Plant Species Classification prediction

Introduction to Computer Vision Project
Plant Seedlings Classification Study

October 04, 2023

By: Yolanda OMalley

Contents / Agenda

- Executive Summary
- Business Problem Overview and Solution Approach
- EDA Results
- Data Preprocessing
- Model Performance Summary
- Conclusion
- Appendix

Executive Summary

Objective:

The aim of this project is to Build a Convolutional Neural Network to classify plant seedlings into their respective 12 categories.

Executive Summary

Actionable Insights:

Using data augmentation to overcome the imbalance problem, it help improve the CNN model.

Batch Normalization and Reducing the Learning Rate has also helped in improving the CNN model.

Executive Summary

Recomendations:

The field of agriculture can benefit the workers in this field, as the time and energy required to identify plant seedlings will be greatly shortened by using CNN model 2.

CNN model 2 could be improved with a better test accuracy than 80%

The confusion matrix of model 2 appears to be improving as well, however there is still some confusion with the black grass & Loose Silky bent classes of plant species.

These model can be further improved by training with different filter sizes and different number of filters.

Data Augmentation can be performed more, and dropout rate can be changed to improve the model performance.

Treating the class imbalance by increasing the weights of the minority classes could help improve the model.

Business Problem Overview and Solution Approach

We will be focusing on:

- Building a Convolutional Neural Network model which would classify the plant seedlings into their respective 12 categories to benefit the workers in this field, *as the time and energy required to identify plant seedlings will be greatly shortened by the use of the CNN model.*

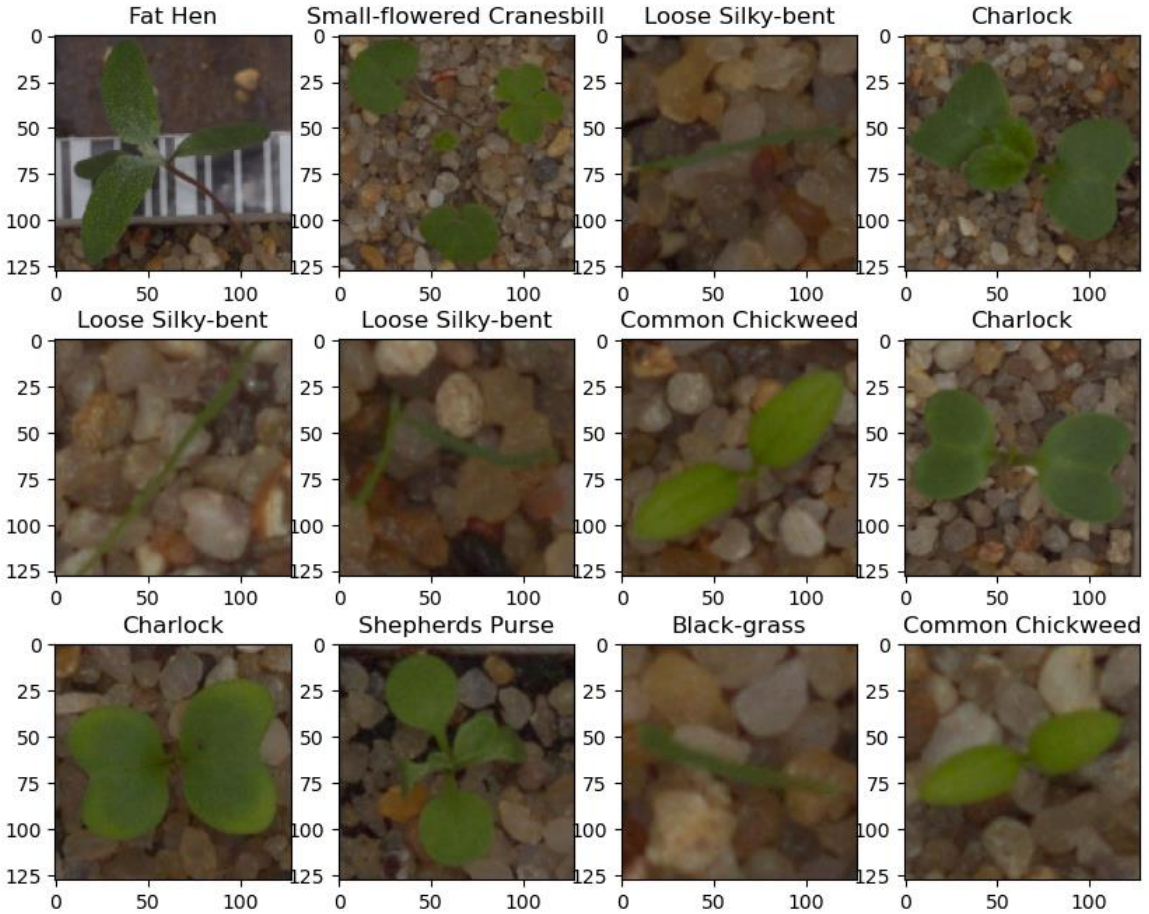
-We will use data preprocessing and EDA using descriptive statistics and visualizations. Convert the BGR images to RGB images. Resize the images. Plot the images before & after the pre-processing steps. Split the data into train and test. Encode the target variables & Apply the normalization

- We will Build the CNN model with Adam as optimizers to Predict *plant seedlings classification* and analyze these predictions to gain insights.

-Model Performance Improvement is used: data augmentation, Batch Normalization & ReduceLRonPlateau

EDA Results

Plotting random images from each of the class

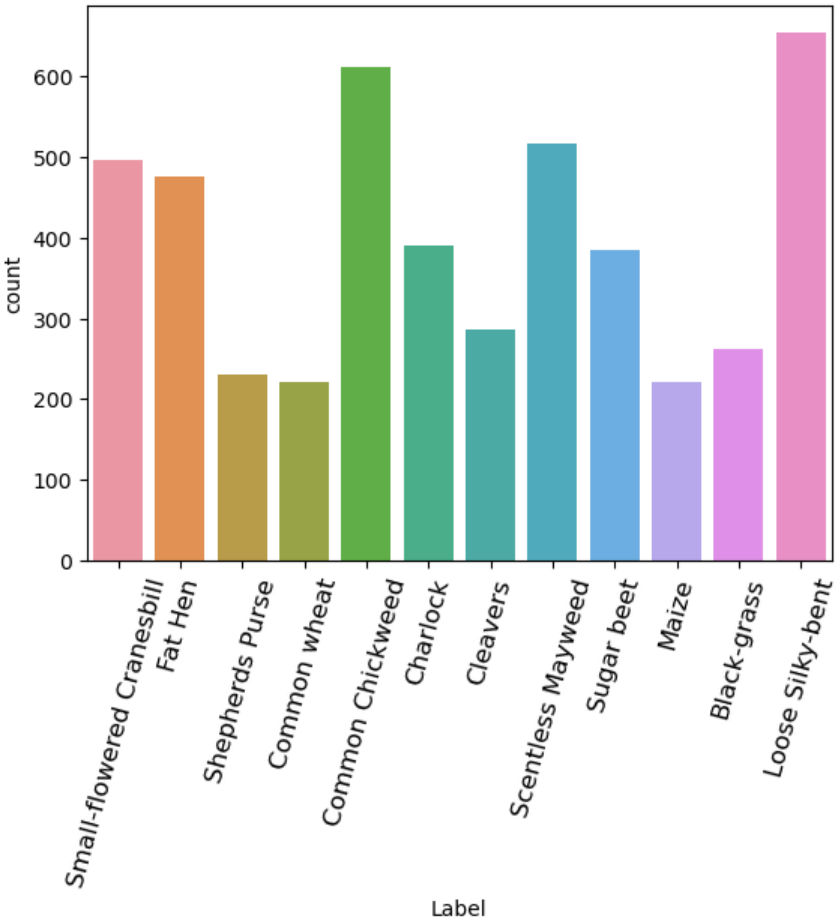


Observations:

Black grass, Loose Silky-bent & Common wheat are very similar shape of plants.

EDA Results

Checking the distribution of the target variable



Observations:

The dataset is imbalanced. Loose Silky-bent has approximately 650 images. Common wheat, Maize, Shepherds Purse & Black grass has approximately between 200 to 250 images.

Data Preprocessing

Converting the BGR images to RGB images.

Resizing images from 128 to 64

Data preparation for modeling: the data was split into train, validation and test.

The target label was encoded categorical features using LabelBinarizer. The data was Normalized using scaling

Model Performance Summary

Overview of model and its parameters:

Model 1: CNN was made with 128,64,32 filters and kernel size 3x3 , padding 'same' with ReLu activation. The CNN takes an input of 64,64,3 and max pooling is added to 3 CNN. Flatten is used. Two fully connected layers are added and Dropout of 0.3. For each transaction, the final layer will output 12 multi-class classification (softmax activation function) and classify 12 plant class. Since this is a multi-class classification problem, we will be minimizing the categorical_crossentropy and we can choose **Adam optimizer** with **accuracy** as the metric. This model is overfit. We have to try another architecture to get the better test **accuracy of 72%.**

Model 2: CNN was made with 64,32 filters and kernel size 3x3 , padding 'same' with ReLu activation. The CNN takes an input of 64,64,3 and max pooling is added to 2 CNN. Batch Normalization is added before Flatten is used. Two fully connected layers are added and Dropout of 0.3. For each transaction, the final layer will output 12 multi-class classification (softmax activation function) and classify 12 plant class. Since this is a multi-class classification problem, we will be minimizing the categorical_crossentropy and we can choose **Adam optimizer** with **accuracy** as the metric. Model Performance Improvement is used: data augmentation & ReduceLROnPlateau. Test a**ccuracy increased with 80%.**

Model Performance Summary

Overview of Model 1 and its parameters

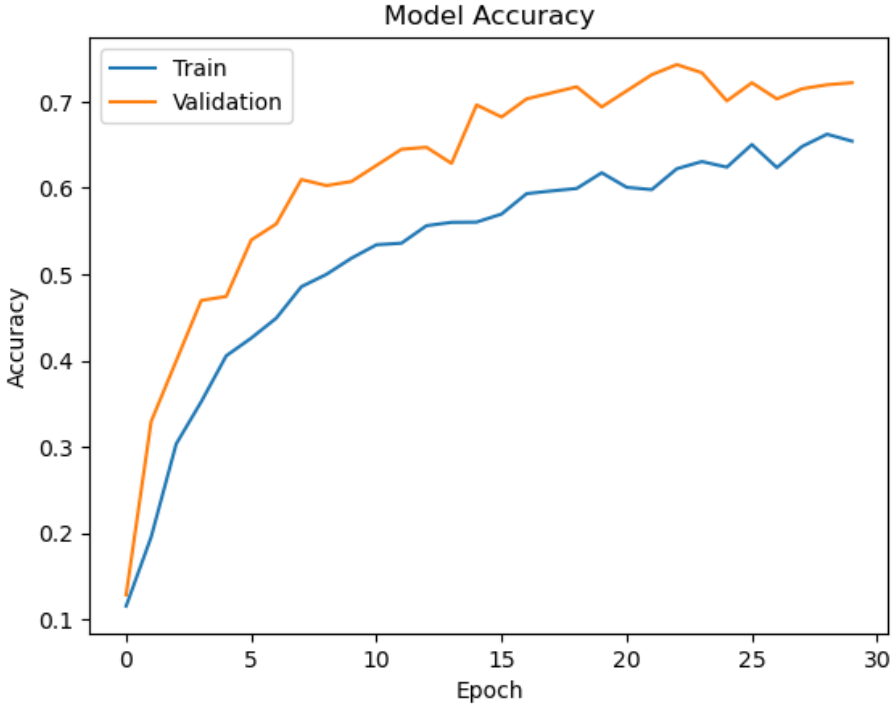
Model 1: "sequential"

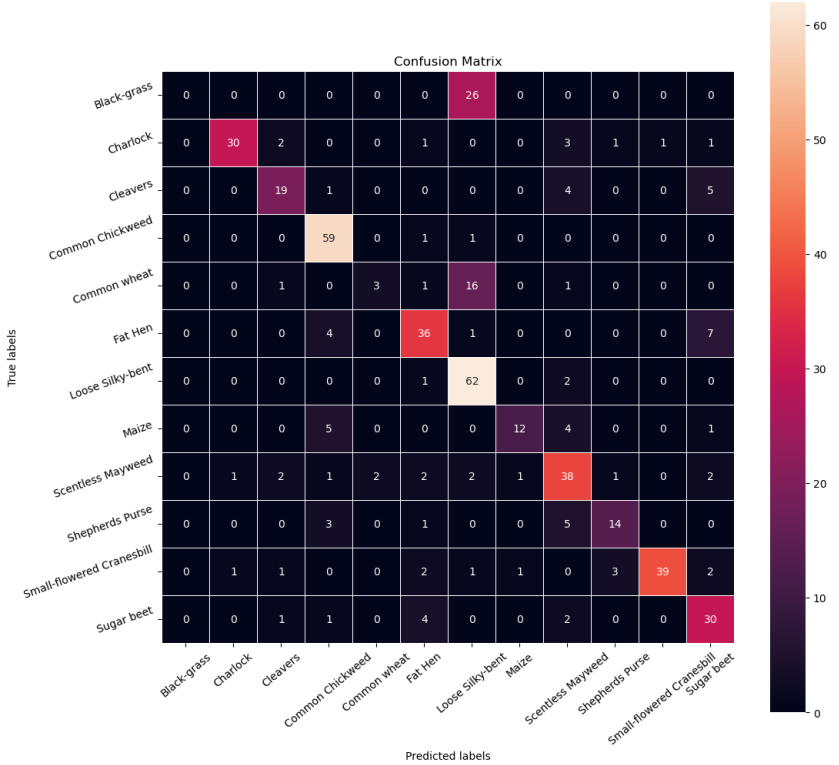
```
_________________________________________________________________
Layer (type) Output Shape Param #
=================================================
conv2d (Conv2D) (None, 64, 64, 128) 3584
max_pooling2d (MaxPooling2D) (None, 32, 32, 128) 0
conv2d_1 (Conv2D) (None, 32, 32, 64) 73792
max_pooling2d_1 (MaxPooling2D) (None, 16, 16, 64)
conv2d_2 (Conv2D) (None, 16, 16, 32) 18464
max_pooling2d_2 (MaxPooling2D) (None, 8, 8, 32) 0
flatten (Flatten) (None, 2048) 0
dense (Dense) (None, 16) 32784
dropout (Dropout) (None, 16) 0
dense_1 (Dense) (None, 12) 204
=================================================
Total params: 128828 (503.23 KB)
Trainable params: 128828 (503.23 KB)
Non-trainable params: 0 (0.00 Byte)
```

Observation: There are 128828 params.

Model Performance Summary

Overview of Model 1 and its parameters





Observation: The training accuracy of the model was good but the validation accuracy was not good. The model seems to overfit on the data. We have to try another architecture to get the better accuracy of 72%. We can also observe that classes black grass and common wheat are mostly misclassified.

Model Performance Summary

Overview of Model 2 and its parameters

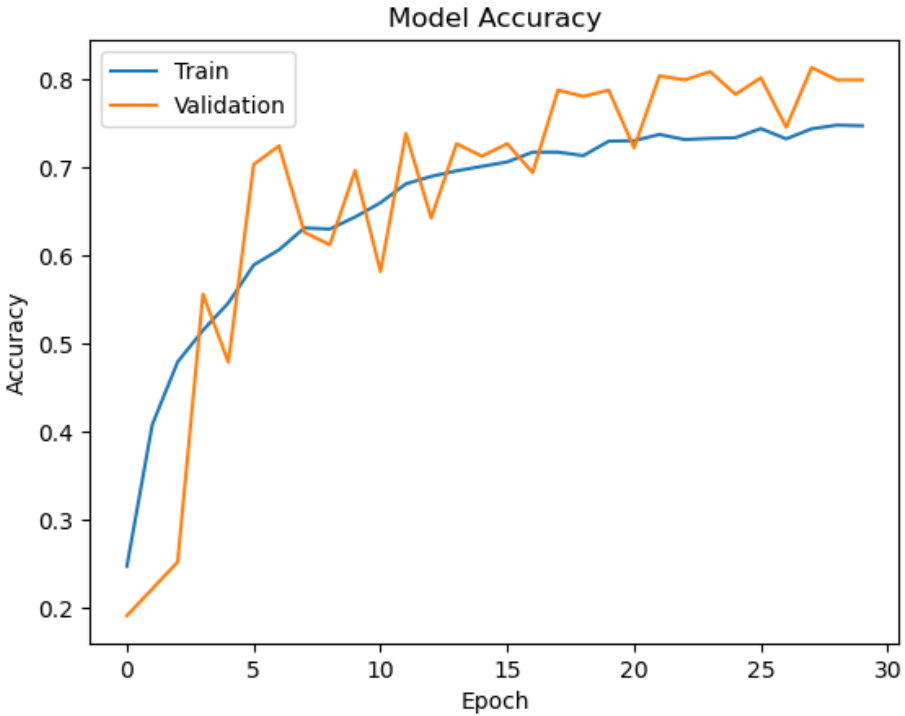
Model 2: "sequential"

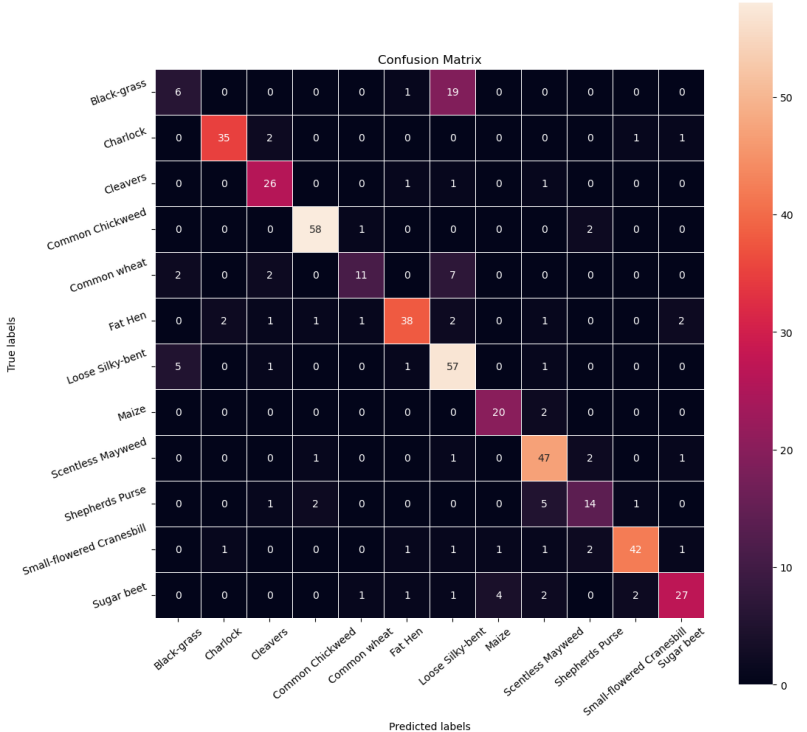
```
_________________________________________________________________
Layer (type) Output Shape Param #
=================================================================
conv2d (Conv2D) (None, 64, 64, 64) 1792
max_pooling2d (MaxPooling2D) (None, 32, 32, 64) 0
conv2d_1 (Conv2D) (None, 32, 32, 32) 18464
max_pooling2d_1 (MaxPooling2D) (None, 16, 16, 32) 0
batch_normalization(Batch Normalization)(None, 16, 16, 32) 128
flatten (Flatten) (None, 8192) 0
dense (Dense) (None, 16) 131088
dropout (Dropout) (None, 16) 0
dense_1 (Dense) (None, 12) 204
=================================================================
Total params: 151676 (592.48 KB)
Trainable params: 151612 (592.23 KB)
Non-trainable params: 64 (256.00 Byte)
```

Observation: There are 151676 parameters.

Model Performance Summary

Summary of the final model for prediction





Observation: The training accuracy and the validation accuracy of the model 2 was good. Data augmentation, Batch Normalization and Reducing the Learning Rate help improve the model's performance. The confusion matrix appears to be improving as well, however there is still some confusion with the black grass & Loose Silky bent classes of plant species.

Model Performance Summary

Summary of key performance metrics for test data in tabular format

Plant classification
0, 'Small-flowered Cranesbill'
1, 'Fat Hen'
2, 'Shepherds Purse’
3, 'Common wheat'
4, 'Common Chickweed'
5, 'Charlock'
6, 'Cleavers'
7, 'Scentless Mayweed'
8, 'Sugar beet'
9, 'Maize'
10, 'Black-grass'
11, 'Loose Silky-bent'

	precision	recall	f1-score	support
0	0.46	0.23	0.31	26
1	0.92	0.90	0.91	39
2	0.79	0.90	0.84	29
3	0.94	0.95	0.94	61
4	0.79	0.50	0.61	22
5	0.88	0.79	0.84	48
6	0.64	0.88	0.74	65
7	0.80	0.91	0.85	22
8	0.78	0.90	0.84	52
9	0.70	0.61	0.65	23
10	0.91	0.84	0.87	50
11	0.84	0.71	0.77	38
accuracy			0.80	475
macro avg	0.79	0.76	0.76	475
weighted avg	0.80	0.80	0.79	475

Observation:

Common wheat plant classification has the highest precision of 0.94 by the Final Model 2. Accuracy is 80%. The model 2 has a good generalization performance.

Model Performance Summary

Summary of key performance metrics for training and test data in tabular format for comparison

Models	Train Accuracy	Validation Accuracy	Test Accuracy
CNN Model 1	65%	72%	72%
CNN Model 2 with Data Augmentation	75%	70%	80%

Conclusions:

- Loose Silky-bent has approximately 650 images.
- The dataset is imbalanced. Loose Silky-bent has approximately 650 images. Common wheat, Maize, Shepherds purse & Black Grass has approximately between 200 & 250 images.
- There are 4750 images of shape 128 x 128 X 3, each image having 3 channels.
- The confusion matrix appears to be improving as well, however there is still some confusion with the black grass & Loose Silky bent classes of plant species.
- The Maize class is the least confused class among all.
- Model 2 was the best model because it predicted the majority of the classes better than the other models.
- The test accuracy of the model 2 is 80%.
- Data Augmentation, Batch Normalization and Reducing the Learning Rate has also helped in improving the CNN model.

APPENDIX

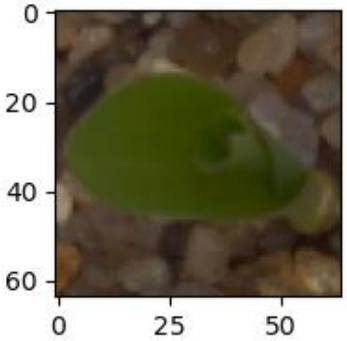
Data Background and Contents

The Aarhus University Signal Processing group, in collaboration with the University of Southern Denmark, has recently released a dataset containing images of unique plants belonging to 12 different species.

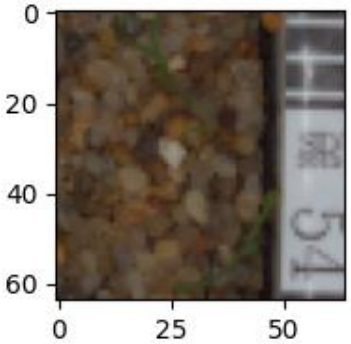
List of Plant species :
Black-grass
Charlock
Cleavers
Common Chickweed
Common Wheat
Fat Hen
Loose Silky-bent
Maize
Scentless Mayweed
Shepherds Purse
Small-flowered Cranesbill
Sugar beet

Visualizing the prediction of Final Model 2

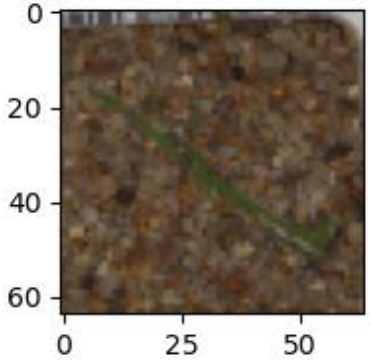
Visualizing the predicted and correct label of images from test data



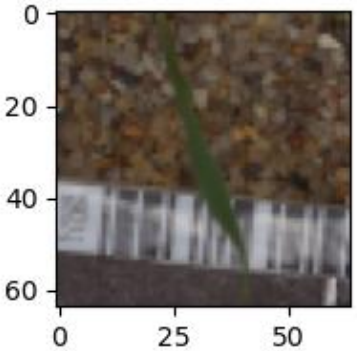
Predicted Label: ‘Maize'
True Label: Maize



Predicted Label: 'Loose Silky-bent’
True Label: Loose Silky-bent



Predicted Label: 'Loose Silky-bent’
True Label: Black-grass



Predicted Label: 'Common weed'
True Label: Common weed

Observation: Model 2 was the best model because it predicted the majority of the classes better than the other models with accuracy of 80%

Great Learning

Happy Learning !