Machine Learning Engineer Nanodegree

Capstone Proposal

Charles Brands
June 11, 2018

Introduction

This project will use image analysis to classify twelve different species of plants. The dataset comes from the Plant Seedlings Classification competition.

Domain Background

Farmers have been spraying their fields with herbicides for decades to reduce weeds. So far one herbicide mixture was used to spray an entire field. This is not a very effective approach. Massive usage of herbicides are expensive, bad for the environment and may lead to resistance in the weeds. Also a herbicide mixture that is effective against one species of weed can have little or no effect on another species. Not to mention that the herbicide may be damaging to the crop it is supposed to protect.

Aarhus University in Denmark is working on a project to create weed maps of a field from a tractor to pinpoint where a certain species of weed resides on a field. They have released a dataset of images of 960 unique plants belonging to 12 species at several growth stages.

Kaggle is hosting this dataset as a Kaggle competition in order to give it wider exposure and to give the community an opportunity to experiment with different image recognition techniques.

My motivation

Growing up in a rural area of the Netherlands I realize that food production is extremely labor intensive and uses significant amounts of herbicides to kill weeds. If a deep learning algorithm could distinguish the weeds from the disired plants then herbicides could be more efficiently and therefore more sparingly used.

Secondly this project is about image classification using a neural network and that field is very

hot right now.

Problem Statement

For this project we have to detect which species of plant is in the picture. Each image contains one seedling which has to be classified into one of twelve categories. The kaggle competition closed three months ago but I can evaluate any solution by measuring the mean multi-class F1 score run against the given test data set just as in the original Kaggle competition.

Datasets and Inputs

The train and test datasets can be found on kaggle. There are 4751 images for training and 794 images for testing. Each image has a filename that is its unique id. The dataset comprises 12 plant species. The list of species is as follows:

Danish	English	Latin	EPPO code
Majs	Maize	Zea mays L.	ZEAMX
Vinterhvede	Common wheat	Triticum aestivum L.	TRZAX
Sukkerroe	Sugar beet	Beta vulgaris var. altissima	BEAVA
Lugtløs kamille	Scentless Mayweed	Matricaria perforata Mérat	MATIN
Fuglegræs	Common Chickweed	Stellaria media	STEME
Hyrdetaske	Shepherd's Purse	Capsella bursa-pastoris	CAPBP
Burresnerre	Cleavers	Galium aparine L.	GALAP
Agersennep	Charlock	Sinapis arvensis L.	SINAR

Hvidmelet gåsefod	Fat Hen	Chenopodium album L.	CHEAL
Liden storkenæb	Small-flowered Cranesbill	Geranium pusillum	GERSS
Agerrævehale	Black-grass	Alopecurus myosuroides	ALOMY
Vindaks	Loose Silky-bent	Apera spica-venti	APESV
EPPO codes are computer codes developed for plants, pests (including pathogens) which are important in agriculture and plant protection.			

Solution Statement

The solution to this problem is a model trained to predict the plant species for a given image.

- Input: An image of a plant.
- Output: The label of the plant on the image.

The 794 test images Will be run through the model and the results will be combined into a single CSV file. This csv file can then be submitted to Kaggle for evaluation of the mean multi-class F1 score. Alternatively I can calculate the F1 score myself as Kaggle has published how they evaluate this project.

Benchmark Model

The results will be benchmarked against other submissions in the Kaggle leaderboard.

Evaluation Metrics

I will use the mean multi-class F1 score as the metric to evaluate my solution.
From scikit-learn.org

> 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:

$$F_1 = \frac{2 * precision * recall}{precision + recall}$$

> In the multi-class and multi-label case, this is the weighted average of the F1 score of each class.

Precision is the ability of the classifier not to label as positive a sample that is negative. In formula:

$$precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

Recall is the ability of the classifier to find all the positive samples. In formula:

$$recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

Project Design

I will attempt to solve this problem in a similar manner as the dog breeds classification project. I will use a convolutional neural network and apply transfer learning using a pretrained network such as Inception-V3, RESNET-50, VGG-16, or VGG-19.

Steps

1. Import the required data sets.
2. Display some sample images for each plant.
3. Separate the data set into training, validation, and testing sets.
4. Do some preprocessing converting to tensors and normalization and so on.
5. Possibly augmenting and oversampling the dataset by using horizontal flipping, rotation, zoom and so on.
6. Create one or more models using the transfer learning approach with pretrained network as described above.
7. Fine tune the model(s) using different optimizers and adjusting multiple parameters.

8. The F1-score will be calculated as described in the section "Evaluation metrics" above.
9. Finally my score will be compared to the leaderboard on Kaggle.

References

Plant Seedlings Classification
A Public Image Database for Benchmark of Plant Seedling Classification Algorithms
Original dataset