Machine Learning Engineer Nanodegree

Capstone Proposal

Francisco Proboste
July 26, 2019

Proposal

Domain Background

This Project is framed in the field of Environmental Sciences and Conservancy. In particular, it is part of the “remote sensing approaches” to monitor nature dynamics. These techniques have been developed for 40 years, mainly in the context of forestry planning. The main source of information in this field comes from satellite images in the visible spectrum and others. At the beginning, experts performed most of the analysis manually, but nowadays image recognition techniques provide unprecedented possibilities for massive analysis of the ecosystem.

Problem Statement

In this capstone project I want to tackle a very particular problem in the forest remote sensing field. Nowadays we are facing massive ecosystem degradation over all the earth, one of these ecosystems are forests. Our cities, food industry and other extractive human activities are destroying large extensions of boreal, template and tropical forests. These forests sustain most of the terrestrial life diversity and also provide oxygen and other environmental services to the entire earth ecosystem. A lot has been talked about where forests are being cut down (Indonesian due to Oil Palm, Amazonian due to cattle rising, etc), however we don’t know much about places where the forest is growing. Knowing about places that are actually recovering could give us clues about how to recover other places, and how should society behave in order to sustain the life diversity on earth. So, I think a nice contribution would help developing technics to identify those places on earth were forests are thriving again.

Datasets and Inputs

There is many different satellite observations of the earth and its forests that are open to the public. In this case we will use images containing information of the visible spectrum. That makes it very democratic and easy to be interpreted, as we stated that one of the goals of this tool is to get more insight about places were forests are recovering. Google Earth Time-lapse shows clearly one of the many sources of satellite images of nature over time (https://earthengine.google.com/timelapse/). Another source of a huge amount of satellite images is tha NASA repository “GLOBE Observer” (https://observer.globe.gov/about). This is a part of a greater organization called the GLOBE Program which is an international program that enable the collaboration in the understanding of the Earth system and the global environment. We will start working with the data of the Sentinel-3 OLCI EFR, that provide full resolution observations (https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S3_OLCI). Time lapses of fixed size images of specific forests will be selected to train the model, and then the model will be tested on a broader set of images. In case we don’t find forests that have recovered over time, we will invert the time in well known degraded forests.

Solution Statement

Using open source satellite visible images, I propose to develop a service that is able to recognize if a forest has increased over a time series. This service will use deep learning and transfer learning to train an image classification tool that will quantify the amount of forest of a picture, and then will analyze if that indicator of amount has increased over time.

Benchmark Model

A benchmark model to compare our forest recovery finder will be using a vegetation index build over information of the NOAA-AVHRR radiometer. Despite I didn’t find a tool that uses a time-lapse of these radiometer images to detect forest recovery, this dataset could be a good benchmark for the first part of our model: the forest index that we will obtain from visual spectrum images. (Repository: https://developers.google.com/earth-engine/datasets/catalog/MODIS_006_MYD13A1)

Another benchmark option is https://vegmachine.net/#. This website provides a web app that allow to do GIS ground analysis including a “green ground” time-lapse.

Evaluation Metrics

The evaluation metric will be the area under the ROC Curve. We will measure the different model tunings of our tool, and compare them with this index represented by the area. ROC curve is useful in this case because our model is mainly a classification one. Given a time-lapse of satellite images, it will have to identify if the forest have recovered or not. That has to be performed intensively over a large region of the earth. In principle, false positives are not more important than false negatives, so I think this “Area under the ROC Curve” will be a very good general approach to measure the performance of the different models.

Project Design

As we can see in Figure 1, I will temporarily split the work of the project in 6 weeks. The first one will be about preparing the ground for the models construction: first the satellite images will be retrieved from the image repositories cited above. Those images will be labeled with some metadata to be defined. In parallel I will select a definitive deep/transfer learning architecture for the image classification and then for the time series analysis.

In the second week the images will be resized to a given fixed final size able which will be the basic frame where we will calculate the forest index and where later we will classify if the forest has recovered or not.

Weeks 3 and 4 will be all about the model tuning, both the forest index predictor and then the forest recovery classifier.

The final week will be about benchmark comparison and wrapping up all the work in a usable forest recovery service, able to be tested by Udacity instructors/evaluators.

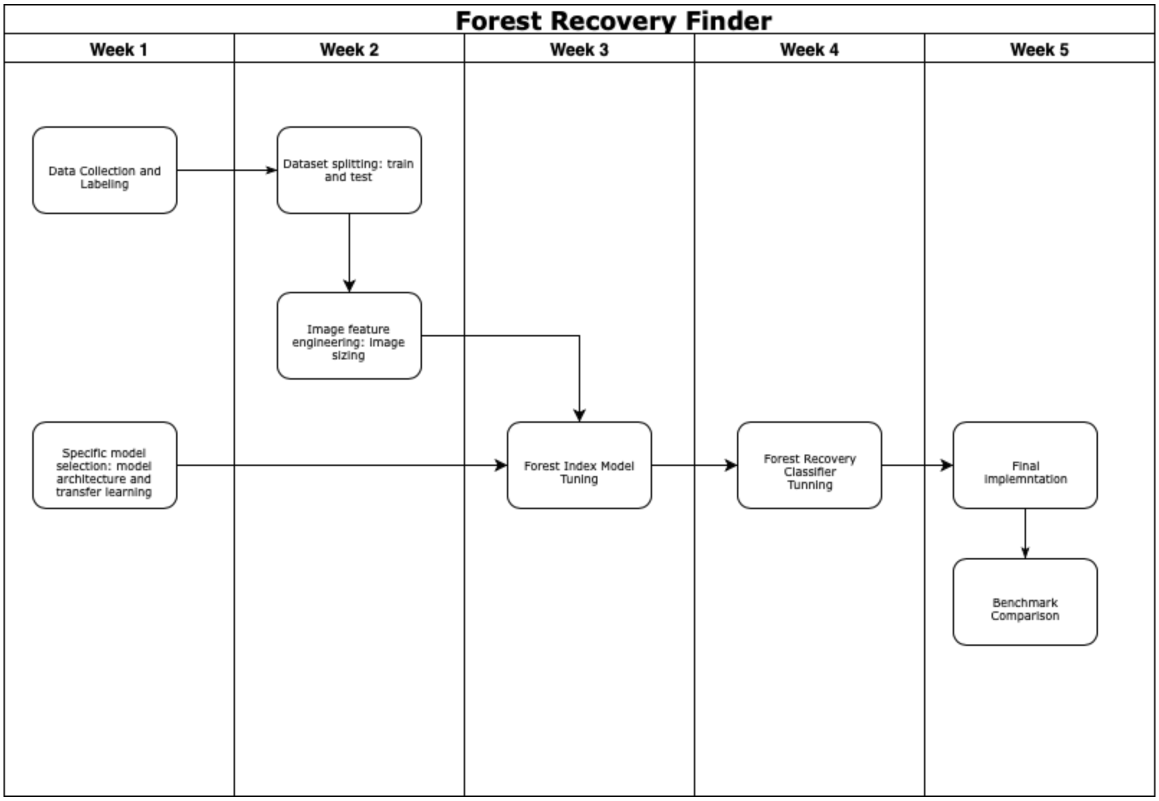
Forest Recovery Finder
Week 1
Week 2
Week 3
Week 4
Week 5
Data Collection and Labeling
Dataset splitting: train and test
Image feature engineering: image sizing
Specific model selection: model architecture and transfer learning
Forest Index Model Tuning
Forest Recovery Classifier Tunning
Final Implemntation
Benchmark Comparison

Figure 1: Project Gantt Chart