Udacity Machine Learning Nanodegree 2020

Capstone Proposal:

Plant Pathology Identification

Author: Angel Peinado Perez, May 1st 2020.

1. DOMAIN BACKGROUND

This project is based on Kaggle competition "Plant Pathology 2020", for identification of the category of foliar diseases in apple trees (https://www.kaggle.com/c/plant-pathology-2020-fgvc7/overview). This competition is a challenge to distinguish between leaves which are healthy, those which are infected with apple rust, those that have apple scab, and those with more than one disease.

2. PROBLEM STATEMENT

Objectives of 'Plant Pathology Challenge' are to train a model using images of training dataset to:

- Accurately classify a given image from testing dataset into different diseased category or a healthy leaf;
- Accurately distinguish between many diseases, sometimes more than one on a single leaf;
- Deal with rare classes and novel symptoms;
- Address depth perception—angle, light, shade, physiological age of the leaf;
- Incorporate expert knowledge in identification, annotation, quantification, and guiding computer vision to search for relevant features during learning.

Misdiagnosis of the many diseases impacting agricultural crops can lead to misuse of chemicals leading to the emergence of resistant pathogen strains, increased input costs, and more outbreaks with significant economic loss and environmental impacts.

Current disease diagnosis based on human scouting is time-consuming and expensive, and although computer-vision based models have the promise to increase efficiency, the great variance in symptoms due to age of infected tissues, genetic variations, and light conditions within trees decreases the accuracy of detection.

3. DATASETS AND INPUTS

Files available for this competition are the following:

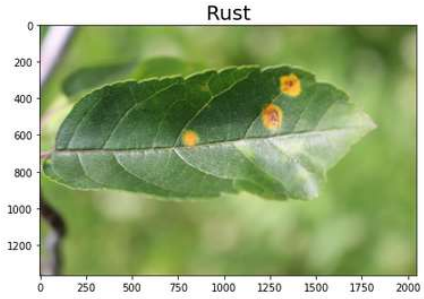
3.1 train.csv (CSV file with information of training images and columns with information about target diseases)

	image_id	healthy	multiple_diseases	rust	scab
0	Train_0	0	0	0	1
1	Train_1	0	1	0	0
2	Train_2	1	0	0	0
3	Train_3	0	0	1	0
4	Train_4	1	0	0	0

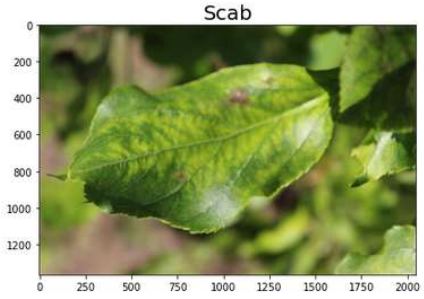
- image_id: name of image files for train
- combination of multiple diseases: 1 label if True
- healthy: 1 label if True
- rust: 1 label if True
- scab: 1 label if True

3.2 Explanation of Diseases

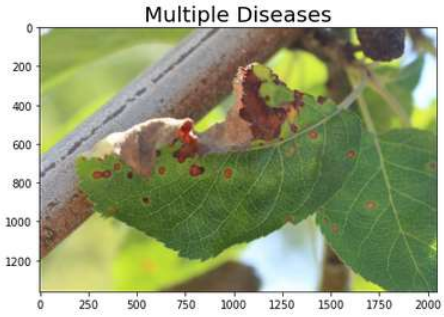
Rust: Common rust (*Phragmidium* spp.) is a fungal disease that attacks roses, hollyhocks, snapdragons, daylilies, beans, tomatoes and lawns. It is most often found on mature plants where symptoms appear primarily on the surfaces of lower leaves.



Scab: Caused by a fungus that infects both leaves and fruit. Scabby fruit are often unfit for eating. Infected leaves have olive green to brown spots. Leaves with many leaf spots turn yellow and fall off early.



Multiple disease: combinations of Scab and rust and/or other diseases



3.3 Images

A folder containing the train and test images (2048 x 1365 pix), in jpg format.

3.4 Test.csv

- image_id: name of image files for testing, without labels information because these are the object of prediction.

	image_id
0	Test_0
1	Test_1
2	Test_2
3	Test_3
4	Test_4

3.5 sample_submission.csv (Object submission file example)

- image_id: name of image files for testing
- combination of multiple diseases: probability of multiple diseases
- healthy: probability of healthy
- rust: probability of rust
- scab: probability of scab

	image_id	healthy	multiple_diseases	rust	scab
0	Test_0	0.25	0.25	0.25	0.25
1	Test_1	0.25	0.25	0.25	0.25
2	Test_2	0.25	0.25	0.25	0.25
3	Test_3	0.25	0.25	0.25	0.25
4	Test_4	0.25	0.25	0.25	0.25

Note: it is asked to submit probabilities of disease

4. SOLUTION STATEMENT

The proposed solution to this problem is the application of Deep Learning techniques that have been proved to be highly successful in the field of image classification.

Specifically, convolutional neural networks will be used with a stack of different kind of layers: convolutional, max pooling, dropout and finally dense layers. First a model from

scratch will be build, trained and tested to check the accuracy, and later transfer learning techniques will be used to improve the accuracy.

It will be used the evaluation metrics described in later section (mainly accuracy metric) to compare the performance of these solution against the benchmark models specified.

5. BENCHMARK MODEL

The benchmark model will be first a model from scratch will be build, trained and tested to check the accuracy.

Later transfer learning techniques will be used to improve the accuracy (i.e: densenet121, vgg16, etc).

6. EVALUATION METRICS

The evaluation metrics to be used will be mainly Accuracy Score: Number of correct predictions / Total number of predictions.

7. PROJECT DESIGN

This project will be developed using Google Colab services, with the GPU available.

Project will have the different stages:

7.1 Data reading and exploration

- Upload of competition datasets to Drive and preparation of roots for data reading.
- Data analysis and visualization: visualization of images and exploration of data statistics

7.2 Data preparation

- Preparation of data to be introduced to Pytorch datasets
- Preparation of Class objects for Pytorch datasets reading (train, validation and test)
- Definition of image transforms for datasets.

7.3 Building of models

- Building of scratch model using a combination of convolutional, max pooling, drop and dense layers.
- Training scratch model and later validation
- Testing of model and evaluation of metrics.
- Building model using transfer learning techniques (some models will be chosen: densenet, vgg, etc..).
- Training scratch model and later validation
- Testing of model and evaluation of metrics.
- Tuning of hyperparameters for metrics improvement.

7.4 Output

- Once the model has obtained enough accuracy: creation of csv file with the format required in the Kaggle competition.

8. REFERENCES

[1.] The Plant Pathology 2020 challenge dataset to classify foliar disease of apples. Ranjita Thapa (1), Noah Snavely (2), Serge Belongie (2), Awais Khan (1) ((1) Plant Pathology and Plant-Microbe Biology Section, Cornell University, Geneva, NY, (2) Cornell Tech). https://arxiv.org/abs/2004.11958

[2.] https://www.planetnatural.com/pest-problem-solver/plant-disease/common-rust/

[3.] https://www.kaggle.com/c/plant-pathology-2020-fgvc7/overview