Plant Seedlings Classification

20.05.2018

Noura Hussein Fekry

Machine learning Nanodegree

Capstone Project proposal.

Domain Background

Image classification has become one of the most important problems that Machine learning and deep learning can solve.

In This project I will use one of kaggle Competition's dataset, this dataset containing images of approximately 960 unique plants belonging to 12 species at several growth stages
The database have been recorded at Aarhus University Flakkebjerg Research station in a collaboration between University of Southern Denmark and Aarhus University. You can find the dataset here

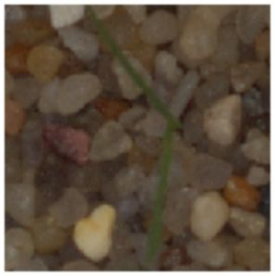
The problem here is the weed seedling is much like crop seedling and our goal is to be able to differentiate between them.

Problem Statement

The problem here is the similarity between different species of plant seedling and weed seedling

As you see in the sample pictures below:






- Machine learning algorithm can learn from huge amount of data and make a prediction with a reasonable accuracy that we can depend on it's output.
- The goal of the competition is to create a classifier capable of determining a plant's species from a photo.

Datasets and Inputs

I will use Plant Seedlings Classification dataset from kaggle you can find it here

Which has 1.6 GB as Training set (which is enough to make predictions)

And has 85.97 MB as test set (to evaluate the model we will use)

A training set and a test set containing them images of plant seedlings at various stages of grown. Each image has a filename that is its unique ID. The dataset comprises of 12 plant species, which are listed below:

Solution Statement :

The solution of this problem is to build and train a model that can classify the new unseen image into one of the twelve mentioned categories accurately.

I intend to make prediction using Neural Network, feeding the model with the training set and try to tweak the number of hidden layer to get a Acceptable accuracy.

I will attend to use CNN using tensorflow.

Benchmark Model :

I will be using a CNN model (like the one we used in “image classification” model in the deep learning Nanodegree) this dataset is very similar to CIFAR-10 dataset which used in “image classification” project (has 10 images for different objects) , and I will benchmark my model against other models from Kaggle Leaderboard to check if my Score is higher or lower.

Evaluation Metrics :

I will evaluate my model on MeanFScore

Given positive/negative rates for each class *k*, the resulting score is computed this way:

F1 = 2*(precision * recall) / (precision + recall)

Precision and recall can be calculated as :

Precision = number of TP / (number of TP + number of FP)

Recall = number of TP / (number of TP + number of FN)

Project Design:

I intend to start with with some data visualization and exploration to have more intuitive understanding, then I will divide data into training and validation sets, then I will divide training into training and testing (I may use k-fold too).

I will try extract features from images using computer vision algorithms to help me reach a acceptable accuracy.

I intend to use CNN model to help me to classify those images

will try tuning parameters for algorithms mentioned in the solution part, the best score of three of them would be considered the model.