Machine Learning Engineer Nanodegree

Capstone Proposal

Mayank Meghawat
January 3rd, 2018

Proposal

Domain Background

Earlier developing a generic specie recognition system was not feasible because of the lack of knowledge or ability to cope with a huge variation in specie appearance.
Botanists have been dealing with species categorization for centuries, but with substantial progress in the field of content-based image retrieval and analysis of images and video. It seems possible now that we might be able to classify some of common species.

The project aims to help researchers acquire insight into the normal variation is some of the most common weed species in Danish agriculture. It is believed that if a classification approach is able to handle this data then it is likely to possess the ability to cope with high intra-class variations and be a step in the right direction towards automatic specie identification, usable for example for site-specific weed management.

Citing the paper: https://arxiv.org/abs/1711.05458
This paper standardizes the evaluation of plant-seeding classification results obtained with the database(https://vision.eng.au.dk/plant-seedlings-dataset/), a benchmark based on proposed scoring.

The target of this proposal 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.

Problem Statement

Given the Plant Seedling Dataset, containing 960 unique RGB images of 12 species, an algorithm needs to developed to classify each of the 12 species separately. This model will help us in better crop yielding and environment assimilation.

Datasets and Inputs

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. (https://vision.eng.au.dk/plant-seedlings-dataset/).

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	(263 images)
⇒ Charlock	(390 images)
⇒ Cleavers	(287 images)
⇒ Common Chickweed	(611 images)
⇒ Common wheat	(221 images)
⇒ Fat Hen	(475 images)
⇒ Loose Silky-bent	(654 images)
⇒ Maize	(221 images)
⇒ Scentless Mayweed	(516 images)
⇒ Shepherds Purse	(231 images)
⇒ Small-flowered Cranesbill	(496 images)
⇒ Sugar beet.	(385 images)

The training set consists of total of 4750 images, varying from 200~650 images per class. And the testing set contains 794 separate images.
The dataset seems good, but in case if more images are required then we will apply data-augmentation techniques on the already provided dataset to extend training data.

The image dimensions also vary from 100~2000 pixel, so will be resizing the images to 299x299.

The following problem has also been hosted as the Kaggle competition in order to give it wider exposure. (https://www.kaggle.com/c/plant-seedlings-classification)

Files descriptor provided in Kaggle competition

- train.zip - the training set, with plant species organized by folder
- test.zip - the test set, you need to predict the species of each image
- sample_submission.csv - a sample submission file in the correct format

Solution Statement

A deep learning solution will be developed using Tensorflow/Keras model and will be trained using training data.

Specifically using Transfer Learning, an already implemented model will be tuned and modified as per our requirement for minimizing multi-class classification loss, which in turn leads to an increased MeanFScore, mentioned as evaluation criteria in Evaluation Metric. Predictions will be made on test data and will be evaluated.

Benchmark Model

The model with the Public Leaderboard Score of 0.99496 will be used as benchmark score, currently this competition is ongoing so benchmark models are not yet public.

Attempt will be made so that the score obtained will be among the top 20% of the Public Leaderboard submission.

As for moving towards the goal, initially a simple deep convolution model will be created, for getting an initial accuracy. Then transfer learning will be applied and models like VGG16, Xception, Inception, etc. will be trained and tested on validation set.

Evaluation Metrics

As mentioned in the Evaluation matrix for Kaggle competition, submissions are evaluated on **MeanFScore**, given positive/negative rates for each class k, the resulting scores are computed.

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

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

F1-score is the harmonic mean of precision and recall

$$MeanFScore = F1 = \frac{2Precision_{micro}Recall_{micro}}{Precision_{micro} + Recall_{micro}}$$

The better the MeanFScore, the better our result.

Project Design

From the problem statement and the dataset, it can be inferred that computer vision will be used to arrive to the solution. CNN deep learning models will be employed for this problem.

Steps towards achieving the goal:

1) For removing the background and enhancing the image, pre-processing needs to be done:
    a) Using RGB & HSV plane, an initial mask will be generated for the given range.
    b) Morphology will be applied using a specified kernel for further enhancement.
2) As a result of pre-processing, a mask will be generated, which provides us the Region of Interest i.e. leaf region, removing the unwanted background.
3) Some Image enhancement technique like Histogram Equalization, etc. will be applied depending on the earlier outputs.
4) A deep convolution model will be created with basic layers, and initial predictions will be made using the model.
5) After applying basic deep learning model, Transfer learning will be applied. Models like VGG16, Inception, etc. will be trained and tested.
6) The Transfer learning architecture with best predictions will be chosen based on accuracy on validation set.
7) Finally, necessary predictions on the test data will be carried out and will be evaluated.