Understanding the Amazon Forest from Space with Random Forest, K-Nearest Neighbors and Convolutional neural network

Derek Ching, Yuetian Li, Guanzhou Song

March 10, 2019

1 Problem description

Understanding deforestation in Amazon has been increasingly important: we are losing an area of forest the size of 48 football fields every minute, and deforestation in the Amazon Basin accounts for the largest share.

In this project, we want to build a model and train it with a Kaggle dataset to obtain a high accuracy to detect and classify where and why the amazon rainforest are deteriorating. Because satellites are taking snapshots from outer space every so often, with high resolution imagery, we can use the algorithms from image classifiers to help us better distinguish between different labels.

In the given data, each pictures has two labels, the atmospheric conditions and the land cover/ land use phenomena. There are 4 types of atmospheric conditions: *clear, cloudy, partly cloudy, haze*, and 13 types of land cover/ land use phenomena: *agriculture, artisanal mine, bare ground, blooming, blow down, conventional mine, cultivation, habitation, primary, road, selective logging, slash burn, water*.

The *input* will be 40K image data provided by Planet Labs, obtained from Kaggle (`https://www.kaggle.com/c/planet-understanding-the-amazon-from-space`). Each image contains multiple labels: one weather and any or none from 13 different land labels. The *output* result for each image will be a list of probabilities to match with the tags for each image. As expected, we want to design a model that would generate accurate *output* predictions with the *input* data.

The reason we are interested in this topic is that image classification is a relatively new and "hot" topic to learn about since it pertains to AI and Machine Learning field. Image Classification is often used in NASA and other big aerospace companies who want to learn how to label satellite images based on various circumstances or changes in the weather and land usage. Also image classification is a big topic in deep learning which is relevant in the field of AI and Machine Learning.

Figure 1: The example of labeled images of different land cover/ land use phenomena

2 Algorithms

We will use 3 algorithms to solve this problem: KNN (K-Nearest Neighbors), Random Forest, and CNN (Convolutional neural network). We will also use cross-validation methods to tune the parameters to achieve the maximum classification accuracy. Derek Ching will work on KNN, Yuetian Li will work on Random Forest, and Guanzhou Song will work on CNN.

KNN is a classification algorithm which hardly does any learning. So getting poor results is quite obvious. It's efficiency depends on the features you're considering from your dataset. If we have a full-furnished datasets with good features, then KNN will give better results. Typically KNN is used for recommender system and image classification.

Random Forest is a robust algorithm that can be used for remotely sensed data classification and regression. It constructs a multitude of decision trees at training time and outputs the class that is the mode of the classes or mean prediction of the individual trees. By adapting this process, it corrects decision trees' habit of overfitting to their training set. Random Forest are used in many machine learning applications, including market prediction and image classification.

CNNs can be thought of automatic feature extractors from the image. While if I use a algorithm with pixel vector I lose a lot of spatial interaction between pixels, a CNN effectively uses adjacent pixel information to effectively downsample the image first by convolution and then uses a prediction layer at the end. CNN are now widely used in computer vision and speech recognition.

In our case, all three algorithms will be working on image classification. As described above, the 3 algorithms should be considered appropriate for this project.

There have been other data scientists and programmers who have used these algorithms before. KNN, CNN, and RF are algorithms often found in classic image classification applications.

3 Results

In order to evaluate our model, we will run each algorithm on the same test dataset and evaluate the result of the project with F_2 measurement score, given by

$$(1+\beta^2)\frac{pr}{\beta^2 p+r} \quad \text{where} \quad p=\frac{tp}{tp+fp}, \quad r=\frac{tp}{tp+fn}, \ \beta=2$$

The goal of this project is to reach at least 70% accuracy on classification result for the images. We will be using decision tree algorithm and logistic regression on weather prediction as our baseline, also we will make comparison between each of the three algorithms.

The risks for not getting all the results, are that these algorithms are very dependent on training data, we cannot get a high accuracy model without being able to train a large set of data. In some cases, a training of small dataset may prove to have different results than a large dataset. (ie, an algorithm may be better at classifying in earlier stages than a long term stage) We will be either be using a AWS server, NEU's discovery cluster, Google Drive or Google Cloud Platform to help aid our training process of large datasets.