Machine Learning Engineer Nanodegree

Capstone Proposal

Khalid Hakami
September 14th, 2018

Proposal

Domain Background

Arabic language, one of the top spoken languages in the world. A language that has lots of unique features compared to other languages. It is written form right to left and the characters are so artistic that each writer can have his own style. For all the beautiful features it has, it became so complex for computers to detect. The research in Arabic computing are not enough and there are lots of potential to improve. As an Arabic speaker, I find myself responsible to use my knowledge in machine learning to help the industry to evolve and provide more solutions to Arabic speakers.

Problem Statement

Arabic words are always written in cursive (joined) style and it has deferent shapes and technique which make it harder to read words. On the other hand, recognizing Arabic characters letter by letter is much easier. Arabic characters are used in many identification problems, such as car plates, employees' IDs and others. An auto recognition system for Arabic characters will help creating security systems, traffic monitoring and so on.

Datasets and Inputs

A data set found in Kaggle. 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 'yeh') ten times. The data was scanned at the resolution of 300 dpi. Each block is segmented automatically using Matlab 2016a to determining the coordinates for each block. 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).
(https://www.kaggle.com/mloey1/ahcd1)

Solution Statement

The proposed solution is to use CNN to detect handwritten Arabic charchters, by providing the input as csv array images.

Benchmark Model

The model will be benchmarked against a random benchmarked model. One suggested model is https://github.com/tahaemara/arabic-characters-recognition

Evaluation Metrics

I will evaluate the performance by looking at the number of correct classified characters and calculate the Accuracy of the model and then compare it to the benchmark model.

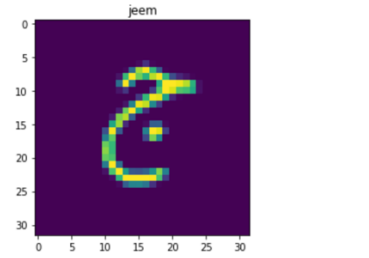
$$Accuricy = \frac{N\ correct}{N\ predicted}$$

Project Design

Data preprocessing

As the data is currently cleaned and correctly labeled and split to training and testing sets. All what I need is to implement an input convertor to make any input images meets the standard format for this model. However, that is out of the scope of this project so it will be optional.

```
import random
for r in range(0,3):
    x = random.randint(0, 3360)
    plt.imshow(testx[x].squeeze().T)
    plt.title(arabic_labels[testy[x][0]])
```



Model

I am going to build a deep learning CNN model to characterize Arabic character.

Testing

I will test the model against the benchmark model and calculate the accuracy for my model.