Machine Learning Engineer Nanodegree

Capstone Proposal

Sandeep Bethapudi
September 28, 2018

Proposal

Domain Background

Can you differentiate a weed from a crop seedling or plant?

Doing so effectively can mean better crop yields and better stewardship of the farms, public spaces, and the environment. Given that weed can destroy crops and use up resources such as water, space, etc., and agriculture needs to be an efficient process to minimize losses and provide a sustainable crop production, it is important that an automated weed management system is used by farmers and others where a differentiation between plants and weeds are needed. I grow a small garden in my backyard with various vegetables, flowers, and plants and have encountered annoying and unwelcomed weeds that are sometimes hard to differentiate from plants. I have also volunteered to clear out invasive plants including weed at my local community parks where we were either given a paper description of the invasive species or were shown a few samples. However, while actually clearing it became obvious that picking the right plant was difficult oftentimes than not. So an image classifier for plant differentiation from invasive species like weed would be extremely useful.

Problem Statement

The problem is that it is indeed difficult to differentiate weed plant from a plant seedling and similar looking plant seedlings from each other given that the plants seedlings are not fully developed yet, so their distinctive features, that could be used for differentiation, many not be fully obvious.

Datasets and Inputs

The dataset was obtained from Kaggle provided by the Arhus University Department of Engineering Signal Processing Group. Kaggle provided 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 of 12 plant species. The goal 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, and sugar beet. Given that the background if undesirable, I potentially need to use masking and segmenting.

Solution Statement

To build and train a model that will be able to classify a new image to one of the twelve plant categories with high accuracy by outputting a single category with highest probability as a prediction of the input plant image.

Benchmark Model

My benchmark mode will be a simple convolutional neural network Keras model or a SVM that will be used to train and test the dataset and will be used in comparing results of my final pre-trained classifier model.

Evaluation Metrics

The evaluation metric listed on Kaggle is the Mean F Score and I intend to use that to evaluate the performance of my classifier. F score The Mean F score is the weighted average of the F scores of each class.

F1 = 2*(Precision*Recall)/(Precision + Recall)

Project Design

For the final classifier: I will first perform exploratory data analysis to get a “big picture” of the dataset and some cleaning and pre-processing. Pre-processing can include image scaling, normalization, one-hot encode uniform aspect ratio, and data augmentation. For my final classifier model, I intended to use a pre-trained convolution neural network such as Inception, Xception, etc., with Keras and extract bottleneck features, and

perform logistic regression on the bottleneck features. Then I'll compare these models before choosing one as my final classifier based on validation accuracy and compare their confusion matrix.