

Machine Learning Engineer Nanodegree

Capstone Proposal

Deep learning for Plant Seedlings Classification

Domain Background

Weed control, which directly affect the efficiency of agricultural, has been researched for several decades. Keeping crop free of weeds is critical in good farm management. Farmers should plan their farm activities well so that they do weeding at the right time and in the right way. For this purpose, to differentiate weed from crop seedling is the essential step. Different seedlings have their unique features on their colors, shape, and outlines. Multiple technologies have been adopted including GPS mapping, ground-based vision identification, remote sensors and so on. This project focuses on to identify and classify the seedlings from the image using machine learning based computer vision. Fortunately Aarhus University and University of Southern Denmark recently released a dataset containing images of approximately 960 unique plants belonging to 12 species at several growth stages. those labeled data provides the researchers an opportunity to experiment with different image recognition techniques(for instance, K-nearest, SVM, CNN, and other high performance deep learning models). Kaggle also helped to organize an online competition providing a platform to let researchers communicate and cross-compare the solutions.

Citation

PAPER: A Public Image Database for Benchmark of Plant Seedling Classification Algorithms
Kaggle Competition: <https://www.kaggle.com/c/plant-seedlings-classification>
(<https://www.kaggle.com/c/plant-seedlings-classification>)

Problem Statement

The topic of this project is to use provided dataset training weed recognition algorithms. The most critical part is to select and adjust the architecture of the model together with its hyper-parameters to push the limits of the accuracy. To get inspiration from previous research results, transfer learning becomes a good start point. The dataset from AU & SDU consists of different seedling images and labeled classification will be used for training and testing separately. The performance will be evaluated on MeanFScore, which at Kaggle is a micro-averaged F1-score.

Datasets and Inputs

Datasets are divided by a training set and a test set of images of plant seedlings at various stages of growth. It comprises annotated RGB images with a physical resolution of roughly 10 pixels per mm. Each image has a filename that is its unique id. The dataset comprises 12 plant species, including: (Samples of each species from the database <https://vision.eng.au.dk/plant-seedlings-dataset/> (<https://vision.eng.au.dk/plant-seedlings-dataset/>)).



Maize



Common wheat



Sugar beet



Scentless Mayweed



Chickweed



Shepherd's Purse



Cleavers



Charlock



Fat Hen



Cranesbill



Black-grass



Loose Silky-bent

File descriptions

train.csv - the training set, with plant species organized by folder

test.csv - the test set, need to predict the species of each image

sample_submission.csv - a sample submission file in the correct format

Solution Statement

The goal of the project is to train a model/classifier that can recognize the seedlings from the image input. Considering the type of the input as the image, Convolutional Neural Networks was considered as a good option to start given its abilities on visual applications. Different architectures of CNN with transfer learning will be evaluated and tuned for the use case. ImageNet database and its relevant projects provide multiple high potential models worthy to take a test. Data visualization method like tensorboard will be used to monitor the performance during the training. The model will be measured using MeanFScore.

Benchmark Model

Kaggle presents a leaderboard (<https://www.kaggle.com/c/plant-seedlings-classification/leaderboard> (<https://www.kaggle.com/c/plant-seedlings-classification/leaderboard>)) to compare the performance among different solutions. Up til know the top 50s have reached an excellent score (>0.98110). The best score is 0.99622. Several kernels (SimpleNet, Xception network, VGG based classifier and so on) have also been published on the kaggle platform.

Evaluation Metrics

Solution is evaluated on MeanFScore (micro-averaged F1-score).

Given positive/negative rates for each class k , the resulting score is computed this way:

$$Precision_{micro} = \frac{\sum_{k \in C} TP_k}{\sum_{k \in C} TP_k + FP_k}$$
$$Recall_{micro} = \frac{\sum_{k \in C} TP_k}{\sum_{k \in C} TP_k + FN_k}$$

F1-score is the harmonic mean of precision and recall

$$MeanFScore = F1_{micro} = \frac{2Precision_{micro}Recall_{micro}}{Precision_{micro} + Recall_{micro}}$$

Project Design

Overview:

A theoretical workflow for this project will be

- Data prepare
- Iteration of:
 - Model training and validation
 - Parameter tuning
- Model test

Data Prepare

The first step is to get available datasets. The dataset will be downloaded from kaggle including a test set and a train set (<https://www.kaggle.com/c/plant-seedlings-classification/data> (<https://www.kaggle.com/c/plant-seedlings-classification/data>)). Data will also be split for validation. Tools like numpy, pandas will be used.

Model training and validation

More and more computer vision challenges have been successfully launched. There are a few of architectures published showing a really good performance on image recognition. Keras has pretrained models of those architectures including Xception, VGG, Inception, etc. It will help me to quickly test different models and evaluate their performances. In this step, tools like tensorflow and keras will be used.

Parameter tuning

Adjust hyperparameters over the course of training runs to achieve an optimal result. Considering the requirements of large computing power, cloud computing service(AWS EC2) may be adopted to accelerate the hyperparameter tuning iteration.

Model test

At last, the model will be test with an all-new test data. The evaluation metrics (MeanFScore) will be calculated by comparing with the ground truth.