Machine Learning Engineer Nanodegree

Capstone Proposal

By Narendra Kumar Muthum
December 15th, 2018

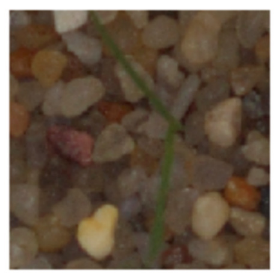
Proposal of Plant Seedlings Classification

Domain Background

The main objective of this project is 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, has recently released a dataset containing images of approximately 960 unique plants belonging to 12 species at several growth stages.






Rapid and accurate identification of weeds at the seedling stage is the first step in the design of a successful weed management program that saves producers and land managers time, money, and reduces herbicide use. How does weed seedling identification provide these benefits? First, weed management is typically much easier, less costly, and more effective at the seedling or juvenile (e.g. rosette) stage than on mature plants. Second, controlling a weed during early growth stages allows desirable neighboring vegetation to grow better, thereby improving overall plant community vigor. Finally, improper identification can result in misapplication of a management tactic such as herbicides or failure to adequately control the weedy plant species at the time that it is most vulnerable. Once a species has been correctly identified, an IWM can be designed that combines the use of biological, cultural, mechanical, and chemical practices to manage weeds. The main goals of an Integrated Weed Management program are to:

♣ use preventive tools to maintain the crop or desired vegetation and limit weed density to a tolerable, non-harmful level,

♣ avoid shifts in the composition of plant communities towards other weeds that may be even more difficult to control,

♣ develop sustainable management systems that maximize environmental quality, productivity, and revenues. Thus, designing a successful Integrated Weed Management (IWM) program requires an understanding of the biological and ecological factors that influence the growth and development of weeds. Part of this understanding is the need to correctly identify all different kinds of weed species.

To automate this process Aarhus University group partner with University of Southern Denmark are hosting this dataset as a Kaggle competition in order to give it wider exposure, to give the community an opportunity to experiment with different image recognition techniques, as well to provide a place to cross-pollenate ideas.

Problem Statement

Determine the species of a seedling from an image Species of a seedling from an images labeled by Aarhus University Signal Processing group by identifying 12 species of objects in the image such as '`Black-grass`', '`Charlock`', '`Cleavers`', '`Common Chickweed`'... along with sand, stones and bar codes.

The goal is to determine or predict the likelihood that a species is from a certain class from the provided classes, thus making it a multi-class classification problem in machine learning terms.

These twelve target species classes are provided in train dataset. The goal is to train a CNN that would be able to classify fishes into these twelve classes.

To quantifiable or measures the model by submissions are evaluated on Mean F-Score, which at Kaggle is actually a micro-averaged F1-score.

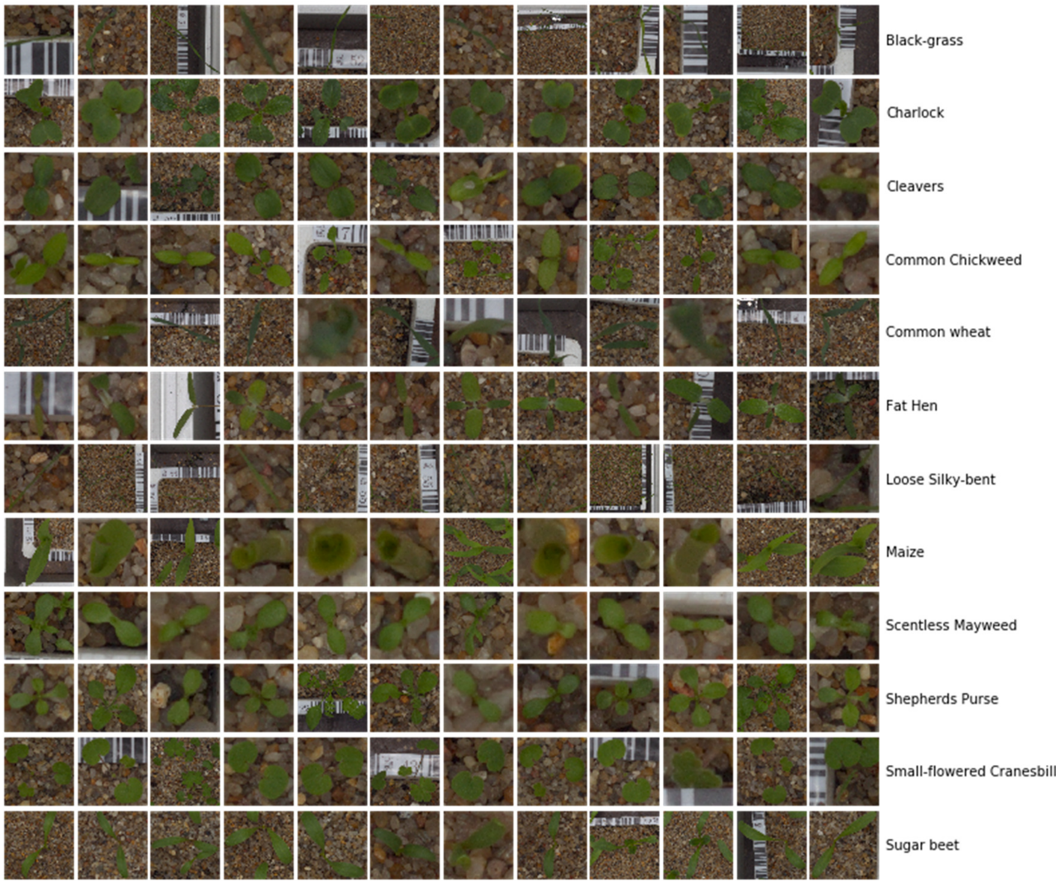
Datasets and Inputs

I extend my appreciation to the Aarhus University Department of Engineering Signal Processing Group for hosting the original data.

They have provided with a training set and a test set of images of plant seedlings at various stages of grown. Each image has a filename that is its unique id. The dataset comprises 12 plant species. The goal of the competition is to create a classifier capable of determining a plant's species from a photo. The list of species is as follows:

```
Black-grass
Charlock
Cleavers
Common Chickweed
Common wheat
Fat Hen
Loose Silky-bent
Maize
Scentless Mayweed
Shepherds Purse
```

```
Small-flowered Cranesbill
Sugar beet
```



For each file in the test set, you must predict a probability for the species variable. The file should contain a header and have the following format:

```
file, species
0021e90e4.png, Maize
003d61042.png, Sugar beet
007b3da8b.png, Common wheat
etc.
```

Solution Statement

As deep learning techniques have been widely using and very effective in image classification over the years. Fortunately many such networks such as VGG-16, Inception-V3, Xception pretrained on imagenet challenge is available for use publicly.

Given that there are 4750 labeled images (1.73GB) showing plants of 12 different types, the goal is to classify correctly the species shown on the 794 images (91MB) of the test set. All images are quadratic but vary in size. I resize them such that each image has the shape (244,244,3) or (299,299,3) according to what pretrained model we feed. Next, I detect and segment the plant parts of the images, then normalize them such that each pixel is defined on the range [-1,1]. An optional step is to generate new images through rotations, translations and axis flipping, augmenting the original data. All images are then fed into a pretrained AGG16, Xception, & InceptionV3 model provided by Tensorflow/Keras and I extract 2048 bottleneck features for each image. Having computed these features once, I train and validate a multi class logistic regression, random forest and fully connected neural network model. Finally, I predict the species classes of the test images and write the submission file. The main important note is that, this work needs more implementation and well fine-tuned and we may not get optimized model at single attempt.

Final step to optimized to minimize multi-class logarithmic loss as defined in the Evaluation Metrics section. Predictions will be made on the test data set and will be evaluated.

Submissions are evaluated on 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

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

For each file in the test set, we must predict a probability for the species variable. The file should contain a header and have the following format:

```
file, species
0021e90e4.png, Maize
003d61042.png, Sugar beet
007b3da8b.png, Common wheat
etc.
```

Benchmark Model

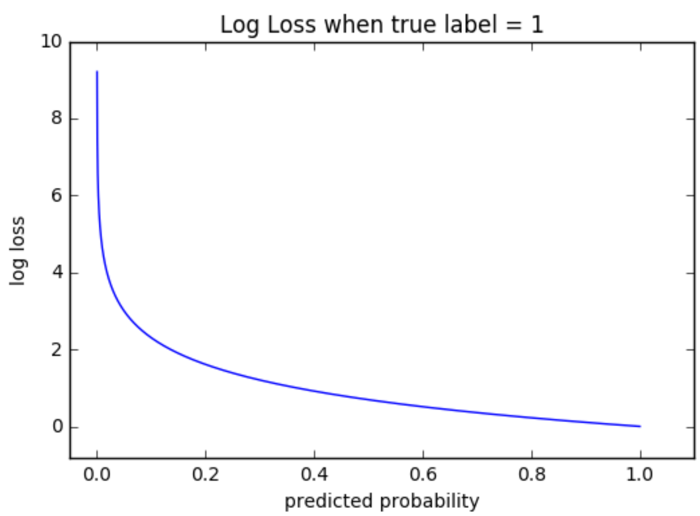
To standardise the evaluation of classification results obtained with the database, a benchmark based on f1 scores is proposed. The dataset is available at https: // vision. eng. au. dk/ plant-seedlings-dataset.

Common to all approaches is the goal of detecting weeds - either in patches or as single plants. Although some systems are commercially available, a true commercial breakthrough of such systems is still to come despite the construction of several prototypes and case studies showing promising results. The reason may be that a general approach enabling robust classification despite varying conditions and species compositions is yet to be discovered. One might ask why this task still poses a problem: botanists have been dealing with species categorization for centuries and substantial progress is reported in the field of content-based image retrieval and analysis of images and video. What makes this problem so hard? The present I believe that one problem is a lack of benchmark databases. Several studies on species recognition contain a description of preprocessing steps such as image acquisition, segmentation and annotation which suggest that researchers have spent time on these topics although each of these tasks is an area of its own. Furthermore, a performance benchmark for classification is proposed, so that using this database will permit easy replication of research results and easy comparison of algorithm performance. I will also try Ensemble methods if the hyperparameter tuning does not improve the score.

Evaluation Metrics

I will select categorical cross entropy as our loss function since the categorical cross entropy is preferred for mutually-exclusive multi-class classification task (where each example belongs to a single class) compared to other metrics. For each image, we must submit a set of predicted probabilities (one for every image). The formula of log loss is then,

$$\text{logloss} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log(p_{ij})$$



I will use the whole dataset as the baseline dataset for models evaluation because the Plant Seeding dataset is pretty small. The baseline model of our project is CNN with VGG 16 and multi-class Logistic regression. I will choose micro-averaged F1 score as the major evaluation matrix since it is selected in the Kaggle competition, and it is easier to compare the performance of our model with previous works. In addition, F1 score could balance the precision and recall and yield a more realistic indication of model performance. I would also use confusion matrix to visualize the prediction results.

Besides baseline models, I plan to experiment with models such as simple Neural Network model, CNN models (pre-trained models) and ensemble model to see their performances. The state of the art F1 score is achieved by a Xception, VGG 16, InceptionV3 models.

Project Design

- **Programming language** : Python 3.7+
- **Libraries :** Keras, Tensorflow, Scikit-learn, Preprocessing

As per above the description and problem statement it can be inferred that computer vision can be used to arrive at a solution. CNN class of deep learning algorithm can be employed for this problem.

Initially data exploration will be carried out to understand possible labels, range of values for the image data and order of labels. This will help preprocess the data and can end up with better predictions. After this necessary preprocess functions will be implemented, training data will be randomized and CNN will be implemented from scratch for further comparison with transfer learning models in Tensorflow/Keras.

Extracting features from the images with the pretrained network and running a small fully connected network output neurons on the last layer to get predictions using with logistic or random forest multi classification models on the extracted features.

Finally necessary predictions on the test data will be carried out and these will be evaluated.

Reference

[1] https://blog.keras.io/building-powerful-image-classification-models-using-very-little-data.html

[2] https://vision.eng.au.dk/plant-seedlings-dataset/

[33] https://keras.io/applications/

[4] https://datascience.stackexchange.com/questions/15989/micro-average-vs-macro-average-performance-in-a-multiclass-classification-settin/16001

[5] Giselsson, Thomas Mosgaard, Dyrmann, Mads, Jorgensen, Rasmus Nyholm, Jensen, Peter Kryger & Midtiby, Henrik Skov (2017). A Public Image Database for Benchmark of Plant Seedling Classification Algorithms.

[5] Deep Learning with Python by francois chollet, Published by MANNING Shelter Island.

[6] Python Data Science Handbook Essential Tools for Working with Data by Jake VanderPlas, O'REILLY

[7] Hands-On Machine Learning with Scikit-Learn and TensorFlow Concepts, Tools, and Techniques to Build Intelligent Systems by Aurélien Géron