Capstone project proposal

For Udacity Machine Learning Engineer Nanodegree

I want to participate in the *Plant Pathology 2020* challenge hosted on Kaggle. The following descriptions are based on the information shown on Kaggle.

Project domain background

Many plant diseases can be detected by visual inspection. An early detection is necessary to avoid further spread and avoid economic or environmental impact. As with other vision based challenges, deep learning seems to be a promising tool to increase the level of accuracy in detecting and classifying diseases while minimizing human efforts through automation.

Problem statement

Based on a training set size of only 1800 images of apple leaves a model needs to be developed, that can automatically detect if a leaf is healthy or not. Therefore the challenge tackles a supervised learning problem.

Dataset and inputs

The dataset is provided by the Kaggle competition and consists of images of apple leaves as well as one of the following classes: "healthy" (ca. 30%), "rust" (ca. 35%), "scab" (ca. 30%) or "multiple diseases" (ca. 5%). In total we have ca. 1800 training images and 1800 test images. The images look like high quality photographs, where the leaves of interest in focus.

Solution statement

I want to use start exploring the data in Jupyer notebooks and then use Sagemaker for the machine learning workloads. I will also try to build a small web frontend to upload the data and send it to the backend for prediction.

In detail, I plan to perform various preprocessing steps such as loading the data, performing some simple resizing/ cropping, eventually apply contrast normalization and then store the dataset in S3.

After that I want to train a convolutional neural network to predict the desired output classes. I'm not sure yet about the model architecture, but plan to try out several ones such as VGG-16 or ResNet. I will try to fine tune a pre-trained model due to the lack of available training data. For training, I will augment the data during runtime.

Benchmark model

As a benchmark I will one the one side use the Kaggle leaderboard but also train a simple benchmark model on my own. For example I could transform each image to a one-dimensional vector, transform it through a PCA and then feed the result into a Random Forest Model. The deep learning model should as a minimal requirement perform better than this easy model.

Evaluation metrics

As evaluation metric this competition uses the AUROC per output column which will then be averaged to a final score. AUROC means "area under the receiver operator curve". The receiver operator curve (ROC) has the false positive rate (FPR) on the x-axis and the ture positive rate (TPR) on the y-axis. For each potential decision threshold of the classifier from 0 to 1 the values for the FPR and TRP are calculated. To receive the AUROC we simply calculate the area under this curve.

Outline of the project design

Since the data is already collected, I plan do directly jump into the data exploration phase. After understanding the data incl. the underlying distribution of the categories I will start to evaluate possible machine learning approaches. In general I would like to use deep learning techniques but I'm not sure yet, if the amount of training data makes this a feasible task. Probably transfer learning or finding additional training data can help here, otherwise I will opt for a simpler approach such as dictionary learning together with a standard classifier such as random forest. After finding a model that was successful on the training data I will validate the results on Kaggle's leaderboard and start a new interation of training based on the outcome.

Sources: Photo by Paweł Czerwiński on Unsplash Competition information: Plant Pathology 2020 FGVC7