github link : https://github.com/nehareddy9399/NLP-Projectgroup11

1.Project Title and Members:

Project Title: Optical Character Recognition using Deep Learning
Project Team number : 11

Team Members:
Keerthana Pinikeshi - 11448714
Neha Kalluri - 11445206
Lahari Kunduru - 11445208

2.Goals and Objectives:

Motivation:

As part of this NLP project, we thought of building something that is very close to the real world and widely used in NLP applications. We found that Optical Character Recognition (OCR) is one of the most popular area of research in NLP applications.

We as a team felt that Optical Character Recognition problem statement will be used by all kinds individuals in their day to day life. Hence we are proposing to work on Optical Character Recognition model to recognize the text in the given input image.

Significance :

The major significance of Optical Character Recognition (OCR) is used in almost all kinds of industries and individuals. Before the advancements in the OCR, we used to look at the hard copies of documents and type the content manually. But now we can process any document or image to extract the text from the OCR. It saves huge time and manpower in almost every industry and every individuals life.

OCR technology is used very often in real world applications. Some of the applications of the OCR is Listed below.

1. Data entry of documents
2. Data extraction from KYC documents
3. Number plate recognition
4. Self Driving Cars
5. Hand Written text recognition.
6. Scanned documents to Searchable PDFs
7. CAPTCHA recognition

Objectives :

The main objective of our project is to build a real world OCR model to recognize the text in the image. As we have researched that , to build a OCR model which recognizes the whole text in the image is not scale-able and required huge computation power along with powerful models and datasets. We have also researched that most of the OCR products performs word by word extraction and concatenates the words to get the whole text extraction. Hence we are going to build the OCR model for the extraction of the single word.

In order to build the OCR model, we have some objectives to follow.

They are :

- **Data Collection :** We need to do some research and pick the right dataset for this project. We have collected some datasets and we would like to spend some more time in the data collection.
- **Data Cleaning :** It is very important clean the data after collection. Sometimes there could be wrong annotations in the dataset. More the time we spend on data cleaning , more the quality results we can get.

- **<u>Data Preprocessing :</u>** In this phase, we will take the image and perform some preprocessing techniques like data augmentation, rotation, adding noise etc.
- **<u>Model Building :</u>** Since the dataset is in the image format, it required to build a deep neural network model instead of machine learning models. We need to build a suitable model architecture.
- **<u>Model Training :</u>** In this section we will train the model and also we have to try various hyper parameters like learning rate, batch size etc.
- **<u>Evaluation of results :</u>** In this section , we evaluate the results by checking the loss and plotting the image and it's recognized text.

Features :

In general, to train a deep learning model, we require GPU computing power. By using Google Colab we can get the free GPU for limited time per day. Colab’s computer power will be sufficient to train OCR model. Colab also provides the option to mount the google drive. Since we are going with huge dataset, we can upload the dataset to the google drive once and we can easily able to mount the drive and use it for the model training.

Since we are training the OCR model from scratch, We require a huge dataset. We would like to choose the dataset that contains at least 100,000 image samples for training the model. It is also important to pick the dataset with good accurate annotations.

Some of the important modules are required for this project are:

1.Tensorflow

2.Keras

3.Numpy

4.Opencv

5.Matplotlib etc.

Unlike the classical structured datasets, we have image dataset as input. Hence we wont be having the numeric or text related columns. But any model cannot process an image directly to compute the predictions. We first need to perform the featurization. Featurization is nothing but to convert the image into an n-dimensional vector and we pass that n-dimensional vector to the neural network in order to get the predictions.Typically we convert the image into numpy array.

The technology offers a comprehensive solution for document capture and form processing. OCR typically employs a modular, open, scalable, and workflow-controlled design. It has capabilities for defining shapes, scanning, picture pre-processing, and recognition.

* The following is a summary of the main characteristics of OCR scanners based on machine learning.

* The item will assist the user in converting handwritten text into machine-encoded text.

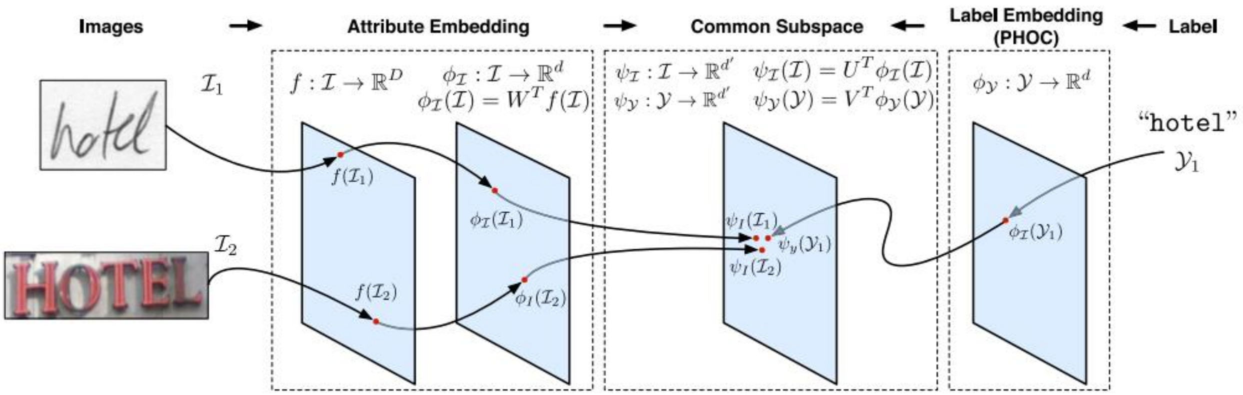
* This will make it easier for the user to keep their content in an editable state without worrying about maintaining it.

* It will be simpler to search for any text or phrase since users may turn their data into machine-encoded text.

* Data loss or unauthorized access to documents in paper format may be avoided with the help of this method.

* In paper format documents, once they are destroyed, data recovery is not feasible. However, if they are transformed into computer text documents, recovery is still possible with the right backup procedures in place.

Project Workflow :


Images
Attribute Embedding
Common Subspace
Label Embedding (PHOC)
Label
$\mathcal{I}_1$
$\mathcal{I}_2$
$f : \mathcal{I} \rightarrow \mathbb{R}^D$
$\phi_\mathcal{I} : \mathcal{I} \rightarrow \mathbb{R}^d$
$\phi_\mathcal{I}(\mathcal{I}) = W^T f(\mathcal{I})$
$\psi_\mathcal{I} : \mathcal{I} \rightarrow \mathbb{R}^{d'}$
$\psi_\mathcal{Y} : \mathcal{Y} \rightarrow \mathbb{R}^{d'}$
$\psi_\mathcal{I}(\mathcal{I}) = U^T \phi_\mathcal{I}(\mathcal{I})$
$\psi_\mathcal{Y}(\mathcal{Y}) = V^T \phi_\mathcal{Y}(\mathcal{Y})$
$\phi_\mathcal{Y} : \mathcal{Y} \rightarrow \mathbb{R}^d$
"hotel"
$\mathcal{Y}_1$
$f(\mathcal{I}_1)$
$f(\mathcal{I}_2)$
$\phi_\mathcal{I}(\mathcal{I}_1)$
$\phi_I(\mathcal{I}_2)$
$\psi_I(\mathcal{I}_1)$
$\psi_y(\mathcal{Y}_1)$
$\psi_I(\mathcal{I}_2)$
$\phi_\mathcal{I}(\mathcal{Y}_1)$

3.References:

The below are some of the initial references that we have researched as part of this project.

https://en.wikipedia.org/wiki/Optical_character_recognition

https://www.flatworldsolutions.com/data-management/articles/key-advantages-ocr-based-data-entry.php

https://www.cloudfactory.com/machine-learning/optical-character-recognition-ocr

https://www.linkedin.com/pulse/15-best-ocr-handwriting-datasets-machine-learning-limarc-ambalina/