Machine Learning Engineer Nanodegree (Udacity)

Capstone Project Proposal

Karim Mohamed Talaat

August 27, 2019

- **Domain Background:**
  - Handwritten Arabic character recognition systems face several challenges, including the unlimited variation in human handwriting and large public databases, Arabic handwriting fonts is an ancient art, and there are a lot of old books and scripts need to digitize that what motivate researchers to build algorithms for this job to save ancient Arabic culture. The image shows an old Arabic script.
  - Handwritten Arabic character recognition (HACR) has attracted considerable attention in recent decades. Researchers have made significant breakthroughs in this field with the rapid development of deep learning algorithms. Arabic is a kind of the Semitic language used in countries of the Middle East as a mother language of millions people, the personal motivation is that I'm interested in the recognition systems that are always very useful.
  - Related academic research is attached with the proposal file, name of the publication is "Arabic Handwritten Characters Recognition Using Convolutional Neural Networks" by: "Ahmed El-Sawy, Mohamed Loey, Hazem EL-Bakry".
- **Problem Statement:**
  - Generally, the Arabic alphabet characters consist of twenty-eight alphabet characters that illustrated in the following table:

ص	ش	س	ز	ر	ذ	د	خ	ح	ج	ث	ت	ب	أ
sad	sheen	seen	zay	raa	thal	dal	khaa	haa	geem	thaa	taa	baa	alef
ي	و	هـ	ن	م	ل	ك	ق	ف	غ	ع	ظ	ط	ض
yaa	waw	haa	noon	meem	lam	kaf	qaf	faa	ghain	ain	zaa	ttaa	dad

  - There is also a big variety of Arabic handwriting fonts, Arabs still use till now. Most of Arab people write Arabic with two special techniques which are "**Naskh**" and "**Reqaa**" and sometimes by mistake and fast writing they write a mix of both. Also Some characters are very confusable with similar character stroke that shown in following table:

Master Stroke	ب	ح	د	ر	س	ص	ط	ع	ف	ل
Similar Characters	ب,ت,ث	ج,ح,خ	د,ذ	ر,ز	س,ش	ص,ض	ط,ظ	ع,غ	ف,ق	ل,ك

- The tiny distinction of stroke structure brings challenges for some similar character pairs, such as sad and dad in the previous table. The difference of sad and dad is the dot that above character dad, so the important role for machine learning comes now, and specially the "Classification Algorithms", which will help us to solve this problem. This problem belongs to the supervised learning.

Datasets and Inputs:

- Dataset used in this project is found on Kaggle, can find it by the following link: https://www.kaggle.com/mloey1/ahcd1
- From the content section on the previous link the details of the dataset is described:
  The data-set is composed of 16,800 characters written by 60 participants, the age range is between 19 to 40 years, and 90% of participants are right-hand. Each participant wrote each character (from 'alef' to 'yaa') ten times. The database is partitioned into two sets: a training set (13,440 characters to 480 images per class) and a test set (3,360 characters to 120 images per class). Writers of training set and test set are exclusive. Ordering of including writers to test set are randomized to make sure that writers of test set are not from a single institution (to ensure variability of the test set).

Solution Statement:

- The solution is to build a Handwritten Arabic Character Recognizer using different classification algorithms to recognize the Arabic letter found in any image.

Benchmark Model:

- The dataset is based on research attached to the proposal file, which achieved 94.9% accuracy, also mentioned some related work in the research that they achieved accuracy 77.25%, 93.92%, 73.4%, 93.8%, 93.3% and 90.73%, so we will choose the lowest accuracy mentioned in this paper **73.4%** as **our benchmark model.**

Evaluation Metrics:

- According to the dataset description, "test set is randomized, shuffled and has a good variability." So I will use accuracy score on test set to evaluate the classifier.

- **Project Design:**
  - First of all, the start will be as the following, loading the data and explore its shape, after that make some visualizations for some random sample from training set and View an Image of an Arabic letter in more details.
  - Preprocessing images: scaling, normalization and dimensionality reduction for classifiers, the dimensionality reduction will be applied is Principle Component Analysis "PCA".
  - Classification Algorithms will be applied like "SVM, Random Forest Classifier, Adaboost, Decision Trees" will be trained and their accuracy will be checked, also grid search may be tried to tune the hyper-parameters for some classifiers.
  - And reshaping the normalized data will be considered as preprocessing for MLP and CNN.
  - Deep learning models that will be used are "sequential" models based on the keras documentation guide here in the following link: https://keras.io/getting-started/sequential-model-guide/
  - It's thought that the input layer will take the image which contains the Arabic letter in a reshaped size of 32x32, some hidden layers and then the output layer which will output 28 different classes according to the number of the Arabic letters.
  - Some approaches may be taken to make some refinement for this like changing the ratio of the "Drop out" or adding "Batch Normalization".