# Machine Learning Engineer Nanodegree

## Capstone Proposal

Dursun KOC
March 3rd, 2018

# Proposal

## Domain Background

A seedling is a young plant sporophyte developing out of a plant embryo from a seed. Seedling development starts with germination of the seed. A typical young seedling consists of three main parts: the radicle (embryonic root), the hypocotyl (embryonic shoot), and the cotyledons (seed leaves). The two classes of flowering plants (angiosperms) are distinguished by their numbers of seed leaves: monocotyledons (monocots) have one blade-shaped cotyledon, whereas dicotyledons (dicots) possess two round cotyledons. Gymnosperms are more varied. For example, pine seedlings have up to eight cotyledons. The seedlings of some flowering plants have no cotyledons at all. These are said to be acotyledons.

The plumule is the part of a seed embryo that develops into the shoot bearing the first true leaves of a plant. In most seeds, for example the sunflower, the plumule is a small conical structure without any leaf structure. Growth of the plumule does not occur until the cotyledons have grown above ground. This is epigeal germination. However, in seeds such as the broad bean, a leaf structure is visible on the plumule in the seed. These seeds develop by the plumule growing up through the soil with the cotyledons remaining below the surface. This is known as hypogeal germination.

The aim of this project is to build a convolutional neural network that classifies different species of plant seedlings while working reasonably well under constraints of computation with help of transfer learning technique.

Giselsson T. M., Jørgensen R. N., Jensen P. K., Dyrmann M. and Midtiby H. S. in their study[1] of “A Public Image Database for Benchmark of Plant Seedling Classification Algorithms” created a database of seedlings pictures, and tried to distinguish them by naïve bayes with a pre-processing of pictures. Now their database is placed in a kaggle competition[2], and I will be using their database provided in kaggle.

## Problem Statement

The goal is to 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 seedling data set is provided by the Aarhus University Signal Processing group, in collaboration with University of Southern Denmark. The dataset containing images of approximately 960 unique plants belonging to 12 species at several growth stages.

## Datasets and Inputs

As it is not feasible to include too many species, the researchers decided to use only a subset of high importance to the Danish agricultural industry. They have used Styrofoam boxes to grow samples. They had 56 boxes, each containing on average 25 samples with only one specie sown. According to the researchers 80 samples of individual species at the same growth stage is sufficient to capture the main variations within a specie, so 4 boxes of each species are sown to allow for 20% germination failure. Images are recorded multiple times over a 20-day period at an interval of 2 to 3 days, starting a few days after emergence. I will be downloading this dataset[3] from kaggle. Below you will find samples of each species from the database:

Maize

Common wheat

Sugar beet

Scentless Mayweed

Chickweed

Shepherd's Purse

Cleavers

Charlock

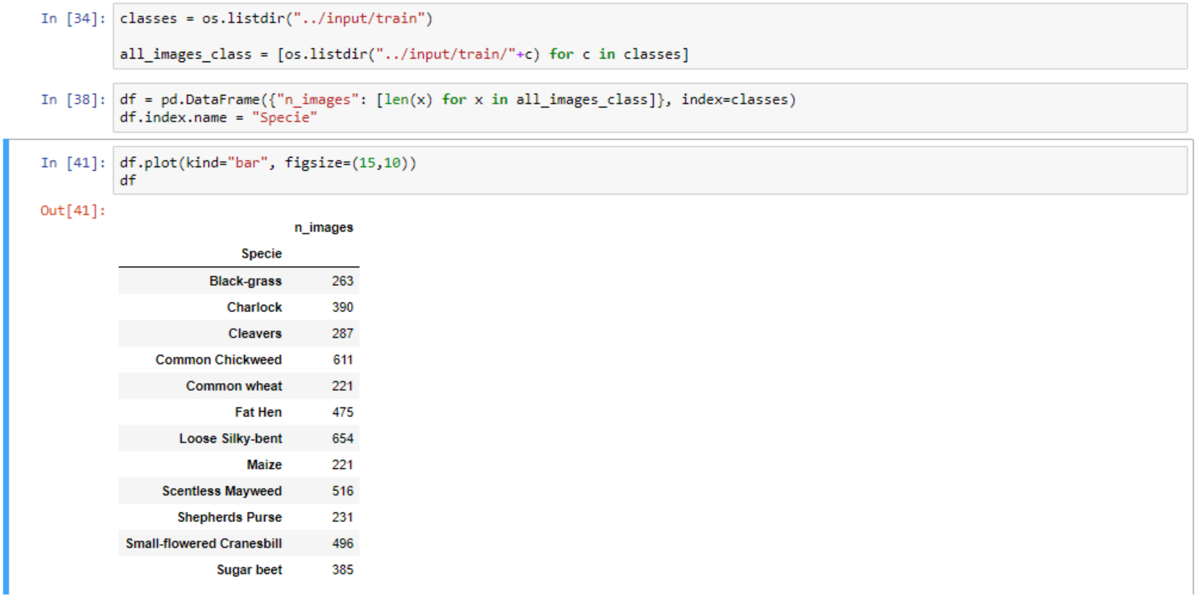
Fat Hen

Cranesbill

Black-grass

Loose Silky-bent

The distribution of each specie in the training data is as follows;



```
In [34]: classes = os.listdir("../input/train")

         all_images_class = [os.listdir("../input/train/"+c) for c in classes]

In [38]: df = pd.DataFrame({"n_images": [len(x) for x in all_images_class]}, index=classes)
         df.index.name = "Specie"

In [41]: df.plot(kind="bar", figsize=(15,10))
         df

Out[41]:
                                    n_images
                           Specie
                      Black-grass        263
                         Charlock        390
                         Cleavers        287
                Common Chickweed         611
                    Common wheat         221
                          Fat Hen        475
                 Loose Silky-bent        654
                            Maize        221
                Scentless Mayweed        516
                  Shepherds Purse        231
       Small-flowered Cranesbill         496
                       Sugar beet        385
```

Some species have more samples than some others; for example, number of “Loose Silky Bent” samples are three times greater than number of “Maize” samples.

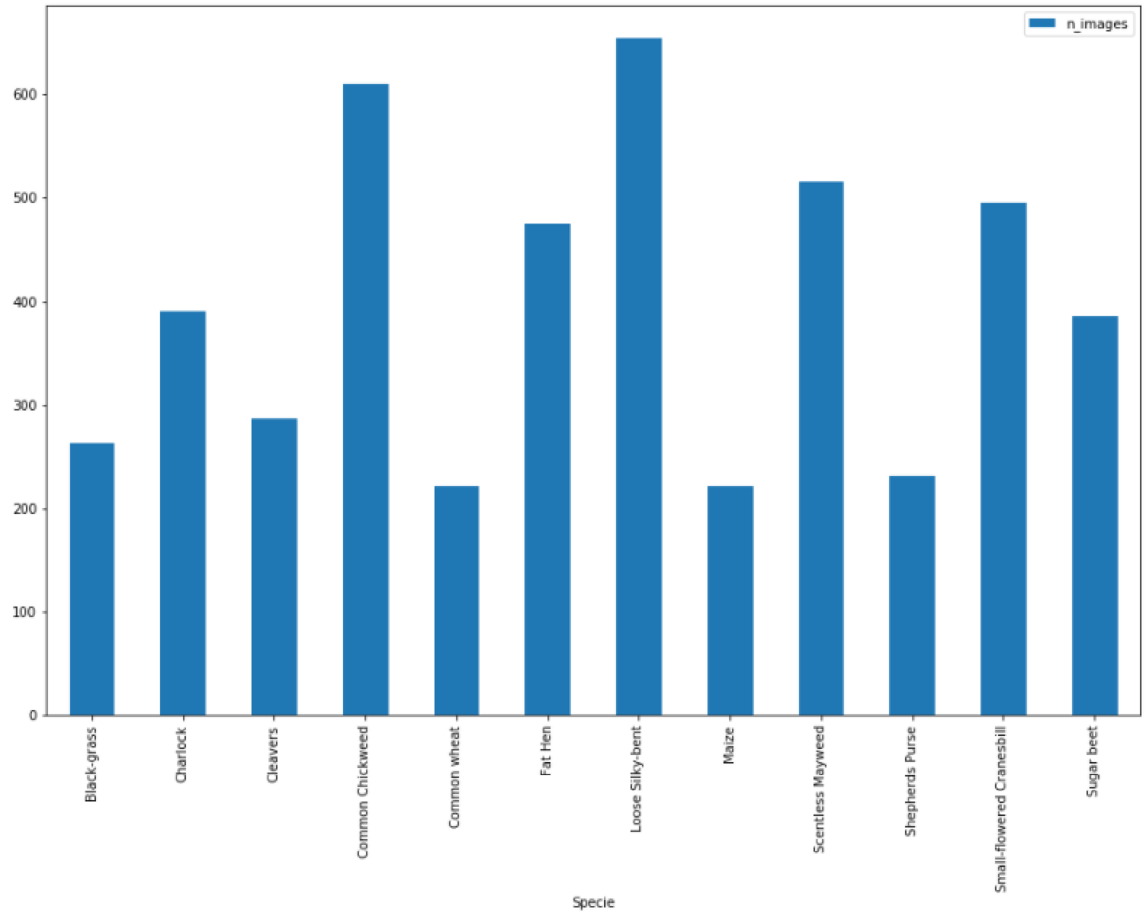


Image sizes varies so we will need to resize them when we input them to the CNN. Lastly; kaggle provides a training set and a testing set, I will use training set to generate my model with success criteria, and use the test set for validation.

## Solution Statement

Deep learning techniques have been very successful in recent years, achieving state of the art results in a wide range of domains, such as voice recognition, image segmentation, face recognition and more. In this project, transfer learning will be used to train a convolutional neural network to classify images of seedings to their respective classes. Transfer learning refers to the process of using the weights from pre-trained networks on large dataset, such as RESNET, Inception-V3 VGG-16.

## Benchmark Model

For a naïve benchmark, for a seedling image belong to any class of 12 classes is equally likely, such a submission yields 0.04534 MeanFScore. Some kagglers using CNN without any transfer learning and augmentation yields around 0.5 MeanFScore. Besides those models I will also create a one-layer CNN to get a baseline score as well.

I think a well-designed CNN, should be able to beat the random choice model, and run much faster compared to a CNN without any transfer learning and should perform better. So, the reasonable score for beating a bare CNN benchmark would be anything >0.5.

## Evaluation Metrics

The metric used for this Kaggle competition is MeanFScore, which at Kaggle is actually a micro-averaged F1-score.

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

$$Precision_{micro} = \frac{\sum_{k \in C} TP_k}{\sum_{k \in C} TP_k + FP_k}$$

$$Recall_{micro} = \frac{\sum_{k \in C} TP_k}{\sum_{k \in C} TP_k + FN_k}$$

F1-score is the harmonic mean of precision and recall;

$$MeanFScore = F1_{micro} = \frac{2 Precision_{micro} Recall_{micro}}{Precision_{micro} + Recall_{micro}}$$

## Project Design

- Programming language: Python 3.6
- Libraries:
    - Keras
    - Tensorflow
    - Scikit-learn
    - Pandas
    - OpenCV
    - Seaborn
    - Matplotlib
- Workflow:
    - In order to feed these pictures to the CNN I will need to pre-process (resize) them to the same size, also I think I will also need to convert them from RGB to BGR or maybe to HSV.
    - I will establish a baseline with a random choice of most probable specie guess.
    - I will train a small convolutional neural network from scratch for comparison.
    - I will extract features from the images with the pretrained network and will run small fully connected network with 12 output neurons on the last layer to get predictions.
    - Finally, I will fine tune the pretrained network by choosing different optimizers and by training the network on this dataset from the convolutional layers instead of the dense layers if it’s computationally inexpensive.