Classification Project Description

November 7, 2021

Handwritten Arabic Digit Recognition

Abstract

To solve for the Ministry of Justice needs for Archive digitization, we developed a tool that classify the Arabic handwritten digits from paper documents using XGBoost Algorithm, then the tool was served using Streamlit.io

Design

Before technology dominated all aspects of the modern day, official documents used to be handwritten by professionals to document important information such as Real Estate Deeds, Financial ledgers .. etc. To properly reserve thousands of crucial documents the Ministry of Justice requires an automated tool that would digitize the document for ease of retrieval and preservation from forgery, decay and misinterpretation.

Data

The MADBase is a modified Arabic handwritten digits database with 60,000 training and 10,000 test images. It was handwritten by 700 different people. Each digit (from 0 to 9) was written 10 times.

Algorithms

Preprocessing:

Scaling between 0 to 1.

Model Building:

- K nearest neighbors.
- Logistic Regression.
- Random Forest Classifier.
- Support Vector Machines.
- Adaptive Boosting.
- Gradient Boosting.
- Extreme Gradient Boosting.

Model Testing:

- Randomized Search cross validation over Random Forest.
- Grid Search cross validation over Gradient Boosting.
-

Tools

- EDA libraries:

Pandas, numpy, open

- Model Building libraries:

Sklearn and XGBoost.

- Model Testing libraries:

Sklearn.metrics:

  - Confusion matrix
  - Classification report
  - Accuracy score
  - F-1 score