Machine Learning Engineer Nanodegree

Capstone Proposal

Thi Nguyen

January 31st, 2019

Predicting Physiological Traits from Hyperspectral Reflectance Measurements Using Deep Learning

Domain Background

As global population is estimated to reach 9.7 billions by 2050 [1], the projected demand for cereal grain is far exceeding the current agricultural output [2]. In order to meet the projected global food demand, the world-wide crop production is required to be double [3]. The efficient use of physiological traits to raise wheat yield potential is the major target for agricultural researchers.

It is desirable to determine the amount of yield for wheat plants early on during the growth of the plants, instead of waiting for the end. A way of predicting the potential yield of plants is to look at the current biochemical and physiological traits of the plants. Measuring Photosynthesis-related traits, such as nitrogen per unit leaf area (N_{area}) and leaf dry mass per area (LMA), require laborious, destructive, laboratory-based methods, while physiological traits underpinning photosynthetic capacity, such as maximum Rubisco activity normalized to 25 °C (V_{cmax25}) and electron transport rate (J), require time-consuming gas exchange measurements.

The project aims to replace the traditional time-consuming laboratory-based methods by a fast and high-throughput deep learning model by using leaf-level hyperspectral reflectance parameters to predict the physiological traits.

Problem Statement

The aim of this project is take simple leaf reflectance measurements, which is data is much easier to collect, and then predicting the biochemical and physiological traits. We aim to develop a deep learning model to assess whether hyperspectral reflectance (350–2500 nm) can be used to rapidly estimate these traits on intact wheat leaves. The proposed model is using gas exchange and hyperspectral reflectance data from 76 genotypes grown in glasshouses with different nitrogen levels and/or in the field under yield potential conditions.

Datasets and Inputs

Datasets: The dataset consists 1185 records supplied by the Center of Excellence in Plant Energy Biology at the Australian National University (ANU). The data is collected by Aus 1, Aus 2, Aus 3 and Mex 1 experiments from two different geographical locations (Mexico and Australia).

Input data:
Hyperspectral reflectance data is for each leaf image. At the raw level there is hyperspectral reflectance curve for each pixel. The data is captured by a FieldSpec®3 (Analytical Spectral Devices, Boulder, CO, USA). Basically, the intensity of the reflected light at different wavelengths. The range of wavelengths measured is between 350 and 2500 nm.

Figure 1 shows what the hyperspectral reflectance images look like for the V_{cmax25} trait, the bold line is the mean and the range is given by the upper and lower lines [4]:

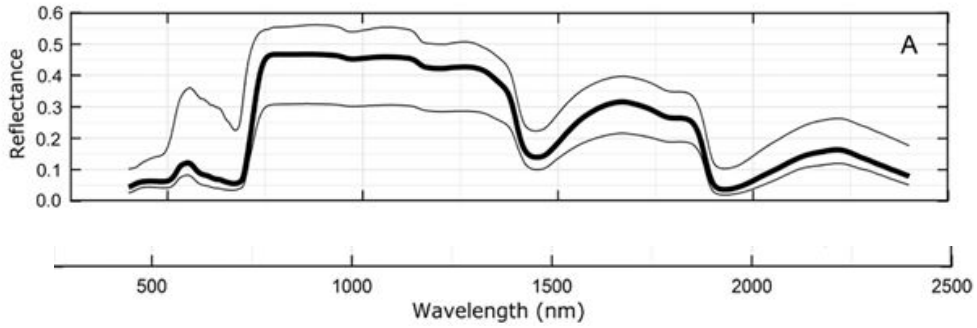

Fig 1.:

Output: 11 biochemical and physiological Traits: LMA, Narea, SPAD, Nmass, Parea, Pmass, V_{cmax}, V_{cmax25}, J (ETR), Photo, and Cond

Solution Statement

I propose to use 1D convolutional neural networks (CNNs) to map the hyperspectral reflectance data into the biochemical and physiological traits. I will try to build two different approaches: a single multiclass model for all traits and different models for different traits.

Benchmark Model

The existing method [4] uses least square regression to map input to output. Depending on which traits are being predicted the R^2 range between 0.42 and 0.93. However, they used a data set of just 200 examples. 100 used for training and 100 used for testing. We will be given data set with thousands of examples. It will come from at least two different geographical locations (Mexico and Australia).

Evaluation Metrics

We evaluate the performance of the benchmark model and the proposed model using the coefficient of determination (R^2) and the relative error of prediction (REP) [2].

The R^2, the model bias is defined as:

$$\text{Bias } (\%) = 100 \times (\bar{\hat{y}} - \bar{y}) / \bar{y} \qquad (1)$$

to represent the percentage of the difference between the mean of the predicted trait, $\bar{\hat{y}}$, and the mean of the observed trait, $\bar{y}$.

The relative error of prediction (REP) is define as:

$$\text{REP } (\%) = 100 \times \left[\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \right]^{0.5} / \bar{y} \qquad (2)$$

to represent the percentage of the root mean square error in prediction, where y_i and $\hat{y}_i$ are observed and predicted traits, n is the number of sample in data set and $\bar{y}$ is the mean of the observed values of traits.

Project Design

Before start building the CNN models, I first will split the data into three different sets: training set, validation set and testing set. Then I will perform data exploration and visualisation. Next, I will build different CNN architectures from very shallow fully connected networks to very deep networks from scratch, and from a multiclass model for all traits to different models for different traits. Hyperparameter tuning then is applied to build the best model. Final step is model evaluation, I plan to compare the proposed models with the existing least square regression model proposed by [4].

References

[1] UN Department of Economic and Social Affairs. 2015. World population prospects. The 2015 revision. Key findings and advance tables. New York: United Nations Department of Economic and Social Affairs, Population Division.

[2] Nguyen HT, Lee B-W. 2006. Assessment of rice leaf growth and nitrogen status by hyperspectral canopy reflectance and partial least square regression. European Journal of Agronomy 24, 349–356.

[3] Tilman D, Balzer C, Hill J, Befort BL (2011) Global food demand and the sustainable intensification of agriculture. Proc Natl Acad Sci USA 108: 20260–20264

[4] Viridiana Silva-Perez, Gemma Molero, et. al., Hyperspectral reflectance as a tool to measure biochemical and physiological traits in wheat, Journal of Experimental Botany, 2017