Machine Learning Capstone Proposal

Shenouda Farouk Wahba
May 20th, 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.[1] [2] Arabic is a kind of the Semitic language used in countries of the Middle East as a mother language of millions people.

academic paper:

Arabic Handwritten Characters Recognition using Convolutional Neural Network

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 for now. Most of educated Arab people write two fonts 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	ٮ	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 bring challenges for some similar character pairs, such as sad and dad. The difference of sad and dad is the dot that above character dad.

Datasets and Inputs

I will use Arabic Handwritten Characters Dataset from Kaggle you can find it here.

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

I intend to make Handwritten Arabic character recognition system using Classification algorithms to recognize an Arabic letter from image.

Benchmark Model

The benchmark model is chosen from one of the kernels of the Kaggle competition. The benchmark model solves the same Handwritten Arabic character recognition problem mentioned in this project by using CNN. It uses the same dataset used in this project. Thus making it a perfect benchmark model for this project. The result of the benchmark model is the accuracy score of the model. The same result can be measured for our model by calculating the accuracy score. And I will Try to build Many classifiers.

Evaluation Metrics

According to 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

I will start the project with loading data and explore it’s shape after that show visualization for some random sample from training set and View an Image in More Detail.

The project separated to two parts:

- The first part using classic classifiers: Support Vector Machines (SVM), *Ensemble Methods (Random Forest, AdaBoost and Decision tree).*
- *The second part using Deep Learning: MLP and CNN.*

Preprocessing images: scaling, normalization and dimensionality reduction for classifiers.

I will train mentioned classifiers and test accuracy for them. Also, I may try grid search to tune hyper-parameters for some classifiers.

I will use normalized data to reshape it as preprocessing for MLP and CNN.

I will use trying different architectures for MLP and CNN until getting the best accuracy on test set.

References:

1. https://en.wikipedia.org/wiki/Handwriting_recognition
2. https://www.researchgate.net/publication/313891953_Arabic_Handwritten_Characters_Recognition_using_Convolutional_Neural_Network
3. https://en.wikipedia.org/wiki/Principal_component_analysis