Machine Learning Engineer Nanodegree

Project Capstone Proposal

Alex Langer, 3rd February 2017

Domain Background

Agriculture and horticulture has seen various applications of image recognition and machine learning. Rice seeds have been classified with Principal Component Analysis (You et al. 2009) and leafs have been classified towards particular species (Wu et al. 2007, Mallah et al. , 2013) by deploying probabalistic neural networks. When mastered, this can lead into various applications such as population control, desease detection[1] or weed elimination[2]. Mastery of these technologies will be crucial to meet rising requirements towards efficient agricultural production and population growth. Due to the sheer variety of plant species, approximately half a million, classification is prone to errors. I will attempt to approach this classification challenge by deploying a deep neural network against images of leaves that belong to one of 99 species.

Problem Statement

Classifying images of leaves as a particular species is a supervised machine learning task (classification) that will require to transform image data into multi-dimensional arrays. As all image data has been labeled with a plant species type, this will be essential to train a deep neural network with dedicated weights and biases for each plant species type. Since we feed more than thousand images into our network, the model should gain some robustness and the capability to reproduce prediction when fed new leaf images. I hope to obtain prediction probabilities above 90% while measuring accuracy will be done by comparing predicted labels (e.g. top 3 predictions) vs. actual image labels.

Datasets and Inputs

The underlying dataset for this exercise has been collected by James Cope, Thibaut Beghin, Paolo Remagnino, & Sarah Barman of the Royal Botanic Gardens, Kew, UK. It was made available through the UCI Machine learning repository and Kaggle (https://www.kaggle.com/c/leaf-classification)
The dataset is made up of 1,584 images of leaf specimens. There are 99 different specimen with 16 samples each. For preprocessing, these images have already been converted to binary black (for the leaves) against white background color. Furthermore, three sets of features are provided per image: a shape contiguous descriptor, an interior texture histogram, and a fine-scale margin histogram. The images are of different dimensions which will require additional preprocessing and shaping in order to feed a n-dimensional tensor with fixed input/output dimensions.

1 Http://peat.technology
2 http://www.bluerivert.com

A test and a training set of image data will be maintained to evaluate model performance.
Given that the number of input images is still rather low, data augmentation techniques may be deployed.

Solution Statement

After performing further preprocessing on the images to feed the network with same size arrays, their pixel data will be transformed to an n-shaped array in order to feed it into a neural network for training. This may be a deep neural network with at least 3 layers. While Xinshao et al (Weed Seems Classification Based on PCANet Deep Learning) utilized PCA analysis to solve this task, I will seek to deploy a Neural network architecture comparable to Goodfellow (Multi Digit Number Recognition, 2014) in order to obtain improved accuracies. The model performance will then be tested by utilzing a testing set on the model and comparing predictions against actual labels.

Benchmark Model

The CIFAR-10 Image set has seen comparable objectives with much more complex image data. The task was to label significantly different objects. Accuracies of > 95% have been obtained with neural networks ranging from 3 to 12 layers. The ambition is to obtain comparable results with 3 to 5 layers given that we have more simple image data as inputs.

Evaluation Metrics

Two different datasets will be maintained: A training and a testing set. The neural network will be trained with the first dataset and then tested against its label prediction for the testing set and the actual labels. We will thus derive a figure for accuray. In order to get a better sense of prediction accuracy, the top 3 predictions per image may also be mapped for a sample group.

Project Design

The first step is to apply image preprocessing where Image sizes will have to be converted to similar size. For computation lightness, pictures have already been converted to binary colors. The data is already labeled but might has to be normalized to reduce standard deviations and enable gradient descent training on a more efficient level.
After completion of preprocessing, a neural network with a minimum three layer setup will be trained so that weights and biases representing the type of leaf and the class per image can be distinguished. The model will be fed with image data in batches to reduce memory usage.