CAPSTONE PROJECT Adamou Nchange Kouotou

Machine Learning Engineer Nanodegree March 7th, 2021

A data entry tool for KYC and customer authentication

Domain :

Finance and Insurance.

Problem statement

Verification of identity is a tedious and reccurent task that employees of financial services companies have to perform on daily basis. Before providing a service to a customer, any financial service agent has to perfom KYC procedure if it is a new customer or authentication procedure if it is an already registered customer. For that, they have to manually enter on a computer software the informations that are on the identity document provided by the client. The tedious and boring nature of this task always results in poor performance of those agents, thus leading to poor quality of data in the customer database of these companies.

Solution statement

The solution that we propose in this project is to use OCR (Optical Character Recognition) to automatize the extraction of information on identity documents. OCR is a computer vision technique used in data science to convert texts from images into machine-encoded texts.

Dataset

To build our identity document information extractor, we will use MIDV-500 dataset which contain images and video clips of 50 different types of identity documents. MIDV-500 is publicly available dataset that can be downloaded from the following server : ftp://smartengines.com/midv-500/dataset/

Because of the time constraint of this project, we will build a data entry tool that supports only three types of identity document. Thus, we will use the following three sub-datasets of the MIDV-500 dataset :

- USA passport cards ;
- Albania ID cards ;

- Brazil passport cards.

Benchmark model

We will benchmark our solution with the *''Two-Step CNN Framework for Text Line Recognition''* model described in this article : https://ieeexplore.ieee.org/document/8999509

Evaluation metrics

The Evaluation metric that we will use to mesure the performance of our model is the *Field Recognition Accuray* of the identity document. We will compare the performance of our model to that of the benchmark model.

$$Fields\ Recognition\ Accuracy\ = \frac{True\ positive}{Dataset\ size}$$

Project implementation plan

To achieve our goal, we decompose the project into the following steps :

- **Step 1 :** build an image classifier to recognize each of the 3 types of identity document of MIDV-500 dataset that we selected.
- **Step 2 :** use PyTesseract library, to extract raw text from an image of identity document and parse it to extract key informations of that identity document.
- **Step 3 :** build a Rest API that takes as input an image, if it is an identity document it returns the corresponding key informations otherwise it returns an alert.
- **Step 4 :** build mobile app that can extract key information of an identity document from a photo of that document.