Seedling Classification

Domain Background

The project proposal comes straight from the domain of agriculture and farming. This project aims to answer the age-old question - Can you differentiate between a crop and a weed? In the country of India, the answer to this question could mean the difference between life and death for a population of about 500 million farmers and agricultural labourers spread across the rural regions.

As Uncle Ben once said - With great power comes great responsibility. With the advent of technology, it is up to us, the Data Scientists and Engineers of this country to use our skills to help make their lives easier. Research is booming in this domain and we have students and organisations working on solving problems by means such as using drones to detect diseases in crops, water scarcity prediction, crop management, yield prediction the Agrobot project and so on. By providing machine learning solutions to the domain of Agriculture and educating the people involved about why this technology matters, we can help alleviate some of their troubles and do our part in pushing India towards a better tomorrow.

Why this project?

A weed is a plant growing where it is not desired. They not only compete with crop plants for plant nutrients, moisture, space and sunlight but also interfere with agricultural operations increasing the cost of labour and tillage. And ultimately affect the yields and quality of farm produce adversely. They cause damage to crops in ways including but not limited to:

1) Increase in cost of cultivation
2) Reduction of crop yield
3) Harbouring of insects and diseases
4) Increase in seepage losses
5) Reduction in land quality

The timely detection and elimination of this harmful flora could be key to improving the quality and quantity of agricultural products each year.

Problem Statement

The Problem Statement chosen for this project is to build a machine learning model which when trained, can take as input, an image of a Seedling similar to that of the Dataset and then accurately classify it as one of 12 different plant species (including weeds and crops) that are used to train the model. This is a dataset that contains both crops and weeds, many of which might seem indiscernible to the human eye. My hope is that the trained model would be able to pick up on the common traits of the different species and help us easily distinguish the weeds from the rest.

Dataset and Input

The Dataset that I will be using for this project is the one freely available on Kaggle here: https://www.kaggle.com/c/plant-seedlings-classification/data . It consists of a training set and a test set of images of plant seedlings at various stages of growth. Each image has a filename that is its unique id. The dataset comprises 12 plant species. It comprises of annotated RGB images with a resolution of about 10 pixels per mm. The goal of the project is to create a classifier capable of determining a plant's species from such a photo. The list of species is as follows:


Maize

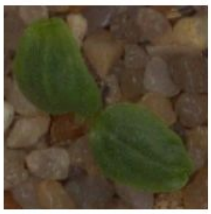

Common wheat


Sugar beet


Scentless Mayweed

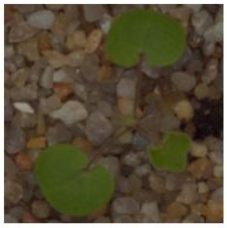

Chickweed


Shepherd's Purse


Cleavers


Charlock


Fat Hen


Cranesbill


Black-grass


Loose Silky-bent

Solution Statement

The Solution that I propose is to make use of a Convolutional Neural Network to pick up on the subtle mathematical features of each plant and hence build a model that is able to distinguish between them as accurately as possible. From my last CNN project, I realised the power of Transfer Learning and here, I intend on exploring the same ideas. I plan on using a pre-trained neural network such as the Xception or Inception V3 models to train the last few layers in an attempt to leverage the pre-learned hidden features for my use case.

In addition to the above-mentioned method for generating results, I will also be using some additional concepts to aid me along the way:

1) Since the data set is pretty small and rather imbalanced, I will also be using Data Augmentation to improve model performance
2) The project will also involve a number of visualization that I will include for easier analysis. I also want to try and make use of a T-SNE plot since I've read a lot about its usefulness and I would like to see it in action.
3) Standard Data Normalization and cleaning techniques will also be used.

Benchmark Model

Since the data and the problem statement are both from Kaggle, we can take a look at the public leaderboard to get a sense for how well others have been able to solve this problem. But before that, I have three approaches in mind:

1) First, consider a random chance, where the model randomly guesses the class labels for each new data point. We can safely expect the results of such an experiment to be well below 1/12 as there are 12 different class labels and the data is not evenly distributed.
2) Consider using a Vanilla CNN built from scratch. This is also something that I intend on trying out just so that I can compare my final model to one trained from scratch.
3) Build a model using Transfer learning and compare my scores with that of the Kaggle Leaderboard.

My goal would be to get as high up on the leaderboard as possible.

Evaluation Metrics

The Kaggle competition description describes using the mean F-score as the metric for the evaluation and that is what I will also be used in this project. So it's worthwhile to know what the F Score really is. The traditional F-measure or balanced F-score (F_1 score) is the harmonic mean of precision and recall:

$$F_1 = \left(\frac{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}}{2} \right)^{-1} = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}.$$

The same evaluation metric will be used to evaluate both the benchmark as well as the final models. More information about F1 score here: https://en.wikipedia.org/wiki/F1_score

Project Design

Here's how I intend on carrying out this project:

1) First, I plan on analysing the data and performing data transformations and normalizations wherever required.
2) Once that is done and I am satisfied with the data quality, I intend on using as many data visualizations as necessary to give me an as clear an idea of the distribution of the data as possible. I want to take this opportunity to enhance my ability to deal with a variety of visualization strategies.
3) I would then set up a benchmark model using a Vanilla CNN and record its performance.
4) Using different approaches and pre-trained models, I would try to leverage their learned features to improve my model's performance, all the while comparing the results with that of the leader board scores.
5) I plan on using a heat map to finally display how well my model is able to predict the class labels for the data in the test/validation set.

References :

[1] http://oer.nios.ac.in/wiki/index.php/Damages_caused_by_Weeds

[2] https://www.kaggle.com/c/plant-seedlings-classification

[3]https://www.quora.com/How-many-farmers-or-agriculturists-are-there-in-India-and-how-many-are-interested-to-work-in-agricultural-throughout-India

[4] https://en.wikipedia.org/wiki/F1_score

[5] https://vision.eng.au.dk/plant-seedlings-dataset/