Machine Learning Capstone Project

Plant Seedling Classification

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

Domain Background:

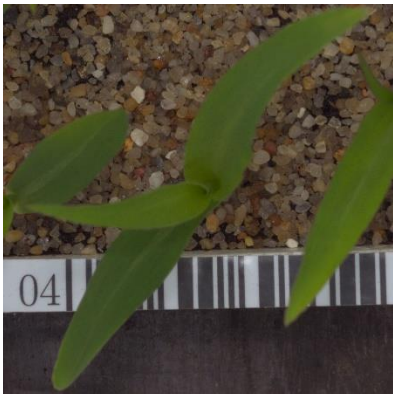
My Capstone project for the MLND is going to be the Plant Seedling classification project obtained from the Kaggle Competition website. Kaggle is a data science/machine learning platform where people, organizations, companies hold competitions to solve a global/personal problem. Usually, the competition awards the winners some prize money depending upon the complication and magnitude of the project. The winners are decided by an evaluation metric such as validation accuracy, f1 score, loss and so on. My project is the Plant Seedling Classification project which is competition active right now on Kaggle. Classifying plants is extremely important in today's world. It helps people, especially botanists and farmers, to organize certain species of plants into their categories. It can be used by scientists studying plants in a forest, by environmentalists analyzing the local environment to determine what types of plants grow in specific areas, and maybe by explorers exploring forests to identify certain species of plants. The project statement on the Kaggle website states that "The ability to do so effectively can mean better crop yields and better stewardship of the environment." My mother was a biology/botany lecturer, and her interest grew on me. Although I do not plan on becoming a botanist when I grow up (Still a junior in high school, so my future at this point is very unpredictable), plants do fascinate me. This project was perfect to deploy my interests to work. There have been past, successful attempts at solving the problem of image classification. Some of them include studies by Brown University (https://news.brown.edu/articles/2016/03/leaves) and other individual contributers such as Biva Shrestha - (https://pdfs.semanticscholar.org/0b59/84670d1a2ea3114b3b1394b5e29c5bb5e8b2.pdf).

Problem Statement:

The goal of this project is to classify images of plant seedlings into twelve different species: Black-grass, Charlock, Cleavers, Common Chickweed, Common wheat, Fat Hen, Loose Silky-bent, Maize, Scentless Mayweed, Shepherds Purse, Small-flowered Cranesbill, Sugarbeet. The images should be trained with and classified into their species for the model to be used later when images in the sample space do not have a label and need to be identified.

Datasets and Inputs:

The training dataset (available at https://www.kaggle.com/c/plant-seedlings-classification/data) entirely consisting of 4750 images came in 12 separate folders that serve as labels to the pictures. The folder names are the names of species and inside the folders are the corresponding images. All photos are quadratic but vary in size, and so the images need to be resized to the similar dimensions. The images are pictures of the plants along with various external elements such as pebbles and in some instances, a bar code (not relevant to the project).



Since not all images have the exact external features placed at the precise location, this will be a factor that needs to be considered when feeding images to the model. The best solution is to eliminate these features to avoid overfitting and a possible decrease in accuracy. Since the dataset only has 4750 images, the accuracy might be lower than usual because of the more the input, the better the accuracy.

The test set is one folder with random images and no labels. The results are to be submitted via an excel document with the image-id and its prediction.

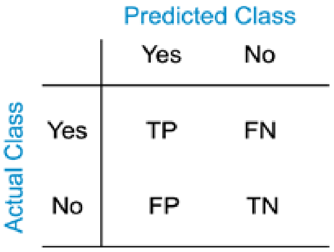
Data should be used because it is one of the essential pieces of machine learning and data science because, without data, machine learning is impossible. For a model to perform a required task, it has to train, and the training happens on data. In this instance, the data consists of pictures and words. Both have to be converted to numbers that the model can interpret and then switched back into to desired output. Regarding Data Distributions, each category has different number of images.

Category	Number of Images
1. Black-Grass	263 (5.5%)
2. Charlock	390 (8.2%)
3. Common Chickweed	611 (12.8%)
4. Common Wheat	221 (4.6%)
5. Fat Hen	475 (10%)
6. Loose Silky-bent	654 (13.7%)
7. Maize	221 (4.6%)
8. Scentless Mayweed	516 (10.8%)
9. Shepherds Purse	231 (4.8%)
10. Small-flowered Cransebill	496 (10.4%)
11. Sugarbeet	385 (8.1%)
12. Cleavers	287 (6%)
TOTAL	4750

For this to be an even distribution, each category(species) should be about 8.3%. or 394 images.

Solution Statement:

One solution to solve this problem is to use a two-dimensional convolutional neural network that studies the image structure, trains and predicts the class of a given image. First, the images should be segmented using OpenCV to eliminate unwanted features such as the pebbles and the barcode. The labels should be binarized (words to matrices) so that the model can understand the label. The optimizer for this model will be “adam” from tensorflow optimizers since it works well most of the times. The evaluation includes training accuracy, categorical cross-entropy loss, and validation accuracy. Finally, testing the model on the test set will also give an estimation of the performance of the model. However, Kaggle does not provide us with labels of the test set and evaluating the solution to the test set might not be possible.



		Predicted Class	
		Yes	No
Actual Class	Yes	TP	FN
	No	FP	TN

Benchmark Model:

The benchmark model will be a tensorflow, two-dimensional convolutional neural network. This network will have the following layers: conv2d, max_pool, dropout, fully_connected, and regression. Filter sizes: 32, 64, 128, 64, 32 and I do think these are enough for my model since there is a risk of overfitting. The relu activation function will be used throughout the model and the softmax for the output layer. The learning rate for this model is going to be 0.001.

Evaluation Metrics:

The evaluation metrics for this project are going to be- Accuracy and Categorical Cross-Entropy loss. Accuracy is usually defined by the confusion matrix.

ACCURACY = TP+TN/TP+FP+FN+TN

Categorical Cross Entropy Loss = $Hy'(y):=-\sum i(y'i\log(yi)+(1-y'i)\log(1-yi))$

Project Design:

In summary, I would use a 2-dimensional convolutional neural network to solve this problem. The first step is to extract and organize all the data into customized arrays so that I can use them throughout my notebook. I will do some image processing such as resizing, masking, segmenting and sharpening using OpenCV. Next, I will binarize the labels and divide the training set into train and validation set. The next step is to define my network and feed my labels and pictures and lastly test my model. The toughest part is determining filter sizes, kernel sizes, and strides. Since there is no pre-determined number, experimenting is the only solution. Using evaluation metrics such as loss and accuracy, I will determine and fix my model for optimal performance.