Satellite image analysis with Machine Learning.

Daniel Atilano

***Abstract*— Among the largest coffee producers in Latin America is Mexico, with an average figure of up to 3.4 million 60-kilogram bags production during the last five years. States like Chiapas, Oaxaca, Veracruz, and Puebla amount to 90 % of total coffee production. However, in this study the developing of satellite image database was carried through in order to show new potential areas for coffee plantation; for both the leading variant bean in the market *arabica*, as well as the the emerging *Robusta* coffee bean. Furthermore, classification attempts with Linear Regression, and CNN proved to be fairly successful for the separation of potential fertile lands from infertile ones. *Index terms*—Coffee production, satellite imagery, Sentinel2 imagery, fertile land for coffee harvest, machine learning algorithms.**

I. Data Set

Satellite imagery makes up one of the most important tools for what is called 'Precision farming'; a term coined in the 90's that refers to more sophisticated methods for crops management [1]. With the never ending toil and trouble of plague infestations, energy monitoring, and fertilizer consumption (whose increasing price range has further destabilized production), this report looks to do further research into satellite images in order to make efficient farming easier.

First, a brief summary into the data used for the machine learning models is in order. In between the satellites that are best suited for crop monitoring is the Sentinel-2 EU mission. This satellite is able to capture light frequencies within a range from 443 nm to 2190 nm distributed with 13 bands [2]. The bands pertinent for analysis are; thus, B8 with B11; responsible for infrared wavelengths, and B2; blue wavelength. The objective is to find regions where infrared light is absorbed so as to reflect back green wavelengths, the more green is reflected the healthier vegetation or crop harvests are [3]. Additionally, the data takes into consideration 5 years of data with coffee harvest dates; from September to March [4], in Mexico.

To get a hold of these images, though, and be able to automatize the process a python script was made to make use of Google's API earth engine. After access has been granted with authorized credentials to the GCP(Google Cloud Platform) and the API, the provided function methods coupled with the geemap library [5] made it possible to mass produce satellite images; albeit with some limitations. The script needs the region of interest (ROI) coordinates to work. To label the data ROI with and without healthy vegetation were selected to provide the images shown in figure 1 and 2.

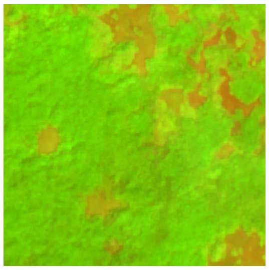


Fig. 1. Fertile region

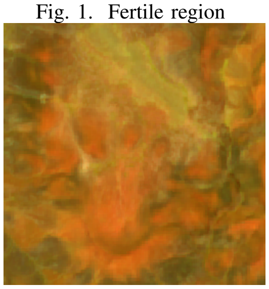


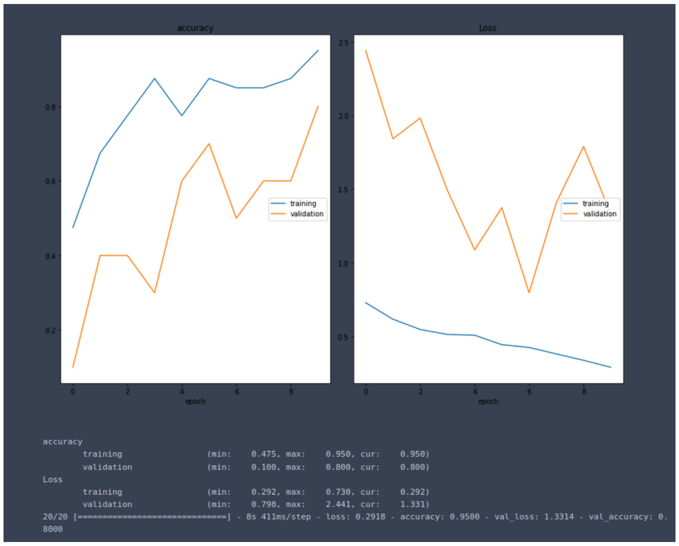
Fig. 2. Non-Fertile region

II. ML models

The CNN model to classify these images uses tensorflow with keras framework that runs in a jupyter notebook with python. The 300 available images are split three-ways into validation, train, and test folders with two class sub-directories each. The CNN uses transfer learning with VGG16 layers with two 128 relu dense layers and a final softmax output layer. When running the CNN, figure 2 shows visible changes of validation loss and training accuracy by the stopping in between 10 epochs.

After the model ran 100 epochs, for each 8-fold , the cost function in figure 6 shows folds that the cost increased in some folds whilst it decreased in others. Finally, the cost converges in the 0.54 y-axis. The average score was also calculated from the scores of all folds.

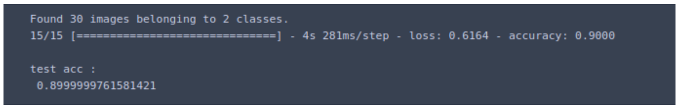
[1] Departamento de Computación y Mecatrónica, Tecnológico de Monterrey Campus Querétaro, Epigmenio Gonzalez #500 76130 Querétaro, México.
Asesores: Arturo Pérez, Alfonso Gómez, Pedro Pérez.
Clave: TC3007 - TC3054
Unidades: 12 unidades
Oficina: Edificio 2, 3er piso.
`bvaldesa at itesm.mx`



```
accuracy
	training         	 (min:    0.475, max:    0.950, cur:    0.950)
	validation       	 (min:    0.100, max:    0.800, cur:    0.800)
Loss
	training         	 (min:    0.292, max:    0.730, cur:    0.292)
	validation       	 (min:    0.798, max:    2.441, cur:    1.331)
20/20 [==============================] - 8s 411ms/step - loss: 0.2918 - accuracy: 0.9500 - val_loss: 1.3314 - val_accuracy: 0.
8000
```

Fig. 3.

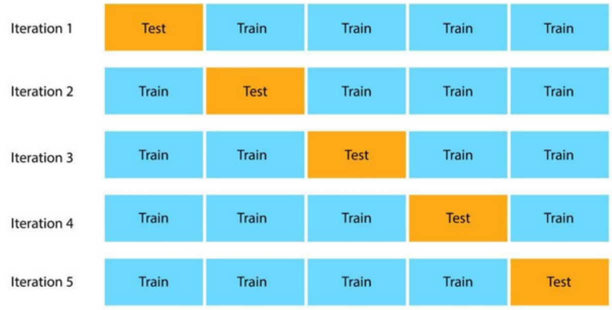
At the early stages of training, the CNN shows the model does not overfit accuracy. In the loss error plot, however, there are signs of overfit maybe due to the small size of the dataset and how susceptible it is to outliers. The accuracy score is shown in Figure 4, as the result when the model runs with a sample of 100 unseen images. The accuracy, though, ranges from 83 % to 90 %.



```
Found 30 images belonging to 2 classes.
15/15 [==============================] - 4s 281ms/step - loss: 0.6164 - accuracy: 0.9000

test acc :
 0.8999999761581421
```

Fig. 4.

The logistic regression model by hand has: an hypothesis function, a sigmoid activation function, a gradient descent function, a predict function, and a final 'model' function that combines them all together. Given the size of the train dataset the K fold technique was used to make the most out of it; the 300 images were split into 8 folds, where the validation fold has 38 images and K-1 train sets have 262 images.



Iteration 1	Test	Train	Train	Train	Train
Iteration 2	Train	Test	Train	Train	Train
Iteration 3	Train	Train	Test	Train	Train
Iteration 4	Train	Train	Train	Test	Train
Iteration 5	Train	Train	Train	Train	Test

Fig. 5.

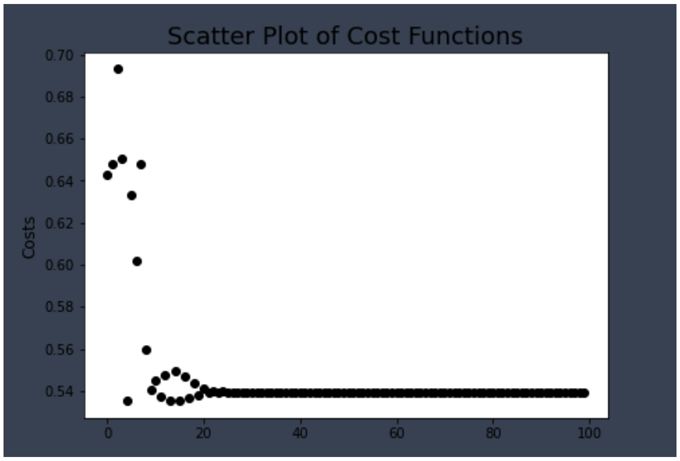
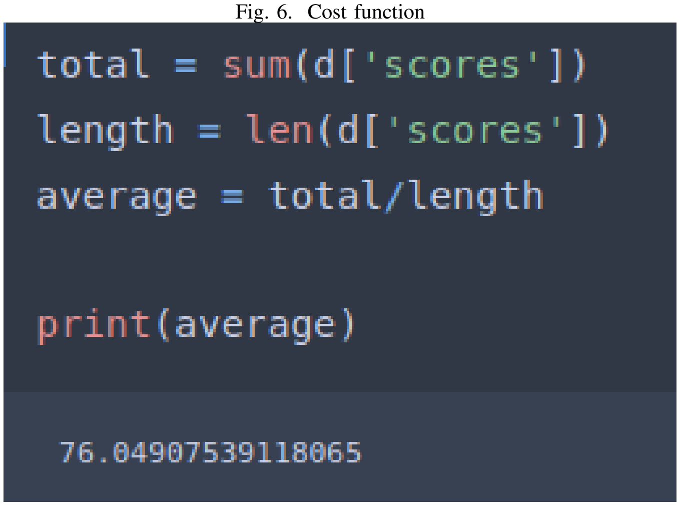


Fig. 6. Cost function

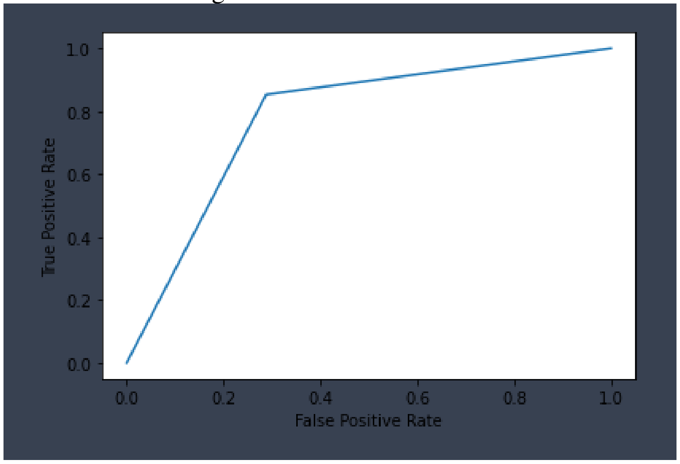


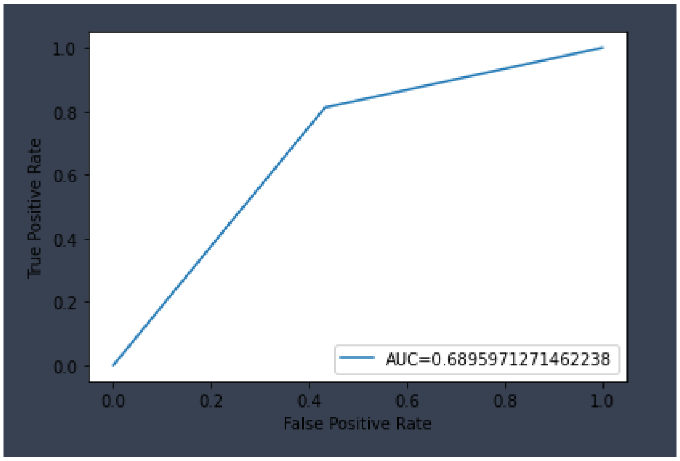
```
total = sum(d['scores'])
length = len(d['scores'])
average = total/length


print(average)
```

```
76.04907539118065
```

Fig. 7. score average

All predictions from both validation set and training set were stored into an array so as to graph two ROC curves, and thus check for overfitting.





There is a significant sign for overfitting probably due to either the images are not readily distinguishable or tests with more validation data are required.

III. Expected contribution

The finding of ideal harvest areas for *arabica* coffee would prompt the potential introduction of new agricultural zones, and reduce costs involved when investing in fertilizers. This report classifies images based solely on agricultural zones (filtered by coffee harvest dates) however, further analysis remains to be done to account for variable more specific to coffee plantation. To determine where not only coffee but healthy vegetation flourishes is the first step.

References

[1] AFP. (January 28, 2021). This is how farmers are using satellites to enhance adaptation. May 23, 2022, from Global Center on Adaptaion https://gca.org/this-is-how-farmers-are-using-satellites-to-enhance-adaptation-2/

[2] SatAgro. Agriculture precisely for you. May 23, 2022, from SatAgro Sp. z o.o. https://www.satagro.eu/?lang=en#pytania

[3] (October 29, 2021) GisGeography, "Sentinel 2 Bands and Combinations"May 23, 2022, from GisGeography. https://gisgeography.com/sentinel-2-bands-combinations/

[4] Mercanta https://coffeehunter.com/knowledge-centre/coffee-seasonality/

[5] GEEMAP https://www.sciencedirect.com/science/article/pii/S0303243421002889?-via %3Dihub