Planet: Understanding the Amazon from Space

Bernardo Casares Rosa[1]

[1]bcasares@stanford.edu (06191000)

February 2018

Abstract

The project represents a solution to the multi-class classification competition from Kaggle: "Planet: Understanding the Amazon from Space" using convolutional neural networks (CNN) with Tensorflow.

1 Introduction

Deforestation contributes to reduced biodiversity, habitat loss, climate change, and other devastating effects. Understanding the location of deforestation and human activity on forests can help governments and local authorities to respond quickly and effectively.

The problem that will be investigated is the Kaggle data-set from planet lab: "Planet: Understanding the Amazon from Space". Kaggle Competition.

Planet lab is the largest constellation of Earth-imaging satellites and the objective is to correctly label 256 x 256 satellite images from the Amazon with several labels from atmospheric conditions, land cover, and use.

2 Data-set

The data-set consists of 40,479 training images with labels and 61,192 test images with no labels (both in TIFF and JPG format). Since there is no way to validate the test samples (competition is closed), the training images and labels were divided into 60 %train 20% development and %20 test sets and used for the project. The TIFF images have lower quality, therefore, only the JPG images were used.

The number of the images into train, development, and test set can be seen below.

Train	Dev	Test	Total
24287	8096	8096	40479

Each image is of size (256, 256, 3), with the channels representing R, G, B.

Figure 1: Example of Labeled Images (from the competition website)

2.1 Labels

Most of the images have several labels. However, there is no guarantee that the labeling is correct for all the images; scenes may either omit class labels or have incorrect class labels (detailed in the data description of the competition).

A histogram of the label frequency can be seen below.

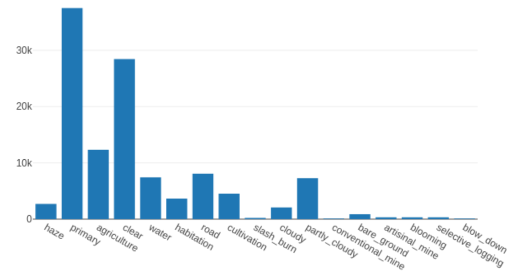

Figure 2: Histogram of Label Frequency (starter code Kernel)

In order to proceed with the analysis, it is important to know if the labels are co-related to each other. The heat-map below shows what percentage of the X label also has the Y label.

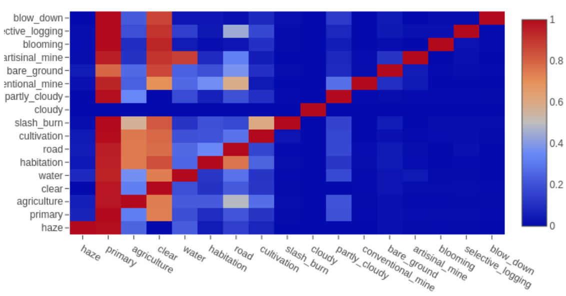

Figure 3: Co-occurrence matrix (starter code Kernel)

The label "primary", which is shorthand for primary rain-forest, has the highest proportion of labels.

It is interesting to note that each image should have exactly one weather label, but the land labels may overlap. The weather labels are: clear, partly_cloudy, haze, and cloudy, and the land labels are: primary, agriculture, water, cultivation, habitation. The co-occurrence matrices plotted as heat map can be seen in the appendices.

3 Approach

For the milestone of the project, only a baseline model optimizing the learning rate was preformed. The baseline model came from the Project Starter Package, and the code used was a modification of the github cs230-code-examples to fit this particular problem.

The steps to solve the problem were the following:

1. Re-size the images from 256x256 to 64x64 for ease of learning.
2. Add data by flipping with a random probability.
3. Build the CNN with several blocks. For each block, the architecture was the following 3x3 conv -> batch norm -> relu -> 2x2 maxpool.
4. Add two fully connected layers at the end.
5. Calculate the loss with sigmoid cross entropy for the 17 classes.

The results of the analysis with a bach size of 32, 10 epochs, 16 channels, using batch norm and momentum ($\beta = 0.9$) are the following:

	Acc	Loss
10^{-4}	0.947156	0.136272
10^{-3}	0.950818	0.127826
10^{-2}	0.949481	0.129836

4 Metric

The kaggle competition (when active) used the F2 score as an evaluation metric. The F2 score is a way of combining precision and recall into a single score – like the F1 score, but with recall weighted higher than precision.

One of the challenges of optimizing for the F2 score is that lower losses don't necessarily lead to higher F2 scores. Therefore, the models not only need to predict the label probabilities, but also select the optimum point to determine whether or not to select a label given its probability.

The F2 evaluation is as follows:

$$F2 = \frac{5}{N}\sum_{n=1}^{N}\frac{P_i R_i}{4P_i + R_i}$$

where P_i and R_i represent the precision and recall for each example. However, since Tensorflow calculates the precision and recall automatically, this project uses a slight modification of the F2 metric, defined as follows:

$$F2 = \frac{5PR}{4P + R}$$

where P and R represent the average precision and recall over all training examples.

The result from the baseline model with the a learning rate of 10^{-3} has the following recall, accuracy, and F2 score:

Loss	Precicion	Recall	F2
0.1278	0.8813	0.8187	0.831

F2 penalizes false negatives more heavily than it penalizes false positives.

5 Future Work

From the current F2 score, it is clear that baseline model can be improved. Some of the important approaches to try are different data augmentation techniques including rotating, transposing, and elastic transforming images as well as a haze removal technique, used by the leader of the competition.

Moreover, it will be interesting to see the results by trying different CNN architectures. Some of the best results from the competition come from a mixture of ResNets and DenseNets with different numbers of parameters and layers, as well an Inception and SimpleNet model. Therefore, the future work for the project include trying, combining, and fine tuning different techniques, and analysisng the results.

References

[1] Kaggle Competition: https://www.kaggle.com/c/planet-understanding-the-amazon-from-space

[2] Kaggle Blog from Winner: http://blog.kaggle.com/2017/10/17/planet-understanding-the-amazon-from-space-1st-place-winners-interview/

[3] Started Code from Kaggle: https://www.kaggle.com/bcasares/getting-started-with-the-data-now-with-docs/edit?unified=1

[4] Started Code from Kaggle: https://www.kaggle.com/anokas/data-exploration-analysis

[5] Jeff Pyke. Understanding the Amazon from Space with Convolutional Networks. http://cs231n.stanford.edu/reports/2017/pdfs/914.pd

[6] Sneh Kudli, Steven Qian, Benjamin Pastel. Kaggle Competition: Understanding the Amazon from Space. http://cs231n.stanford.edu/reports/2017/pdfs/913.pdf

6 Github Repository

bernardocasares/CS230_Final_Project
https://github.com/bernardocasares/CS230_Final_Project

7 Appendix

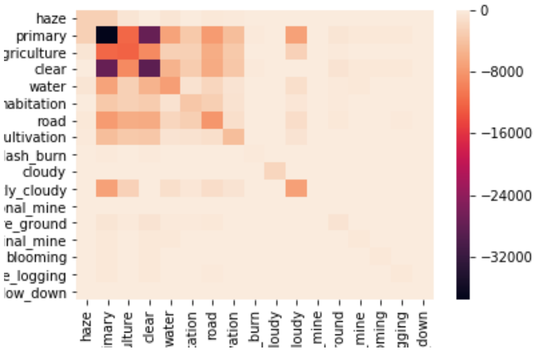

Figure 4: Co-occurrence Matrix for All the Labels (starter code Kernel)

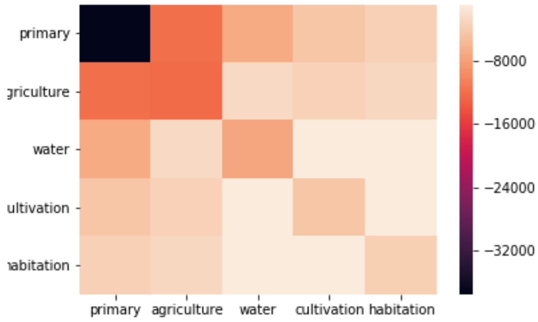

Figure 6: Co-occurrence Matrix for the Land Labels (starter code Kernel)

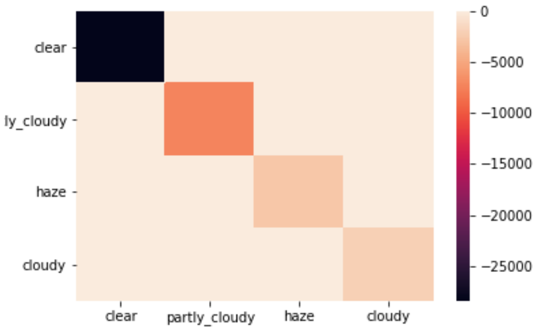

Figure 5: Co-occurrence Matrix for Weather Labels (starter code Kernel)

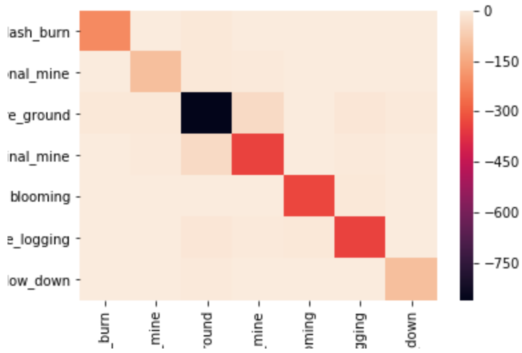

Figure 7: Co-occurrence Matrix for Other More-rare Labels (starter code Kernel)