Machine Learning Engineer Nanodegree

Capstone Project Proposal

Joel Newswanger
December 7, 2017

Proposal

Domain Background

One of the major challenges faced by agriculture is weeds. Weeds compete with crop species for nutrients, sunlight, and water, and often out-compete crops for these, due to their greater growth rate. Since they are a drain in the production of crops the presence of weeds greatly decreases the efficiency of agriculture economically and environmentally.

Crops are especially vulnerable to weed encroachment when they are seedlings. This is because once a plant is grown it will shade the ground and take up the water and nutrients, making it harder for other plants to grow in the same area. This means that whichever plant grows larger first will dominate, and as before mentioned, the weeds grow faster than crops. The otherwise empty field full of seedlings is a prime place for weeds to spring up and overtake.

Current methods for dealing with weeds in fields that have crop sprouted in them are full of problems. To be able to spray the weeds with herbicide, which is the conventionally preferred way, the field must either not have any crop in it yet, or be planted with an herbicide resistant crop. Some people have problems with such crops, overuse of pesticides causes environmental problems, and weeds can become herbicide resistant. Manual removal of weeds from amongst crop seedlings is very difficult and can be not great for the soil too.

A proposed solution includes robotic weed exterminators (not a technical term) that roam the fields, finding and killing weeds, like this one designed by Tijmen Bakker. These robots would need to visually identify and classify weed and crop seedlings, roast the weeds with a laser, and pass harmlessly over the seedlings. The laser works well because heating a plant to a high temperature is a reliable way to kill it to the roots, the water in it conducts the heat throughout the whole plant. Precision sprays of herbicide, and manual extraction are also possibilities for different applications.

Problem

These robots will need to be able to identify and classify crop and weed seedlings from images in order to know which ones to protect and which to unleash their destructive capabilities on. For this the designers look to machine learning and computer vision. Many CNNs have proved very useful in computer vision/image classification problems. Since the data set available for pictures of seedlings is rather small a good and compute-time efficient way to make a CNN for it is to use the bottleneck features from a pre-trained very large CNN like xception or ResNet.

Dataset

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 as a Kaggle Competition. The species are 3 common crops and 9 common weeds as seedlings. The photos are looking straight down at the seedling from above, and below the plant are pebbles that are pretty standard in all the pictures, a factor that I think will make this demonstration easier than a real life scenario would be. The data is provided as files labeled by species that each contain pictures of that species, and a test set is included to enter in the Kaggle competition leaderboard. The images vary in dimension. I will also test my CNN on some images "from the wild" produced by the same authors of the original data.

Kaggle Competition Data

A Public Image Database for Benchmark of Plant Seedling Classification Algorithms

Solution Statement

My solution to the problem will be to create a CNN from the bottleneck features of a pre-trained CNN; one of xception, ResNet, VGGNet, or inception using some of the methods in this Keras Blog post. Using Keras I will construct the additional layers and train them. Once trained, my CNN will be able to look at a picture of a plant and return the probabilities of it being each species. Since the autonomous weeding robots only care whether the plant is a weed or a crop the probabilities can be summed. Thus it can decide to kill a weed even if it can't determine which of two weed species it actually is. For testing against the Kaggle leaderboard the results will need to be returned per species, but in the agricultural terminator application, accuracy can be increased by only returning if its a weed or not.

Benchmark Model

Kaggle provides a benchmark model for the competition leaderboard. This benchmark is set at a mean f-score of 0.09823 on the testing data provided which is only slightly better than randomly guessing. I will make a simple CNN from scratch to use as a better benchmark. I expect this will be at least 50%.

Evaluation Metrics

The mean f-score is used to quantify the performance of the CNN, and is used for the Kaggle leaderboard. F-score if the harmonic mean of precision and recall. At Kaggle the f-score is micro-averaged to make the mean f-score. For simplicity I will use the same metric as Kaggle uses for my own evaluations as well.

Project Design

To start I will need to pre-process the data, including image augmentation. I will then choose a pre-trained network that has a good amount of similar data to get bottleneck features from and run the convolutional part on my training and validation data once and save the bottleneck features.

Next, I will train a fully connected model on the data from the bottleneck features. I will experiment with different arrangements of layers until I have a pretty good model. I will test it against the Kaggle leaderboard benchmark to see how I am doing. I will next try to use logistic regression on the bottleneck features and compare the results with the NN and use the better one.

I will then work to fine tune the model as best as possible by trying dropout, weight decay, and fine-tuning some of the convolutional blocks. Finally, I will test it again against the leaderboard, and the “in the wild” data to see how it would perform.