Machine Learning Engineer Nano-degree
Capstone Proposal

Ejiroghene Precious Oghojafor
September 2019

Classify seedling species from an image

Domain Background

How do we differentiate a weed from a crop seedling? The ability to do so effectively can mean better crop yields and better stewardship of the environment. The Aarhus University Signal Processing group, in collaboration with University of Southern Denmark, released a dataset containing images of approximately 960 unique plants belonging to 12 species at several growth stages.




The field of research where this project is derived is Computer Vision. Computer Vision is an interdisciplinary field of science that aims to make computers process, analyse images and videos and extract details in the same way a human mind does. Deep Neural Networks have a (DNN) have a great capabilities for image pattern recognition and is widely used in Computer Vision algorithm, and Convolutional Neural Network (CNN) is a class of DNN which is mostly commonly applied to analysing visual imagery. This methodology can be improved and also used for human detection which can be applied to various discipline such as self driving cars and robotics. This can also be used in image classification in photos. Here is a gentle introduction to Convolutional Neural network for image classification. There have also been several research paper that attempt to use using deep learning for image classification.

Problem Statement

The problem being investigated here is a classification of seedling from an image using Convolutional neural Network (CNN). This can help improve better crop yield and better stewardship of the environment.

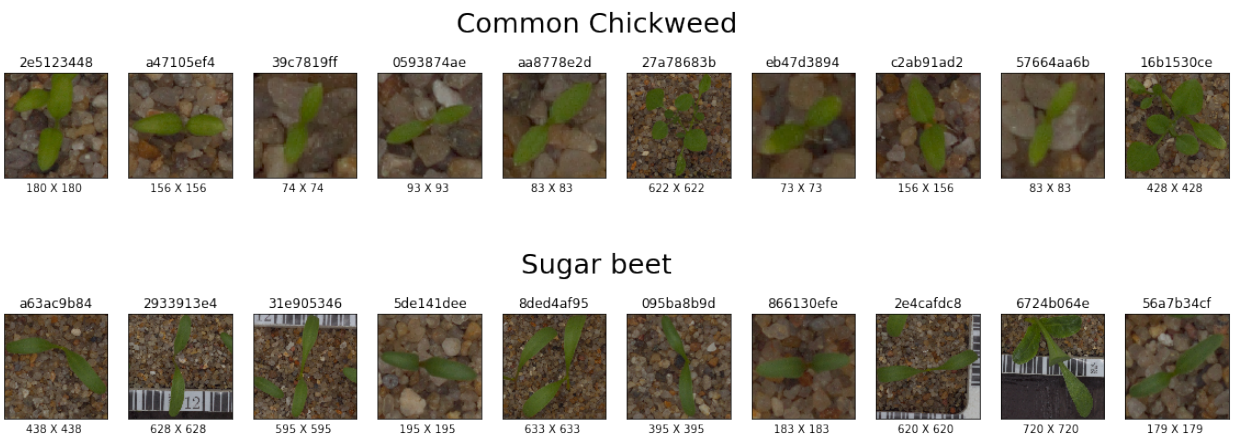
Datasets and Inputs

Train Dataset Test Dataset

The dataset are hosted by kaggle for both training and test. The training and test set contains images of plant seedlings at various stage of growth. Each image has a filename that is its unique Id. The dataset comprises 12 plant species. Some of them include, Black-grass, Charklock, Cleavers, Common Chickweed etc. Also each image has a colour layer. In the train dataset I have: 12 species in different directories each containing the following species and number of images

Black-grass: 263, Charklock: 390, Cleavers: 287, Common Chickweed: 611, Common Wheat: 211, Fat hen: 475, Loose Sliky-bent: 654, Maize: 221, Scentless Purse: 231, Small-flowered craneshill: 496, Sugar beet: 385. While the test data consist of 794 unlabelled data. I intend to use about 0.1 split for the train data.

Sample image files



Solution Statement

The solution to this problem will be a model trained to predict a species for any image. It will have an input which would be an arbitrary image and the output would be a predicted label which will be one of the 12 species. The success of the model can be measure by running the test images, The mean multi-class F1 score will be evaluated.

Benchmark Model

As this is a kaggle competition a bench mark model would be a the best kaggle score.
I can also evaluate my model using K-Fold cross validation.
I will be benchmarking against VGG16 or Inception V3 or both

Evaluation Metrics

The F1 score can be used as the evaluation metrics. The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0. The relative contribution of precision and recall to the F1 score are equal. The formula for the F1 score is:

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

Project Design

Motivation for the design gotten from Udacity dog breed classifier.

Programming language: Python 3.6

Library: pandas, numpy, sklearn, seaborn, keras, PIL, CV2

Workflow:

1) Explore the data: This involves importing the dataset, view samples of training data, number of images in the train dataset, species and image count.

2) Data Preprocessing: Here is will try to resize, normalise my images

3) Model Building: I will try to build a basic CNN and then, I will use transfer learning on top of VGG16 or Inception V3 network. This is because using a pre-trained network is efficient in CNN.

4) Algorithm Building: Once I have my model I will try to create an image classification algorithm that load and predict class using my model.

Reference

1. https://arxiv.org/abs/1711.05458
2. https://vision.eng.au.dk/plant-seedlings-dataset/
3. https://github.com/udacity/dog-project
4. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html
5. https://www.researchgate.net/publication/323894638_Machine_Learning_framework_for_image_classification