Plant Species Classification using V2 Plant Seedling Dataset (Kaggle)

Project Proposal

Udacity Machine Learning Engineer Nanodegree

Jubin Mohanty

2nd July 2021

Domain Background

As the University of Pretoria stated, the successful cultivation of crops broadly depends on weed control strategies. It has been widely observed that tons of cultivated crops are wasted due to crop infestations. A zero percent loss in crops is entirely a rare event; generally, there are 10 to 100% losses as per the current weed control practices. (Pretoria, n.d.)

Additionally, for farmers, it is of vital importance to detect weed in the initial first to six weeks of plantation because both weed and crop vigorously search for nutrients and water in the soil for this specific period.

According to the research paper published by Thomas Mosgaard Giselsson and co, there is no robust computer vision system out there that can classify ground-based weed species. (Giselsson, Jørgensen, Jensen, Dyrmann, & Midtiby, 2017)

In this project, I have developed a CNN model which can detect plant species with up to 92% of accuracy.

Problem Statement

Here, I am using a dataset that contains 5,539 images of crop and weed seedlings. The images are grouped into 12 classes. Moreover, these classes represent common plant species from Danish Agriculture. (Dyrmann)

Each class contains RGB images, which reflect various stages of plant growth.

Using this dataset, the goal is to build a model to further classify weed seedlings and crops.

Moreover, I have also planned to further extend this project by integrating this Deep Learning API with a web or mobile application, where a farmer will have to upload the image in the mobile or web application, and the API will be able to predict the species of the weed or crop. Thus, a farmer can take

appropriate action. In this way, if a farmer found a specific seedling as a weed, it can be destroyed before it infested the actual crop.

Datasets and Inputs

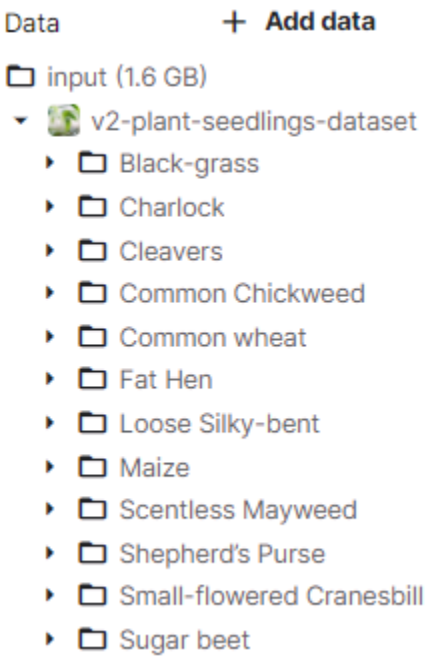
The dataset contains 5,539 images of crop and weed seedlings. The images are grouped into 12 classes. Moreover, these classes represent common plant species from Danish Agriculture. (Dyrmann)

Each class contains RGB images, which reflect various stages of plant growth.

The actual data set can be found out at the link mentioned below.

https://www.kaggle.com/vbookshelf/v2-plant-seedlings-dataset

Input Dataset Structure:



Solution Statements

While developing the CNN model, I used an InceptionResNetV2 as my pre-trained mode, which then further feed to a hidden dense layer with 128 layers, and finally to the output layer consists of 12 layers after flattening the pre-trained model.

After 35 epochs, the best model is selected as per the callback parameter, which results in significant improvement in training and validation accuracy.

Also, the validation accuracy is about 92%, which is better than what I set it as a benchmark.

Inception-ResNet-V2: It is a convolutional neural network which is trained on more than 3 million images from ImageNet database. It has 164 deep layers and can classify images into 1000 categories. (inceptionresnetv2, n.d.)

We can classify new images using this, and that is why I use it as my pretrained model.

Also, it performed better than VGG19 and 16.

Benchmark

As a benchmark for this classification model, I followed Kaggle evaluation metrics. As per Kaggle, for this competition project, meanFScore, which is actually a micro-averaged F1-score, is considered as a benchmark. It is nothing but the harmonic mean of precision and recall. (Plant Seedlings Classification, n.d.)

Moreover, I took the leaderboard score as a benchmark which is 0.99. Also, while developing the CNN architecture, my first goal was to secure validation accuracy around 70-80% by using only custom layers.

After achieving this, I incorporated a pre-trained model, which is already fine-tuned to optimize the model further, which increases the model accuracy.

Evaluation Metrics

In this project, the evaluation metrics that I have used to validate the model and improve model accuracy are Validation Accuracy, Precision, Recall, F-1 score, and Confusion Matrix.

Also, Kaggle use weighted F-1 score as the benchmark to evaluate this specific classification model.

Project Design

I divide the project design into 7 parts:

1. Data Preprocessing: Here, my primary motive was to balance the data, and segregate into train and test after image resizing.
2. Data Augmentation: In this step I artificially expand the training dataset to create modified versions of the image. In this way, it will improve the performance of the model to generalize.
3. Data Loading: It will basically load images to train, validation and test object from the respective class(species) folder containing images, which is further used while training, and prediction.
4. Developing the CNN model: I used an InceptionResNetV2 as my pre-trained mode, which then further feed to a hidden dense layer with 128 layers, and finally to the output layer consists of 12 layers after flattening the pre-trained model.
5. Train the model: In this step, I **train** the model, where I pass training tensor data, epochs, steps per epoch, validation data and steps, and callbacks as its parameter values.
6. Finally, created an object for predicting species of test images using predict_generator method.
7. Also, I created a submission csv file, which contain test image file names and predicted species names.

References

Dyrmann, M. (n.d.). *COMPUTER VISION AND BIOSYSTEMS SIGNAL PROCESSING GROUP.* Retrieved from https://vision.eng.au.dk/plant-seedlings-dataset/

Giselsson, T. M., Jørgensen, R. N., Jensen, P. K., Dyrmann, M., & Midtiby, H. S. (2017, Nov 16). *A Public Image Database for Benchmark of Plant Seedling Classification Algorithms.* Retrieved from https://www.researchgate.net/publication/321095442_A_Public_Image_Database_for_Benchmark_of_Plant_Seedling_Classification_Algorithms

inceptionresnetv2. (n.d.). Retrieved from mathworks: https://www.mathworks.com/help/deeplearning/ref/inceptionresnetv2.html;jsessionid=75bcffbe95974730090da60233c1

Plant Seedlings Classification. (n.d.). Retrieved from kaggle: https://www.kaggle.com/c/plant-seedlings-classification/overview/evaluation

Pretoria, U. O. (n.d.). *Important Weeds in Maize*. Retrieved from https://www.up.ac.za/sahri/article/1810372/important-weeds-in-maize