Machine Learning Engineer Nanodegree

Capstone Proposal

Mayur Selukar December 10th, 2018

Proposals

Plant Seedlings Classification *Determine the species of a seedling from an image*

Domain Background

Weeds are among the most serious threats to the natural environment and primary production industries. They displace native species, contribute significantly to land degradation, and reduce farm and forest productivity.

Invasive species, including weeds, animal pests, and diseases represent the biggest threat to our biodiversity after habitat loss. Weed invasions change the natural diversity and balance of ecological communities. These changes threaten the survival of many plants and animals as the weeds compete with native plants for space, nutrients, and sunlight.

Weeds typically produce large numbers of seeds, assisting their spread, and rapidly invade disturbed sites. Seeds spread into natural and disturbed environments, via wind, waterways, people, vehicles, machinery, birds and other animals.

The ability to differentiate a weed from a crop seedling effectively can mean better crop yields and better stewardship of the environment.

The Aarhus University Signal Processing group, in collaboration with the University of Southern Denmark, has released a dataset containing images of approximately 960 unique plants belonging to 12 species at several growth stages.

Problem Statement

This is a multiclass classification problem with 12 classes representing different plant species Input is a given image and the goal is to classify its species.

I will be tackling this as an Image Classification problem and plan to use the CNN deep learning model. Further on I will use the transfer learning technique to improve accuracy. Data augmentation will also be performed to make the model more generalized and accurate.

The target here is one of the following 12 species

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

- Shepherds Purse
- Small-flowered Cranesbill
- Sugar beet

Datasets and Inputs

The dataset is hosted on Kaggle and is free to download. The data is provided by The Aarhus University Signal Processing group, in collaboration with the University of Southern Denmark Please visit this link to get the data via kaggale.

Solution Statement

The solutions will be the prediction of the species of the plant. First I will use a pre-trained model to extract bottleneck features of the input image The models used will be Xception others in mind are MobileNet and VGG16. Representing models of three drastically different sizes.
follow this link for more info

These features will be given as an input to either a sequential model with dense layers (ANN) and a logistic model to make the predictions.

The one with the higher evaluation score will be selected for fine-tuning.

Benchmark Model

I will take the following VGG16 with 2 Additionally added densely connected layers with softmax activation at the end as the base model for the classification. I will try my best to beat this models performance. with regards to the evaluation metric below.

Evaluation Metrics

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 Reference Click here

Project Design

Before even start training models, I will first take a glimpse of the data see what the shape and how is the data organized. Once the data is loaded successfully I will plot some example images from each category for

Reference. This data will be then sent through a data generation which will augment the images to generate new ones.

To train models, As mentioned earlier I plan to choose 3 different models to compare. Because this is a classification problem, a few models for bottleneck feature extraction in my head are Xception, VGG16, and MobileNet These bottleneck features are then fed to another model for classification some which come to mind are a regression, decision trees, Dense Net and random forest. Using cross-validation I can find which model performs best, and then use that one to tweak relative parameters.

I expect to spend 60% of the time on data cleaning, testing and training different models and 40% of the time on tweaking parameters. The final accuracy will be calculated against the test dataset provided by kaggle